Fakultät für Informatik
Technische Universität München

Political machines: machine learning for understanding the politics of social machines

Orestis Papakyriakopoulos

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Florian Matthes

Prüfende der Dissertation:
1. Prof. Dr. Simon Hegelich
2. Prof. Dr. Jürgen Pfeffer

Die Dissertation wurde am 03.03.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 15.05.2020 angenommen.

Dedicated to Chloe

Acknowledgements

First of all, I would like to thank my advisor, Simon Hegelich, who offered me the opportunity to pursue a doctoral degree in his group for the last three years. I am grateful for his advice, guidance and support. I also thank him for sharing important experiences with me these years, as well as for creating a flexible and open research environment. I would also like to express my gratitude to my thesis co-advisor, Jürgen Pfeffer, for advising and supporting me, as well as for reviewing this thesis.

Next, I would like to acknowledge the support of my colleagues and co-authors, especially Morteza Shahrezaye and Juan Carlos Medina Serrano. They contributed to an enjoyable work environment, exchanged with me scientific ideas and supported me personally and professionally during this thesis. I would also like to thank Martina Drechsel for dealing with all bureaucratic issues related to my research and Florian Schmidt, who assured the existence of the required infrastructure, always being available to deal with unexpected issues.

Beyond that, I thank all my friends outside my immediate academical environment that helped me stay motivated and complete this thesis; either in Munich, Greece, or another part of the world. Most importantly I want to thank my family, Irini, Nefeli, and Minos for giving me the required energy and support in research and personal difficulties. The same applies for my partner, Arwa, who helped me substantially with this thesis and was always next to me.

Abstract

The advent and consolidation of the internet together with the extension of hardware and software capabilities has created new sociotechnological ecosystems, in which individuals and social groups participate. Social media platforms, internet services and artificially intelligent applications constantly interact with the human factor, together co-creating social machines. As in every social environment, politics constitute social machines. Political processes, events and interactions both appear in social machines or are their objective. Therefore, when setting politics in the epicenter of investigation, social machines are transformed into political machines.

Five dimensions of politics take place on political machines. 1. The communication of participants always takes place by the use of text, images or other forms of symbols, which are always bound with the sociopolitical conditions existing in a society and carry predispositions and biases. 2. The application of artificial intelligent algorithms per se influences human behaviour and decision making. 3. Participants might explicitly talk about politics, making political communication a constituent part of many political machines. 4. The design of platforms and services imposes certain limitations and opportunities to the participants of the machines. 5. Legal boundaries and the state control the feasible space for the formation of political machines.

The first scope of this thesis is to define political machines and to evaluate how computational social scientists have analyzed them until now (chapter one). Given the scientific state-of-the-art, the next scope of the thesis is to uncover further properties of political machines given the 5-point framework set above. Research objective of the study is the investigation of political dimensions of algorithmic influence. The thesis achieves that by applying data-intensive machine learning techniques. Chapter two introduces the basic scientific consensus on knowledge discovery from databases (KDD). It also describes the probability and statistical foundations required for studying large scale social behaviour. Then, it provides an overview of machine learning and natural language processing techniques that can be used for the evaluation of political interactions and appear as part of them. The thesis introduces text-based methodologies for understanding politics. It presents three NLP techniques: topic models, word embeddings, and neural networks. The thesis also provides an overview of recommender systems, models extensively used by social media platforms for content personalisation.

Chapter three focuses on data-driven microtargeting, the technological state-of-the-art for performing political campaigning. By analyzing the legal frameworks in USA and Germany, it provides an overview of the possibilities and limits of the technique. Then, the chapter investigates how machine learning techniques are able to detect people's attitudes based on their social media activities, and discusses the ethical, political, and economic implications for the society.

Chapter four investigates social media platforms as ecosystems of political communication. It illustrates that the existence of hyperactive users influences the political discourse, because they become opinion leaders and have an agenda-setting effect. By performing simulations, the chapter shows that hyperactive behaviour can seriously influ-

ence the platforms' recommendation systems. Given the above, it questions the efficiency of existing social media platforms as spaces of fair political communication.

Chapter five focuses on the politicization of algorithms when used in decision making. It illustrates biases immanent in word embeddings, a set of natural language processing techniques for improving the quality of generative and predictive models. It shows that models trained on embeddings discriminate individuals and social groups and discusses ways of mitigating the traced biases. Given that word embeddings are widely used by commercial companies, the chapter discusses the challenges and required actions towards fair algorithmic implementations and applications.

The above case studies provide following contributions: chapter three analyzes the ethical, legal and political limits and possibilities of data-driven microtargeting. Chapter four evaluates how the interaction between social media platforms and algorithms influences political communication, uncovering immanent issues and dangers. Chapter five uncovers biases of text-based algorithmic implementations in decision making. It discusses methodological and ethical issues and proposes mitigation techniques.

Based on the above contributions, the discussion focuses on future research directions. It states issues where political machines are understudied and require scientific attention. Finally, it formulates the need for the formation of civic machines. That is, the design and analysis of political machines that protect the rights of social groups and assure the just ethical and political interaction of individuals and technological applications.

Zusammenfassung

Das Aufkommen des Internets sowie die Erweiterung der Hard- und Softwarekapazitäten haben neue soziotechnische Ökosysteme geschaffen, an denen Einzelpersonen sowie soziale Gruppen beteiligt sind. Die Interaktionen der Menschen in und mit sozialen Medien, Internetdiensten und Anwendungen künstlicher Intelligenz erzeugen hierbei sogenannte soziale Maschinen (social machines), wobei, wie in jedem sozialen Umfeld, auch hier die Politik immanent ist. Politische Prozesse, Ereignisse und Interaktionen erscheinen ständig in sozialen Maschinen oder sind deren Ziel. Wenn das Epizentrum der Untersuchung die Politik wird, so transformieren sich soziale Maschinen in politische Maschinen (political machines).

Fünf Dimensionen der Politik finden auf politischen Maschinen statt. 1. Die Kommunikation der Teilnehmer erfolgt immer durch die Verwendung von Texten, Bildern oder anderer Symbolik, welche immer mit den gesellschaftspolitischen Bedingungen verknüpft sind, und dabei Veranlagungen und Vorurteile mit sich bringen. 2. Die Anwendung von Algorithmen der künstlichen Intelligenz beeinflusst das menschliche Verhalten und Entscheiden. 3. Die Teilnehmer von technologischen Ökosystemen diskutieren teils explizit über Politik, womit die politische Kommunikation ein Bestandteil vieler politischer Maschinen wird. 4. Die Gestalter von Plattformen und Diensten führen für die Teilnehmer der Maschinen bestimmte Einschränkungen und Möglichkeiten ein. 5. Rechtliche Grenzen sowie der Staat kontrollieren hierbei den Raum der Möglichkeiten für die Bildung politischer Maschinen.

Der erste Abschnitt dieser Arbeit definiert politische Maschinen und gibt ein Überblick, wie Computational Social Scientists politische Maschinen bisher analysiert haben (Kapitel 1). Angesichts des wissenschaftlichen Standes der Technik beschreibt die Arbeit weitere Eigenschaften politischer Maschinen angesichts des oben genannten 5-Punkte-Rahmens. Ziel der Arbeit ist die Untersuchung von neuen Dimensionen des politischen Einflusses von Algorithmen. Sie erreicht dies durch den Einsatz von datenintensivem maschinellen Lernens. Kapitel 2 stellt den grundlegenden wissenschaftlichen Konsens über die Erkenntnisgewinnung aus Datenbanken dar und beschreibt die statistischen Grundlagen, die für die Untersuchung von großen Studien über das soziale Verhalten erforderlich sind. Anschließend gibt es einen Überblick über maschinelles Lernen und Techniken der natürlichen Sprachverarbeitung (NLP), die für die Bewertung politischer Maschinen verwendet werden können und als Teil davon erscheinen. Die Arbeit stellt textbasierte Methoden für das Verständnis von Politik vor. Es werden drei NLP-Techniken vorgestellt: Topic Models, Word Embeddings und Neuronale Netze. Die Arbeit gibt auch einen Überblick über die Funktion von Recommender Systemen (Empfehlungsdienste), die bei sozialen Plattformen für die Personalisierung von Inhalten angewendet werden.

Kapitel 3 konzentriert sich auf datengesteuertes Microtargeting, den technologischen Stand der Technik zur Durchführung von politischen Kampagnen. Durch die Analyse der rechtlichen Rahmenbedingungen in den USA und Deutschland gibt es einen Überblick über die Möglichkeiten und Grenzen der Technik. Anschließend wird untersucht, wie

Techniken des maschinellen Lernens dazu in der Lage sind, die politischen Einstellungen der Menschen basierend auf ihren Aktivitäten in sozialen Netzwerken zu erkennen. Es folgt eine Diskussion der ethischen, politischen und wirtschaftlichen Implikationen für die Gesellschaft.

In Kapitel 4 werden soziale Netzwerke als Ökosysteme der politischen Kommunikation untersucht. Es zeigt sich, dass sogenannte hyperaktive Nutzer einen hohen Einfluss auf den politischen Diskurs haben, da sie zu Meinungsmultiplikatoren werden und eine Agenda-setting Wirkung entfalten. Eine Analyse anhand von Simulationen zeigt weiter, dass das Verhalten von hyperaktiven Nutzern die Recommender Systeme der Plattformen stark beeinflussen kann. Vor diesem Hintergrund stellt die Arbeit die Effizienz bestehender sozialer Netzwerke als Räume fairer politischer Kommunikation infrage.

Kapitel 5 konzentriert sich auf die Politisierung von Algorithmen, wenn sie bei Fällen automatisierter Entscheidungsfindung verwendet werden. Es veranschaulicht Vorurteile, die in Word Embeddings immanent sind - eine Reihe von Techniken der natürlichen Sprachverarbeitung zur Verbesserung der Qualität von generativen und prädikativen Modellen. Es wird gezeigt, dass Modelle, die auf den Embeddings trainiert werden, Individuen und soziale Gruppen diskriminieren, woraufhin Wege diskutiert werden, um die verfolgten Vorurteile abzumildern. Da die Anwendung von Word Embeddings bei kommerziellen Unternehmen weit verbreitet sind, werden in diesem Kapitel die Herausforderungen und erforderlichen Maßnahmen für die Implementierungen von fairen Algorithmen und Anwendungen diskutiert.

Die oben genannten Fallstudien enthalten folgende Beiträge: Kapitel 3 analysiert die ethischen, rechtlichen und politischen Grenzen und Möglichkeiten des datengesteuerten Mikrotargetings. Kapitel 4 bewertet, wie die Interaktion zwischen sozialen Netzwerken und Algorithmen die politische Kommunikation beeinflusst und dabei immanente Probleme und Gefahren aufdeckt. Kapitel 5 untersucht Vorurteile textbasierter algorithmischen Implementierungen bei Fällen der automatisierten Entscheidungsfindung. Es werden außerdem methodische und ethische Fragen diskutiert und Techniken vorgeschlagen wie Biases reduziert werden können.

Basierend auf den oben genannten Beiträgen konzentriert sich die Diskussion der Arbeit auf weitere Themen, an denen in Zukunft geforscht werden soll. Sie stellt Problematiken politischer Maschinen vor, die wissenschaftliche Aufmerksamkeit erfordern. Schließlich formuliert die Arbeit die Notwendigkeit der Bildung von Civic Maschines. Das heißt, die Gestaltung und Analyse politischer Maschinen in einer Weise, dass die Rechte sozialer Gruppen geschützt werden und die gerechte ethische und politische Interaktion von Individuen und technologischen Anwendungen zugesichert wird.

Publications

This thesis, *Political machines: machine learning for understanding the politics of social machines*, includes five peer reviewed papers. Table 1 presents their information. The first two papers study political microtargeting on social media platforms. The third and fourth paper analyze the interplay between political communication, social media platforms, and recommender systems. The fifth paper investigates biases in text-based algorithmic decision making models.

Table 1: Dissertation related publications

Published Paper	**Journal / Conference**	**Year**
[1] Social Media und Microtargeting in Deutschland	Spektrum Informatik	2017
[2] Social media and microtargeting: Political data processing and the consequences for Germany	Big Data & Society	2018
[3] Distorting political communication: The effect of hyperactive users in online social networks	IEEE INFOCOM Workshops	2019
[4] Political communication on social media: A tale of hyperactive users and bias in recommender systems	Online Social Networks and Media	2020
[5] Bias in word embeddings	ACM FAT*	2020

The following list includes publications I contributed to *during* this dissertation:

- J. C. M. Serrano, S. Hegelich, M. Shahrezaye, and O. Papakyriakopoulos. *Social Media Report: The 2017 German Federal Elections*. TUM University Press, 2018
- S. Engelmann, J. Grossklags, and O. Papakyriakopoulos. A democracy called facebook? participation as a privacy strategy on social media. In *Privacy Technologies and Policy: 6th Annual Privacy Forum, APF 2018, Barcelona, Spain, June 13-14, 2018, Revised Selected Papers*, pages 91–108. Springer, 2018
- M. Shahrezaye, O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich. Estimating the political orientation of twitter users in homophilic networks. In *2019 AAAI Spring Symposium Series*, 2019
- J. C. M. Serrano, M. Shahrezaye, O. Papakyriakopoulos, and S. Hegelich. The rise of germany's afd: A social media analysis. In *Proceedings of the 10th International Conference on Social Media and Society*, pages 214–223. ACM, 2019
- M. Shahrezaye, O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich. Measuring the ease of communication in bipartite social endorsement networks: A proxy to study the dynamics of political polarization. In *Proceedings of the 10th International Conference on Social Media and Society*, pages 158–165. ACM, 2019

Contents

1 Introduction

The bulk of mankind believe in two gods. They are under one dominion here in the house, as friend and parent, in social circles, in letters, in art, in love, in religion; but in mechanics, in dealing with steam and climate, in trade, in politics, they think they come under another.

- Emerson, On Fate [11]

Technological advances always cause unforeseen social transformations. The implementation and diffusion of any technological application, regardless of whether it is the radio, the internet, or the car, transform social life and the social structure. New communication paths open and individuals, social groups and political actors exploit them towards their personal goals. In periods when many technological innovations appear simultaneously, the social equilibrium shifts to new states, radically reforming the function of society ethically, politically and economically.

One of these such periods marks the advent of the third millennium. The consolidation of the internet as the main communication network in society, the development of hardware of increasing computational efficiency, the invention of the smartphone, and the general social datafication [12] has created new communication and information processes impacting all aspects of society. Algorithmic decision making (ADM) systems, artificially intelligent algorithms, online social networking platforms, data sharing systems and predictive tools invaded everyday life, leading to the transformation of human behavior, socialization, the market and political conduct.

One of the most intensive reformations caused by technological advancement is the degree of digitization of social processes. The new computational and storing capabilities lead to constant transformation of individual behavior into metadata being processed and stored, while integrated internet access on any type of device facilitates permanent information exchange. Hence, technology, society and communication are coupled in a complex and continuous way, generating new forms of sociotechnological systems. The ecosystems, in which individuals and technology participate and interact, can be defined as *social machines* [13]. Because politics and power relations emerge in any type of social interaction [14], social machines can be analyzed as *political machines*.

Because data-intensive algorithmic implementations play a cardinal role in political machines, the objective of this thesis is to uncover properties of political machines related to human-algorithm interactions. To achieve that, I exploit the vast amounts of available data in political machines by applying machine learning models. The thesis studies multiple influence processes appearing in political machines, investigates political and ethical dimensions, and recognizes technical, legal, and regulatory gaps in ensuring fair and inclusive social conditions for algorithmic applications in society.

In the following, I present properties of social machines. I also present which influence processes transform social machines into political machines. Then, I describe how computational social science investigates political machines. Given the existing

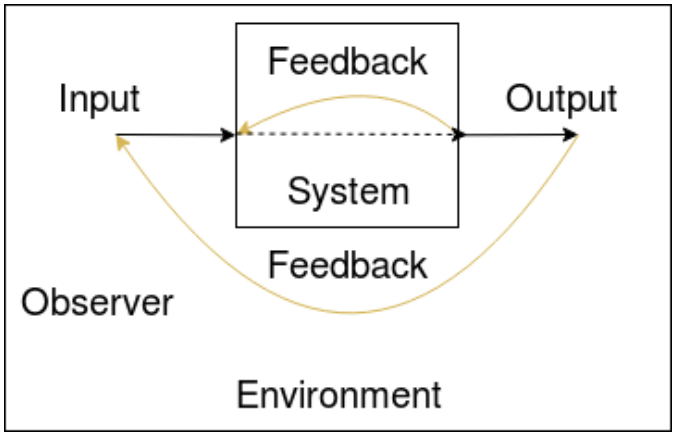

Figure 1.1: First order cybernetic system according to Gregory Bateson and Margaret Mead [21].

research gaps, I explain the thesis' contributions in providing new knowledge about human-algorithm interactions in sociotechnological ecosystems.

1.1 Social Machines

Understanding the political properties of social machines requires a framework that provides a holistic overview of how sociotechnological ecosystems behave. The most prominent scientific theory that deals with systems and their behaviour is cybernetics [15]. Cybernetics does not investigate systems by just looking them as a set of inputs, outputs and interacting components. On the contrary to other theories, it seeks to understand systems as they exist in a given environment, how their state changes according to the environment, what the systems' identity is, which constraints exist, and what processes of feedback, communication and control result in the transformation and self-organization of the system [16, 17, 18, 19, 20] (Figure 1.1). Cybernetics does not reduce systems to things, but to ways of behaving. It does not question *what is this thing?* but *what does it do?* [19].

In cybernetics, communication is not reduced to human or animal communication, which corresponds to an explicit exchange of symbols and signs [22]. Each type of interaction or impact between elements, systems, or the environment, can be treated as information that updates the related entities about differences taking place [23, 24], resulting in a form of communication of higher order. The realization of difference, or else feedback, is critical in cybernetics, because it is able to uncover operators and operants in the system, i.e. what changes what and how [19]. By studying feedback loops, the cybernetician is able to uncover the purpose of elements and systems, as well as their specific structure and intrinsic organization [20].

The dynamic analysis of feedback loops contributes to the detection of emergent properties of the studied systems. For example, in an apparatus that boils water the switch causes the heater to turn on/off, and the heater, similarly, causes the switch to turn on/off. This instantaneous causality is transformed into *circularity*, which defines the emergent identity of the system [25]. This identity is coupled with associated regularities, invariant trajectories of the system, or equilibria which are called *eigenbehaviors*

[17]. These eigenbehaviors change when new information/feedback is presented either in the form of constrain or allowance transforming operators and operants and changing the system's closure [19]. The closure is defined by the structure of communication and of control (regulations) that separate a system from the environment and guide it to self-organize in a new eigenbehavior [17]. The system's closure thus regulates regulations, making the system *autonomous* in respect to its environment [17].

Wiener developed the theory of cybernetics when trying to understand human-machine interaction in anti-aircraft weapons [26]. Because of the speed of the aircraft and the delay in the person triggering fire, the miss-rate was really high. To understand and solve the problem he conceived a system consisting of the airplane, the operator of the anti-aircraft gun and the anti-aircraft itself. The operator communicated with the aircraft by sight, getting information about its position. They then aimed the gun according to it. Because the aircraft moved, the operator got new information (feedback) and changed the aim of the gun, a process that took place repeatedly, constituting a circularity. Because firing the gun always took some clusters of the second, when the missile arrived at the target, the plane had already moved to a new location. This circular behavior led to a regularity, an eigenbehavior, represented by the failure of the anti-aircraft to hit its target in most cases. For Wiener, altering the communication and control processes in the system in a way that the gun hit each target, would denote the system having a new eigenbehavior.

From cybernetic systems to social machines

The example that inspired Wiener was not just a cybernetic system, but also a primitive social machine. Social machines are cybernetic systems that include both the human factor and technology as participants in immanent processes [13]. They are not per se machines, nor do they depict mechanistic deterministic phenomena. On the contrary, they are systems of systems [27] in which humans and technology are detached from their materiality, forming complex patterns of communication and control. Online social networks, algorithmic decision making systems, autonomous cars, are all types of social machines, with individuals, software and hardware constantly interacting and resulting in emergent system states.

In social machines, computability is not an exclusive right of the machines, nor is sociability an exclusive right of the humans. The cybernetic framework allows treating human behaviour as also computable, and the technological participation as sociable too. For example, human behaviour is projected into metadata fed into recommendation algorithms, deep learning models, or computer vision software. Similarly, the decision of an ADM System to hire or fire an individual replaces the human resources manager in a company's social network. Given that all these interactions are projected into forms of communication and control, they appear in the same cybernetic domain regardless of their initial materiality. What complements their regulation is the design frameworks that guide the behaviour of individuals and the application of technologies (figure 1.2). The design framework includes the values, infrastructure, exact algorithms, interfaces, regulations, and any other material or non-material property that constitutes the systems' elements behavior [28]. For example, social media platform design is usually based on companies' business models, often prioritising interactions that promote efficient advertisement placement instead of the optimal interaction between users [29].

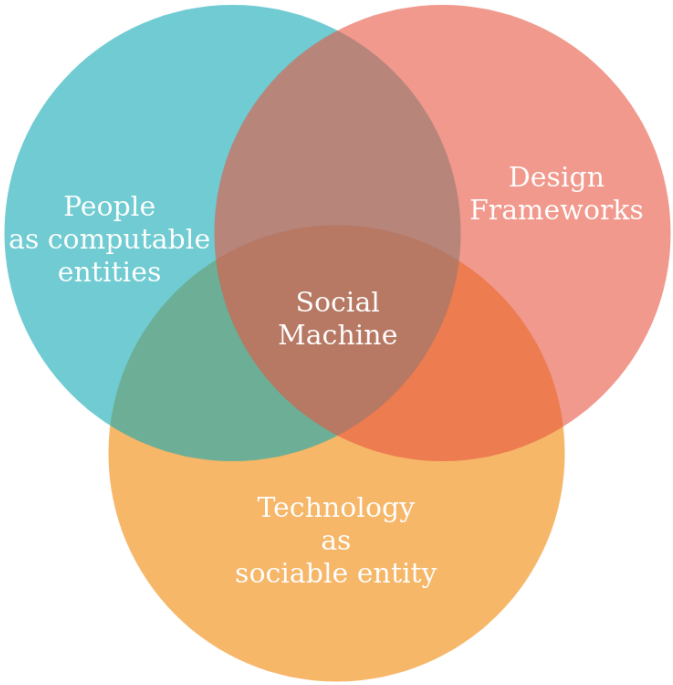

Figure 1.2: Social machines are shaped by human behavior, technological implementations and design frameworks.

Studying social machines from a holistic framework such as cybernetics becomes even more important given the nature of contemporary human-algorithmic ecosystems. In the era of *ubiquitous computing*, individuals and computers are integrated in systems of circular communication and control. Individuals are constantly enhanced by algorithms integrated into most technological artifacts, be those navigation tools, social network platforms, or search engines. This pervasive, persistent, invisible and continuous existence of algorithmic applications in every aspect of human life [30, 31] distorts the classical limits between personal social autonomy and connectivity [32], violates assumptions of classical causality and generates a networked space-time [29] in which the role of operator and operand are constantly exchanged between humans and algorithms.

A piercing example of this transcendental process where technology is no longer an aiding tool for humans but where humans also become an aiding tool of technology is *social computing*. The efficiency of contemporary data-intensive algorithms is mainly based on the quality of input data, which should reflect clearly, in a detailed and unbiased way, every aspect of social behaviour. Thus, humans transform themselves into datafied artifacts, offering every aspect of their lives to algorithms in order to optimize the latters' function. This can not only be seen in the rise of social machines largely based on crowd sourcing [33], where people actually do the creative work and show computers what to learn and how [34, 27], but also in the emergence of new structures in social and political conduct. For instance, state-of-the-art campaigning is based on the generation of data-intensive models about the electorate and the extraction of information from them about voter interests and behaviors, which are used for the generation of ads and adapting parties' profiles [35, 36]. Thus, the electorate is transformed into data for the algorithms, which then provide specific inferences to political actors, which are

then transformed into actions that influence the electorate, generating a circular loop containing social and computational mechanisms, in which the notions of cause and effect become inapplicable.

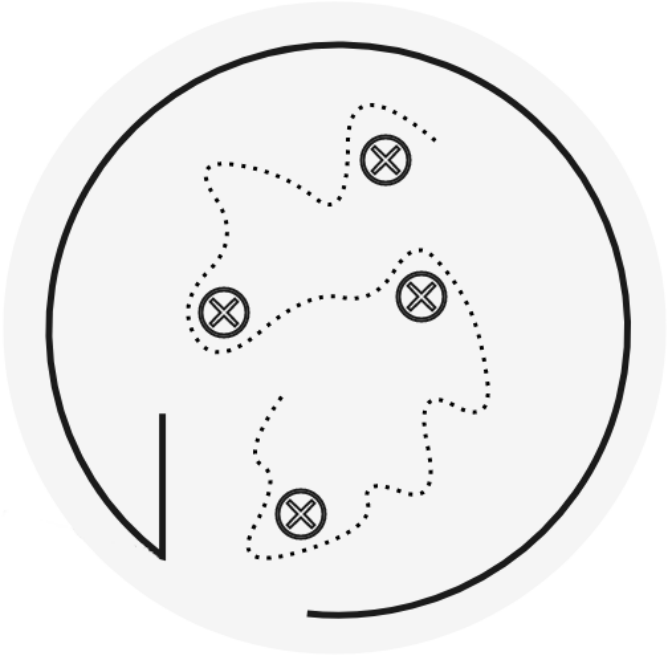

Figure 1.3: Sociotechnological entanglement: Individuals and technology are participants in social machines, interacting by processes of communication and control, and forming stable states that define systemic behavior.

What remains constant in such human-algorithmic ecosystems is not the materiality of humans and algorithms, where computability and sociability become interchangeable [37], but the phenomenological system eigenbehaviors formed by communication processes, which are dependent on how human and algorithms interact and influence each other [38, 39]. The participants of these systems wayfare in time-space, generating dynamic meshworks of interactions, extracting information, and adapting their behavior [29], often generating fabrics of sociability and memory [40]. For example, a dating app's emergent community is dependent both on the users' behaviour and generated data on the platform and the ability of its algorithm to match people according to their attitudes. Similarly, the deployment of an ADM system for recidivism purposes is only feasible if it is able to remember and retrieve people's general behaviour based on the data it was trained on.

Given the above, social machines are systems in which humans and technology are coupled, resulting in sociotechnological entanglement [29] (figure 1.3). As participants in the systems, individuals and technology constantly interact through communication and control, translating the systems into potential points of stability [41]. Such systems should be analyzed with appropriate scientific tools, in order for someone to understand their properties, as well as for the society itself to normatively decide on their design and function. Next, I analyze how these social machines are transformed to political machines, i.e. locate political properties of social machines. I limit the analysis to two classes of social machines, social media platforms and ADM systems.

1.2 Political Machines

The study of politics denotes the study of influence and of the influencial [42]. Social machines de facto include uncountable influence processes, because individuals and technology interact constantly on them. For example, a recommendation system changes the behavior of the users on a platform, while users' behaviour changes the system's recommendations recursively. Similarly, how a social network is designed, what its purpose is, and what the interacting possibilities of users are, change both user behaviour and platform's algorithmic design. A commercial or political actor per se deploy an ADM system in order for their decisions to be influenced by the model's results. Influence processes on social machines appear constantly and can be of various types, concerning different participants. Given that, I develop a five-point framework for understanding political properties of social machines. Because political processes constitute social machines, I will refer to them and analyze them from now on as political machines.

On political machines five main types of influence take place: A. symbolic influence, B. political conduct, C. algorithmic influence, D. design, and E. regulatory influence. Each of the above categories contributes to the formation and identity of political machines, how the system components interact and change, and how systems reach their equilibrium. In the following, I explain each of the types of influence on social media platforms and ADM systems, which are at the epicenter of this investigation.

A. Symbolic influence

The most basic form of influence on political machines is bound to human cognition and is of symbolic nature. The individual, in order either to perceive the world, or to explain and communicate it, deploys symbols [43]. Human language and thought consists of words, which are nothing other than symbols composed by signifiers and signifieds [44, 45]. These symbols always carry social conditions and meanings, informing the individual about the world and influencing their behaviour. For example, the explicit inclusion of a text for opting in to a platform's terms and conditions has the potential to change the decision of a user in using that service. Similarly, making a list of binary genders for a user to select from when they create an account functions as a proxy of social power, social group (in)visibility and illustrates existing social inequalities [46, 47]. That also happens when users converse on platforms, even about non-political issues, with the generated text and discussions revealing and reproducing dominant social group attitudes and perceptions [48], which are then passed on to the reader.

Symbolic influence does not only appear in the context of language. Non-verbalised information, in form of stimuli, such as seeing shapes or colours, can be responsible for influencing an individual. Such information is stored in human memory as mental representations or information schemes [49], which are reactivated, retrieved and deployed depending on new incoming information. Therefore, the appearance and structure of a UX can always influence participants in political machines, changing their behavior. For example, a platform's UX color can influence how much time a user spends with a service [50], or how much and in which way they would interact with it [51].

Because implicit or explicit symbols always appear in social machines, symbolic influence is the most subtle and penetrating type of influence. It is always there, but the extent of its impact is practically impossible to quantify. Nevertheless, in specific

cases with appropriate experimental design, researchers can investigate and understand properties of symbolic influence [52].

B. Political conduct

The most straightforward way that politics appear on social machines is when political conduct explicitly takes place on them. Within social media platforms and ADM tools, that can happen in multiple cases and in different ways.

Although most prominent social media platforms were not designed to foster political discussions, nowadays they serve as central spaces for political exchange, campaigning and communication. Users utilize platforms to comment on civic and political issues, externalize their political ideologies and form online groups of political action [53, 54]. This ample space for political interaction generated hopes and promises for a more diverse, open and democratic political discourse [55]. Social media platforms were treated as space for more autonomous political acting [56], which could contribute to the diffusion of alternative voices that were systematically repressed by authoritarian regimes and power structures [57, 58, 59]. These expectations were largely generated because social media can be seen as a space for information, connection, mobilization, deliberation and diversity [60].

Nevertheless, social media as spaces of political conduct became mainstream and got exploited by various political actors. Contemporary politicians and political parties maintain pages and profiles on social media platforms in order to interact with the electorate [61]. Simultaneously, they deploy large scale political campaigns in order to influence public opinion, especially in the form of personalised advertising [62]. The openness of social media is exploited by other actors as well, with automated, fake, and militant accounts spreading misinformation [63, 64].

All the above features constitute social media as political machines of high complexity. Even the political processes taking place explicitly on them are of different forms, with multiple participants using the services towards their individualistic goals. In this context, many debates take place to investigate whether social media actually contribute to the democratisation of society or actually have a negative political impact [65, 66, 67, 68, 69].

Besides political conduct taking place in social media, a variety of ADM systems are deployed for political purposes. Political parties hold large databases containing demographic and personal information, on which they train data-intensive models used for decision making in political-campaigns [36, 35]. In many cases, computer vision tools assist police in detecting suspects [70], while there are many companies developing systems that quantify the propensity of individuals' committing crimes. These systems are often used by the criminal justice system [71]. Especially in legislature, ADM systems are increasingly developed and deployed for automating legal evaluations and detecting violations of law [72].

It is clear from the above that technology and political conduct intersect both on social media platforms and in ADM systems, with political actors using technology for their own goals. Given the complexity of interactions, many open questions exist about the direct politicisation of technology and its recursive impact on the political processes taking place.

C. Algorithmic influence

Since the thesis focuses on human-algorithm interaction on social machines, it is important to understand how algorithms directly influence individuals. Algorithmic influence is a cardinal part of many political machines, because services and political actors explicitly deploy algorithms for automating processes, affecting individuals and the society. The diffusion of algorithms in the society takes place in both political and non political settings, with ubiquitous computing covering every aspect of socialization. From google maps to online content suggestion, human behaviour is constantly reshaped by algorithmic implementations.

On social media, platforms' designers deploy algorithms to 1. suggest personalized content to users, 2. to place targeted advertisement, and 3. to filter and review the contents generated by users. All three different algorithmic implementations have the potential to change human behavior.

By selecting which contents are going to be visible to a user's news feed, an algorithm results in reality construction [73]. What a user perceives about the world changes in respect to the selected pieces of information, leading to an algorithm-mediated subjective knowledge. That knowledge is then transformed into actions, with users forming opinions about the world and actively behaving according to them in the online and offline world. In this context, it has been largely hypothesized that algorithms can result in filter-bubble phenomena [74]. Filter bubbles are segregated opinion clusters formed by the algorithms, in which users only come in contact with conforming opinions but not to opposing ones, a social setting that can easily lead to opinion polarization.

Even if this content curation does not lead to polarization, it always introduces a bias, because the algorithm-mediated reality is always dependent on an algorithm's structure and the related input data. This leads to the emergence of data politics [75] that concern themselves with how algorithms function [76], what biases they introduce [77, 78], and how they can influence the individuals and social groups [79, 80, 81]. Data politics do not restrict themselves on recommendation algorithms on social media, but also include the platforms' services for personalized advertisement in the form of microtargeting [36, 35]. These opaque algorithms place advertisements to users according to demographic and behavioral criteria with the aim of efficiently influencing user behaviour. Microtargeting is a state-of-the-art technique in political campaigning, while the platforms' business models largely depend on convincing commercial companies and political actors to rent these services for advertising.

Another dimension of algorithmic influence on social media is related to content filtering algorithms. Companies largely use automated processes that scan uploaded images, videos, and text and search for contents that violate the platforms' terms and conditions. These algorithms therefore decide what is allowed to become part of the open discourses and what is not, how freedom of speech is constituted on the platforms, and coordinate the development of user behavior.

Algorithmic influence is equally present in ADM algorithms. ADM systems are largely used for risk assessment and warning [82] and for automated tasks such as image recognition, speech understanding, medical consulting, and predictive policing [83, 84, 71]. ADM systems result in bidirectional algorithmic influence. First, they per se influence the behavior of the users who deploy the models, because they generate knowledge that is exploited in multiple decision making processes. Second, in the case that algorithms are also making decisions about individuals and social groups, their decisions influence

these groups. For example, a hiring algorithm does not only influence the company by suggesting a candidate, but also influences the candidates themselves, deciding who will get a job or not [85]. An algorithm that recommends treatment type for patients to a doctor not only helps the doctor in taking the optimal decision, but also chooses whether patients should get operated on or not, how long their recovery period will be, etc.

Algorithmic influence always takes place, both on social media and in ADM systems. Because algorithmic impact is dependent on multiple parameters such as input data, social structures, and designers' choices, there is an open question how and under what conditions algorithms remain neutral media in decision making and content management. Multiple cases show that algorithmic implementations discriminate individuals and social groups, resulting in unfair decisions and influencing public opinion [86, 87, 88, 89, 90]. Therefore, algorithms have the potential not only to change human behavior, but also to perform these changes in an asymmetric way, which often violates ethical norms and social expectations. Given the above, algorithmic influence is of high complexity and dimensionality; it is a challenge for researchers to understand it and for political actors to regulate it.

D. Design

The fourth dimension of politics on social machines is related to the systems' design. How symbolic influence, algorithmic influence, and political conduct take place, depend on the structure of the social machines, which are largely given by the design of their components. Each component of a social machine, regardless whether it is a social platform or medical imaging tool, takes its final form given the designers' objectives and existing environmental constraints [91]. This final form contributes to the resulting equilibrium in a social machine. For example, the design principles of a social credit system influence the behavior of citizens in a society, setting the peoples' feasible action space, and forming their social goals [92]. Similarly, a social media recommendation system suggests contents to the users in a way that aligns to the company's business model goals.

Design constraints have also a huge impact on the formation of social machines. Hardware or software limitations can result in discriminative model predictions [93] even if that was not part of the designers' intentions. The ability of a model in predictive medicine to make good decisions is dependent on the available data, which because of privacy constrains be scarce, and therefore lead to a model deployment with lower predictive ability.

Besides the designers' goals and environmental constraints, a parameter that strongly influences the formation of social machines are design ethics. The decision of tech companies to gather data about users' interests, traits, demographic and behavioral information, and exploit them into developing better algorithms is always dependent [94] on the owners perception of what is ethical, what is necessary for achieving their goals, and what is allowed by the state. The fact that companies do not disclose how their systems work, maintaining high level of opacity in every aspect of the models development and deployment, is a design property that obstructs the understanding of the systems and forms the relations of accountability and transparency [93] between the state, users, and systems owners. Especially in cases of auditing algorithms and trying to trace their potential discriminatory impact or political influence, such design properties obstruct researchers from interpreting phenomena and the society to govern them [95].

Issues like the above raise questions about how to ideally design social machines that serve the society in an optimal way, given that many technological ecosystems are driven by financial incentives [96], having unknown transformative effects in politics and society. For example, although political communication largely takes place on social media, these are not public, nor do they try to always remain politically impartial [7, 97]. The same applies for ADM systems that are deployed either by the state or are of high social importance [98], which many times formulate inferences in terms of mathematical probabilities [99]. Arguing, justifying and legitimizing an action based on a calculated probability is sometimes ethically inadequate.

The above cases are only some examples about how design values, creators' incentives and environmental constraints can influence the formation of political machines. The analysis of each social machine can reveal multiple design properties that constitute the participants' interactions on them. Therefore a detailed analysis is necessary for an exact political evaluation of them.

E. Regulatory influence

The last form of influence on political machines is related to regulatory frameworks. Because any type of social machine is embedded in a society, and since societal function is controlled by institutionalized processes [100], political and legal structures set the space in which social machines can function. In algorithmic applications on social media and on ADM systems, legislature decides how these systems should be deployed, and how can the interests of the designers and the public be protected. The main regulatory issues for algorithmic applications on social media and ADM systems are related to data property and privacy, algorithmic opacity, and discrimination of social groups. In the following, I provide an overview of these topics.

- **Data property & privacy**
  One of the main reasons for the current intensive application of algorithms in the society is the datafication that takes place since the beginning of the third millennium. The vast amount of created data related to human behavior can be exploited and used towards multiple ends, with new business models being born unstoppably. Data are generated, collected, processed and combined for decision making, political and commercial acting, raising questions about whom these data actually belong to, what kind of rights a data collector has, as well as if the collection and data processing violates individuals' rights on privacy [101]. For answering such questions, states possess various different regulatory frameworks that define what is allowed and what is not [102].

- **Algorithmic opacity**
  One of the main designers' rights in algorithmic implementations is their legal protection in not disclosing their models' inputs, structures, and outputs. The reason behind that is that a developed model can provide its owner better opportunities in the market, therefore its features can remain secret under the fear of competition. Nevertheless, the resulting algorithmic opacity obstructs the auditing and understanding of such systems, especially when it comes to algorithmic impacts that violate the law [103].

- **Discrimination**
  Legal frameworks interfere on social machines regarding discrimination in two ways. First, as already discussed, algorithmic implementations might result in discriminatory decisions against individuals and social groups. Especially for ADM applications, practice proves that such events can happen frequently, raising questions about the extent to which existing legislations are violated regarding protected social groups, and individual rights and freedoms [104]. Second, on social media and other online platforms, algorithms are deployed for content filtering. Algorithms filter out videos, images, and text both according to the designers imperatives and each country's legislation. In this process, questions arise about who is legally accountable for contents that were wrongly not filtered, contents that were mistakenly filtered, as well as on the companies' freedom to choose what to filter.

Overall, algorithmic implementations in social machines remain largely unregulated. Given that, a lot of discussions take place about algorithms and their definitions [105, 106], current and ideal functions [107], how regulations could prevent algorithmic biases [108], as well as who should be accountable in cases of misconduct [95, 109].

1.3 Computational social science for political machines

The above five-point framework of political machines structures influence processes and provides a guide for interpreting interactions on sociotechnoligical ecosystems. Nevertheless, for understanding them in detail, more detailed scientific tools are necessary that can map interactions and integrate them in existing and new theories. Because this thesis investigates human-algorithm interactions on political machines, these tools should be able to analyze and efficiently structure the huge and diverse amount of generated social data. The field that combines data-intensive modeling and social theories is computational social science [110], on which I built upon my research questions, methods, findings and contributions.

Computational social science has extensively analyzed various social machines, uncovering political properties and behaviours that constitute these systems. In the following, I provide an overview of important findings of computational social science in social media and ADM systems, with a focus on the impact and function of algorithms on them.

Computational social science for social media

Researchers have investigated multiple political dimensions of social media regarding political communication, misinformation campaigns, polarization, and political campaigning. They have also analyzed the limits and possibilities of using social media for creating unbiased inferences about the world and general behavioral properties that constitute these systems.

Pfeffer et al. [111] described how discussions and information are diffused in social networks, while Castillo et al. [112] showed what properties and life dynamics news article distribution possesses on social platforms. Singer et al. [113] and Kwak et al. [114] studied how social media population' behavior changes over time, and Malik et al. [115] illustrated how platform design affects user behaviour. Furthermore, many

researchers have designed models for predicting user behaviour given specific platforms and case study properties [116]. The above studies seek to understand general patterns of behaviour, while studies exist that try to understand user behaviour in specific areas. For example, Hegelich et al. [61] investigated how politicians use social media to interact with the public, and Wagner et al. [117] investigated how the gender ratio in Wikipedia contributors influence the generated content.

A set of studies dedicated themselves to understanding properties of spreading misinformation by original or artificial users. Varol et al. [118] investigated how promoted campaigns dynamically develop on the platforms. Thieltges et al. [119] illustrated general properties of manipulation attempts on social media, while Del Vicario et al. [120] studied the properties that constitute successful misinformation diffusion. Comparing real and automated accounts, Vosoughi et al. [121] investigated their tendency to distribute false news. Numerous studies investigate automated accounts explicitly. For example, Hegelich et al. [122] studied the conditions that constitute social bots as political actors, Ferrara et al. [123] and Wagner et al. [124] developed general models for detecting social bots, and Thieltges et al. [125] described ethical dimensions in bot detection.

An important issue extensively studied by computational social scientists lies in the intersection of information diffusion, recommender systems, and opinion formation and concerns group formation and polarization of users on the platforms. Many researchers have studied and described how recommender systems function on big social networking platforms [126, 127, 128], as well as how they influence user attitudes and preferences [129, 130, 131]. These studies set the algorithmic factor in the center of their analyses. Other studies focused on understanding political interactions and user behaviour [132, 133, 134, 135], as well as social group formation on the platforms [136, 137, 138, 139], with the human factor being under the limelight.

A third set of investigations analyzed the behaviour of social groups under algorithmic influence, in order to uncover whether public opinion on social media is constituted by filter bubbles or echo chambers. Echo chambers denote the formation of segregated opinion clusters given the tendency of individuals to socialize with people who have similar attitudes to them [140]. Filter bubbles refers to the phenomenon in which algorithmic personalization is responsible for the formation of these segregated opinion clusters [74]. Researchers have performed numerous studies trying capturing when and how polarized clusters emerge, and are still debating when, how, and if recommender systems contribute to the phenomenon [10, 141, 142, 143].

Besides user behaviour on social media, most of the services are platforms for personalized advertisement. Political parties, candidates and actors exploit this property and use the available tools for performing data-driven political microtargeting. Computational social scientists have uncovered multiple properties and issues with this campaigning technique. Endres [144] and Kruikemeier et al. [145] investigated the impact of microtargeting on individual knowledge creation, while Bodó et al. discussed its limits and capabilities to influence individuals. Schipper et al. [146] conducted simulations to investigate the efficiency of microtargeting in comparison to other information settings, and Hegelich et al. [147] investigated campaigning strategies during election periods. Besides technique efficiency, further studies [148, 149, 150] set the feasibility space of microtargeting given legal, financial, and ethical boundaries. They investigate the in-

terplay between data protection, individual freedom and rights, as well as parties and platforms' rights and responsibilities.

Apart from analyzing phenomena on social media, researchers have investigated the importance and quality of social media data. The use of social media data comes with biases related to the sampling processes defined by the platforms [151, 152], as well as to their actual power to represent real world phenomena [153, 154]. Both properties should always be taken into consideration by scientists, in order to assure their research results' validity. Another issue connected to the quality of the data is *which* data are available to researchers. Previous studies show that is important not only to have a high amount of data, but also a great variety of features for adequately modeling social machines [155, 156, 157].

Overall, computational social scientists have analyzed multiple properties of social media as political machines. Nevertheless, the complexity of phenomena and the constant advancement of technologies generates new questions.

Computational social science for ADM systems

The rise of social computation resulted in the generation of additional data sources that decision makers can exploit. Next to classical ADM systems used in areas such as mechanical and electrical engineering, and weather forecasting, new systems are being developed for various purposes such as autonomous vehicles, healthcare, economics, finance, employment, policing, and public administration [158, 71, 159, 160]. These systems exploit data-intensive algorithms, generating inferences about individuals that were not possible to do so before. Most researchers developing these models are primarily interested in testing their efficiency and accuracy in comparison to other models and the human factor [71, 158, 161, 162, 163, 164]. Nevertheless, computational social scientists focus on ethical consequences of these methods. Whether they are fair, whether they discriminate individuals and social groups, and how should these systems be regulated.

Many studies prove that ADM systems treat individuals and social group unfairly, e.g. in computer vision [165], predictive policing [71], text-based models [166, 167], and algorithmic personalization [78]. Because of the uniqueness of each case study and the form of discrimination, multiple researchers investigate how and under what conditions an algorithmic decision is unfair. For example, Heidari et al. [168] investigate fairness under a Rawlsian conception, while Barocas et al. [160] compare definitions of procedural fairness and distributive justice. Kusner et al. [169] propose a framework of counterfactual fairness, and Joseph et al. [169] investigate fairness given the contextual bandits problem.

The election of an appropriate fairness definition is important, because it actually dictates how an algorithm should be designed. Depending on the definition, researchers have developed multiple statistical criteria that algorithms should follow. These might be independence, separation, sufficiency [160], statistical or predictive parity [170]. Other design techniques might account for latent associations in data that might lead to discrimination but are not directly visible [171], or use causal modeling for controlling the impact of specific variables on models' predictions [172].

Many design techniques are developed according to legal criteria. States hold regulations about how and when an individual is discriminated against [173, 174, 175], and can dictate whether someone belongs to a protected class [160, 176] and how should they be treated. Therefore, computational social scientists seek methodologies that comply

to such rules, as well as investigate why those rules might not be factual [177, 178]. To that end, scientists are developing new tools that promote algorithmic explainability [179, 180], because algorithmic bias can be generated at different stages of the model design process [181, 182]. This might be because of input data, modeling techniques, or designers imperatives and interpretations. Therefore, detecting biases can contribute to understanding issues within ADM systems and performing the required actions towards fair algorithmic implementations.

Explaining and understanding algorithms is not only related to fairness but also to accountability and transparency. Given that legal frameworks, algorithmic design, and algorithmic influence interact with each other, scientists are trying to pose the correct questions to be answered. Towards that end, researchers have analyzed the interaction between data regulations, accountability, fairness, and the right to explanation [183]. Furthermore, they seek to uncover further cases of algorithmic bias [178], how ADM systems influence when they are in production [184, 185], and to detect further challenges in algorithmic fairness in order to form regulations and systems that conform to social imperatives [186, 187]. Although computational social scientists have analyzed many properties of ADM systems, the most recent development of these political machines denotes a terrain full of unknown unknowns to be discovered and understood.

1.4 Theoretical contributions

In the previous sections, I introduced a framework for analyzing the politics of complex sociotechnological ecosystems and defined the scientific background in which the thesis builds on. Drawing from cybernetic theory, I pointed out that human and algorithmic interactions can be seen as social machines (1.1). Social machines are systems in which algorithms and individuals are participants, influencing each other and shaping these environments as a whole in a unique way. Because influence processes are the essence of politics and such processes constitute social machines, I claimed that they can be analyzed as political machines. I defined a five point-framework that classifies political processes in political machines (1.2). This framework includes A. symbolic influence, B. political conduct, C. algorithmic influence, D. design, and E. regulatory influence. Because the thesis focuses on social media and ADM systems as political machines, I provided examples of influence processes on these systems. As the study of political machines can be facilitated by analyzing the enormous amount of algorithmic and human traces generated between their interactions, I adopted computational social science as the theoretic and methodological background for my dissertation (1.3). I then provided an overview of existing studies about politics on social media and ADM systems, and illustrated that fertile ground exists for further investigations.

Given that computational social science poses questions about political interference of social platforms' algorithms on political discourse, as well as discusses the ethical dimensions of algorithmically generated decisions, this thesis seek answer to the following question:

- **Research Question:** What are unobserved political dimensions of algorithmic influence on social media and ADM systems?

I investigate this question by building on existing research performed by computational social scientists, and taking into consideration the conditions of systems' design and

regulation, and the processes of political conduct and symbolic influence taking place. For answering this question I analyze three unique case studies:

- A. Social media and data-driven microtargeting
- B. Political communication and recommender systems
- C. Word embeddings and unfair algorithms

The investigation of the above case studies results in the following theoretic contributions:

A. Social media and data-driven microtargeting

- I examine the legal and technical possibilities and restrictions of collecting and analyzing users' political data in USA and Germany. I prove that in both cases social media data can be used for political purposes, although data gathering and exploitation in Germany is a much more complicated process than in the US.
- I build machine learning models for processing the vast amount of users' political data on social media and show that standard machine learning techniques such as topic modeling are able to structure datasets and locate exact users' political interests.
- I locate a political power asymmetry in political microtargeting, with tech companies as data holders prevailing as key stakeholders in political campaigning. I discuss tension points and issues on how political campaigning is performed in the digital era.

B. Political communication and recommender systems

- I investigate how algorithmic recommendations influence political communication on social media. For achieving that, I first analyze user behaviour on social media and prove that a major amount of the discussions are shaped by hyperactive users. Hyperactive users are users that are overproportionally active in relation to the mean. I prove that hyperactive users become opinion leaders and have different political interests than regular users. I then discuss the implications for political discourse equality.
- I study how asymmetric user behavior influences recommendation systems and illustrate issues in models' training process. By performing simulations, I replicate properties of actual systems used by social media companies and uncover vulnerabilities. I insert adversarial examples and attack the models, manipulating the generated political suggestions in user networks.
- I discuss emerging political problems related to the algorithmic influence of political communication, the opacity of the systems deployed by tech companies and discuss how politics should take place on digital platforms.

C. Word embeddings and unfair algorithms

- I investigate biases related to word embeddings, a standard set of natural language processing models used for the improvement of ADM systems' inferences. I show that models trained on different datasets contain different types of biases depending on input text. By tracing sexism, homophobia and xenophobia in the embeddings, I illustrate that these biases are diffused unaltered to further machine learning models resulting in discriminative decisions.

- I investigate mitigation techniques and show that mitigating biases at the embeddings level is not sufficient. The unfair impact of algorithms should always be investigated, quantified, and mitigated at a system's end-product. Only in this way can scientists and policymakers assess models, given that biases can emerge in various stages of model development and deployment.

- Because many applications use word embeddings uncontrollably, I discuss the political dimensions of such ADM systems usage and discuss potential regulatory and auditing steps.

The thesis exploits machine learning techniques for performing the above analyses. In the next chapter, I provide an overview of the data collection and data processing steps, by describing the statistical tools I use in the presented case studies. In chapters three, four and five I attach the actual investigation as published in scientific journals and conferences. Finally, in chapter six, I present the thesis' results, and mention future work directions in the study of political machines.

2 Machine learning for political machines

2.1 Knowledge discovery in databases

Knowledge discovery in databases (KDD) [188] is an interdisciplinary research process that aims to generate new knowledge from the vast amount of unstructured data existing in today's world. The KDD process comprises following steps (see also figure 2.1):

1. A data curator collects data from one or more sources in a raw format. That means that data might be noisy, incomplete, or in a format that does not allow it to be further processed. Therefore, after data gathering, the curator structures it, repairs it and transforms it into a coherent and manageable format.

2. The now cleaned data can be stored in a data warehouse, from which they can be easily accessed. The data warehouse denotes both the existence of appropriate hardware and software infrastructure, which can ensure the data be deposited safely and its efficient and consistent retrieval.

3. The third step of the KDD process concerns data selection. A data scientist chooses a part of the data based on a research question or an exploratory task. Based on the nature of the posed question, the analytical skills of the scientist and hardware and software constraints, a subset of the available data is further processed and transformed into a format that allows their statistical processing.

4. Next, the scientist applies sets of statistical and machine learning algorithms (data mining) than can reveal information that was not visible before. The aim of the model application is the extraction and recognition of patterns, associations and statistical inferences based on the available data.

5. Finally, the scientist integrates the new information into narratives about the world. In contrast to the classical scientific method, KDD does not follow classic deductive reasoning [189]. Based on the available datasets, it either tries to create accurate predictions [190], or inductively generate knowledge about the world [191]. Therefore, after systematic interaction with and visualization of the model results, the scientist holds new subjective knowledge about the world [192]. Given the evaluation of the models' results, steps 3 to 5 can be repeated until the final predictions or built narratives satisfy the scientist's expectations.

In this thesis I concentrate on steps 2 to 5 of the KDD process. After collecting and formatting data, I use the process to achieve two purposes. First, I apply state of the art machine learning algorithms and statistical tools to create knowledge about the processes taking place in political machines. Second, I simulate the KDD process as it takes place in political machines, in order to understand how algorithmic applications impact political machines. In the following, I present the statistical and machine learning foundations used to achieve my purpose.

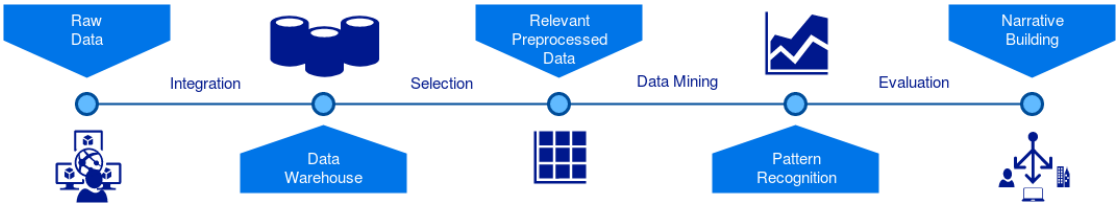

Figure 2.1: The Knowledge Discovery in Databases Process

2.2 Foundations of probability theory, statistical inference and machine learning

Understanding political machines is not a trivial task. Political interactions on social media comprise trajectories of millions of users, that need to be studied in a coherent and efficient way. Similarly, the application of data-intensive algorithms within services for decision making and for behavioral influence of people is based on a huge amount of input data. The investigation of algorithmic impact thus requires a framework that can detect regularities of interactions. In this thesis, I use probability as a framework that quantifies randomness and uncertainty in a principled formal way [193]. I then employ statistical tests and machine learning techniques in order to extract knowledge about the phenomena under investigation.

Statistics, probability and distributions

The KDD process involves the analysis of input data to extract implicit, previously unknown and potentially useful information [194]. This happens, because input data themselves already contain unstructured information about the world. To transform unstructured information into structured patterns a data scientist employs data mining, which itself relies on statistics, i.e. *mathematical procedures for organizing, summarizing, and interpreting information* [195].

The study of political machines aims to analyze political interactions of individuals and social groups across sociotechological ecosystems. To answer a specific question based on probability, a definition of a population P is necessary, i.e. the set of individuals and entities of interest according to the research question. Because of the complexity of phenomena, usually complete data on the population is not available. Instead, models use a sample S of individuals, which is ideally a representative subset of P. A variable X is a property that has a specific value for each specific individual or entity in S. Variables can be numerical, ordinal, or nominal. A numerical variable is for example the number of interactions a user has generated on an OSN. Ordinal data depicts properties grouped in hierarchies, such as the educational level of an individual. Nominal data consists of properties that contain purely qualitative information, e.g. the appearance of a word in a user comment. Usually there is a naturally occurring deviance between the sample and the actual population properties called a sampling error.

A data set is a collection of measurements about the properties of S, and each datum is an observation about S. Given a data set, a data scientist can employ descriptive statistics to summarize and simplify information. The scientist can also use inferential statistics, i.e. techniques to draw generalizations about population P based on sample S.

The simplest statistical technique to draw knowledge from a dataset is the presentation of existing sample properties in the form of a frequency distribution, which shows the number of observations associated with each unique value of a variable. When the distribution is normalized by the total number of observations, it is called relative frequency distribution. In political machines, the extraction and evaluation of frequency distribution can contribute towards understanding user behavior or algorithmic performance, as I show in chapter four.

The distributions obtained from observations are called *empirical distributions*. In order to evaluate them, a data scientist can compare them with formal *probability distributions*, drawing inferences on the data generating process (DGP). DGP signifies the ontic relations existing in the population from which the data were drawn from. A probability distribution is a formalization of a relative frequency distribution. Each possible value of a variable X, also called outcome ω, is assigned to a probability or likelihood of appearance. The formalization indicates how the total probability of 1 (appearance of any outcome) is distributed over all possible outcomes existing in a population. Depending the nature of variable X, the sample space Ω (set of all possible outcomes) is either countably finite or countably infinite. If it is finite, it can be modeled by a discrete probability distribution, otherwise by a continuous probability distribution. An example of a countably finite space is the sex of a person (male or female), while a countably infinite is the the time they spend on an online social network (OSN).

A discrete distribution function is defined as follows. Given is a random variable X and sample space Ω that contains only finitely possible amount of outcomes. A discrete distribution function of X is real-valued function f with domain ω that satisfies:

1. $f(\omega) \geq 0$, for all $\omega \in \Omega$
2. $\sum_{\omega \in \Omega} f(\omega) = 1$.

For any subset A of ω, the probability of A is defined as $P(A)$ given by

$$P(A) = \sum_{\omega \in A} f(\omega).$$

Similarly, given X and a sample space Ω that contains infinitely possible amount of outcomes, a continuous distribution is characterized by a density function f. f is a real-valued function that satisfies

$$P(a \leq X \leq b) \int_a^b f(x)dx$$

, where $a, b \in E$ and $E \subset R$ for which the density function makes sense.

One of the most important distributions in statistics is the normal or gaussian distribution, whose density function is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

μ denotes the random variable's expectation given by its weighted average and σ^2 the distribution's variance given by the expected value of the squared deviation from the mean of X. A normal distribution with μ and σ^2 is denoted with $N(\mu, \sigma^2)$. The normal distribution is useful because of the central limit theorem (CLT):

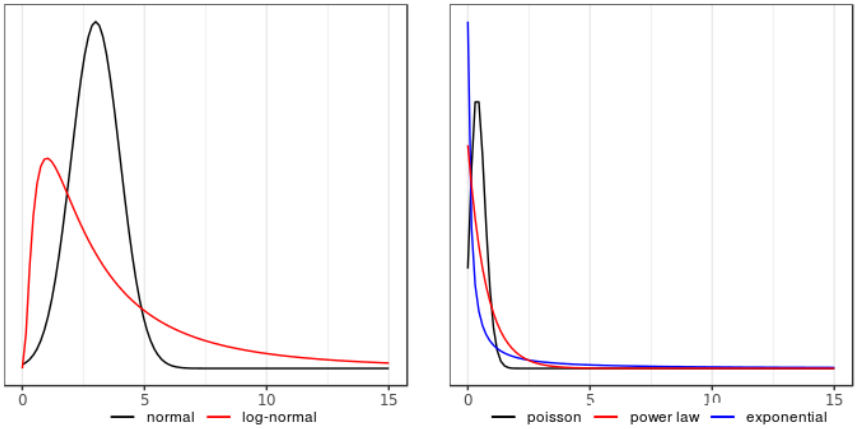

Figure 2.2: Left: comparison of the normal with the log-normal distribution. Right: Three fat-tailed distributions: exponential, power law, poisson

.

- Let $X_1, ..., X_n$ be a random sample of size n of a set of independent and identically distributed (i.i.d) random variables drawn from a distribution with μ and finite σ^2. The difference of the sample average S_n with the true average of the population μ will converge in distribution to a normal $N(0, \sigma^2)$ as n approaches infinity.

That holds even if the original variables themselves are not normally distributed, a property exploited in various statistical tests and metrics that will be presented later. The normal distribution is a symmetrical distribution, because densities left and right of the mean are mirrored. Nevertheless, activities on political machines often do not follow the normal distribution. As do many social phenomena [196], social behavior in political machines is often skewed, with the majority of individuals having the same behaviour on one side of the distributions scale, while the minority of behavior lies on the other. In many cases, the given skeweness becomes extreme, resulting in frequencies of activities that can be described by fat-tailed distributions. Social phenomena that can be described by fat-tailed distribution are of extreme interest, because single and rare outcomes can have a disproportional impact on a population. Fat-tailed distributions are applied to understand phenomena in various disciplines as economics, social network analysis and climate research [197, 198, 199]. In political machines fat-tailed distributions can be also useful. For example, their comparison with empirical user activities on social media provides insights to phenomena such as political communication, as I show in chapter four. In the following, I present four basic heavy-tailed distributions and their properties.

- **Log-normal distribution**
  A positive random variable X is log-normally distributed if the logarithm of X is normally distributed: $ln(X) \sim N(\mu, \sigma^2)$. The respective probability density function is given by

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

where μ, σ the respective location-scale parameters. As I show in chapter four, the user frequency of interacting with a social media platform can be described by a log-normal distribution.

- **Poisson distribution**
  A discrete random variable X is Poisson distributed with parameter $\lambda > 0$ for $x = 0, 1, 2, ...$ if its probability mass function is given by:

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!},$$

  where λ is equal to the expected value and variance of X. The Poisson distribution can be used to describe properties that have a large amount of possible outcomes, each of which are rare. An example could be the time interval between different users flagging a post on an OSN as misinformation [200].

- **Power law distribution**
  A continuous random variable X follows a power-law distribution if it has the probability density function

$$f(x|a, x_{min}) = Cx^{-q} \text{ for } x \geq x_{min},$$

  with

$$C = (a-1)x_{min}^{a-1},$$

  where $a > 1$ and x_min a cut-off parameter. Power law has multiple applications, but it is interesting for this thesis because general human activities on social networks can often be approximated by the distribution [201].

- **Exponential distribution**
  A continuous random variable X follows a power-law distribution if it has the probability density function

$$f(x| = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

  where λ a rate parameter. The exponential distribution can be used for modelling time intervals between the occurrence of a specific event such as users' incoming messages.

The above distributions are just a subset. Numerous other continuous and discrete distributions exist that can be useful for evaluating political machines. Given that, a data scientist should have formal criteria that help them decide what distribution is ideal for their data. For that, multiple tests exist, all of them based on the *likelihood function*.

- A parametrized family of probability density (mass) functions with observed outcomes x for a set of random variables X and distribution parameters θ can be described by a function $f(x|\theta)$. When θ is fixed, it is a probability density function expressing the probability of outcomes taking place. When x is fixed and θ variable, the function serves as a likelihood function. It illustrates how likely

it is that the data come from a DGP following a distribution with the specific parameters θ. This is formalized by the equation

$$L(\theta|x) = f(x|\theta).$$

For a specific sample and a given distribution, someone can calculate the *maximum likelihood* of the distribution, i.e. find the parameters θ that maximize $L(\theta|x)$. The likelihood is used to compare the goodness of fit of different distributions on the same data by performing statistical tests, which are going to be introduced next.

Regardless of whether a sample can be consistently described by a probability distribution, there is always the case that it might include *outliers*. Outliers are observations, whose position in the variable space differs significantly from the rest. [202]. They might be a result of sampling errors, or of a different DGP, or even that the general DGP includes the generation of rare events that deviate significantly from the other observations. Given each case, researchers have developed different techniques for tracing outliers. For political machines, outlier detection can provide significant information about the interactions taking place. For example, automated accounts that spread misinformation on social media platforms can be treated as outliers, and understanding their behaviour can result in important knowledge on the nature and impact of political machines.

Statistical hypothesis testing

Although the KDD process creates the conditions for inductive knowledge creation based on datasets, the collected data sometimes is just a minor part of the total DGP. In order to investigate assumptions about the relationship of the samples with theoretical properties of the phenomena studied, researchers formulate statistical hypotheses and explore them by applying of statistical tests. For example, because of privacy regulations a data scientist might hold only a part of user activities on a social media platform. In order to infer the total platform population behaviour, a data scientist makes certain assumptions about the correspondence between the sample and the total population distribution. According to them, they can draw a formal inference about the population properties. This is done by using statistical hypothesis testing.

The data scientist formulates a hypothesis about a population, and investigates the plausibility of that hypothesis based on the available sample. The hypothesis is usually called *null hypothesis* (h_0), and presupposes the absence of a specific property, relationship, or difference on a sample and its parameters. For example, a null hypothesis might be that there is no statistical relation between the time a user spends on a platform and the number of posts they generate. To statistically investigate the hypothesis, the data scientist calculates a test statistic. The test statistic is a random variable calculated from the sample distribution and is able to quantify the agreement between the null hypothesis and the sample. The agreement is given by a probability value called p-value, which gives how likely it is that the sample obeys the null hypothesis. By setting a probability threshold, also called alpha level, and comparing it with the p-value, the data scientist rejects or not the null hypothesis. If they decide to reject the null, then they assume that the alternative hypothesis H_1 holds, i.e. that there is a property, relationship, or difference on a sample and its parameters.

Given the data scientist's decision on the alpha-level value, there is always the possibility of a false inference. There are two types of errors that a scientist can make in hypothesis testing, Type I and Type II. A Type I error denotes that a scientist rejects a null hypothesis that is actually true. In a typical research case, a Type I error means that the researcher concludes to the existence of a property, association, or difference in a sample when in fact there is none. A Type II error denotes that a researcher fails to reject a null hypothesis that is in fact false. That means that the scientist concludes to the absence of a property, association, or difference in a sample, although they actually exist in the true population.

Depending on the hypothesis under investigation and the nature of available sample data, different test statistics can be calculated. Tests might be one-sample, two-sample or paired. A One-sample test is used when a hypothesis about a population is investigated given a sample. A Two-sample test is useful when comparing the similarity or difference of properties existing in two different samples. A paired test is a two-sampled test that can be used when there is a fixed correspondence between observations among the two samples. Tests might also be parametric or non-parametric. A test is parametric when a data scientist assumes that the sample follows a specific statistical distribution. Therefore, several conditions should be met in order for the test statistic to be reliable. This does not hold for non-parametric tests, where a data scientist makes no assumptions about properties of the samples under investigation. In this thesis I employ one-sample, two-sample, parametric and non-parametric tests to investigate properties of user behavior on OSNs (chapter four), as well as to locate differences on ADM models' predictions for different social groups (chapter five). In the following, I present prominent statistical tests to analyze political machines, their assumptions and possible applications.

Parametric tests

- **t-test**

  The t-test encompasses a family of statistical tests that hypothesize that the test statistic follows a Student's t-distribution under the null hypothesis. Common t-tests are the one-sample t-test, the two-sample t-test and the paired t-test.

  The one-sample t-test is used to compare the mean of a population to a specified theoretical mean as formulated in the null hypothesis μ. Let X represent a random variable with n observations, mean m and standard deviation s. The comparison of the observed mean m of the population to a theoretical value μ is performed by the statistic:

  $$t = \frac{m - \mu}{s/\sqrt{n}}$$

  The t-test holds when the random variable consist of continuous, random, identically and normally distributed observations.

  The two-sample t-statistic is used to compare the means of two unrelated groups of samples. For example, someone can compare the average popularity of posts about two different political topics. Given are samples A and B with mean and size m_A, n_A and m_B, n_B respectively. The t test statistic is used to test whether the sample means are different can be calculated by:

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

, where the variance is given by

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}.$$

The two-sample t-test assumes that the samples consist of continuous, random, identically and normally distributed observations of equal variance. It also assumes that the two samples are independent to each other. In the case of unequal variance, an adaptation of the test exists, called *Welsch's test*.

The paired t-test is used to compare the means of two unrelated groups of samples in cases where the two samples are not independent. It can be performed when each observation on one sample is related (paired) to an observation in the other sample. For example, someone can compare how a specific user group interacts on two different social media platforms. Let d represent the differences between all pairs in the two samples. The average of the difference d is compared to 0. If there is any significant difference between the two pairs of samples, then the mean of d is expected to be far from 0. The test statistic value is calculated as

$$t = \frac{m}{s/\sqrt{n}}.$$

The paired t-test assumes that the samples consist of continuous, random, identically and normally distributed observations.

- **Chi-squared test**

  Chi-squared tests are a family of statistical hypothesis tests that assume that the test statistic is a chi-squared distribution when the null hypothesis is true. Common Chi-squared tests are the Chi-squared test for independence, the Chi-squared test for variance and the Chi-squared test for goodness of fit. The statistic is calculated by the general form

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

  , where O_i is the number of observations of type i and E_i the expected count of the specific type of observations. Depending on the type of chi-squared test, the above equation is generalized or adapted accordingly.

  Chi-squared tests use similar calculations and the same probability distribution to answer different questions. For example, Chi-squared tests for variance are used to determine whether a normal population has a specific variance. Chi-squared tests of independence are used for testing whether two categorical variables are associated or not. Chi-squared goodness of fit tests are used to determine the adequacy of a probability distribution to describe sampled data.

- **F-test**

  F-tests are a family of statistical hypothesis tests that assume that the test statistic follows an F distribution under the null hypothesis. They are commonly used to decide if collections of data, both samples or model predictions, differ in terms of variability. The general F-statistic for model comparison is given by

$$F = \frac{\text{explained variance}}{\text{unexplained variance}}.$$

  When comparing the expected values of a quantitative variable within several predefined groups, it is defined by

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}.$$

  F-tests are useful when comparing the predictive ability of linear regression models, or when investigating if there is a difference between properties of multiple groups.

Non-parametric tests

- **Kolmogorov–Smirnov test**

  The Kolmogorov-Smirnov test investigates the similarity of the empirical sample distribution to a reference probability distribution (one-sample K–S test), or compares the empirical distribution of two samples (two-sample K–S test). The statistic for the one-sample test for n identically and equally distributed observations is given by

$$K_n = \sup_x |(F_n - F)(x)|$$

  , where $\sup_x$ the supremum of the set of distances resulting fromt the empirical cumulative distribution function $F_n(x)$ and the reference distribution $F(x)$.

  Similarly, for the two-sample test the statistic for two sets of n and m identically and equally distributed observations is given by

$$K_{n,m} = \sup_x |(F_n - F_m)(x)|$$

  , where $F_n(x)$ and $F_m(x)$ are the empirical cumulative distribution function for each sample. KS-tests can be used both for matching empirical distributions of user activities with formal theoretical ones, as well as to detect behavioral similarities and differences between users of different groups.

- **The Kruskal–Wallis H test**

  The Kruskal–Wallis H test is a method for investigating if samples originate from the same distribution. It can be applied on two or more independent samples of equal or different sample sizes. The statistic is given by the equation

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

, where n_i is the number of observations in sample i, r_{ij} is the rank (among all observations) of observation j from sample i, N is the total number of observations across all samples, $\bar{r}$ is the average rank of all observations in sample i, and $\bar{r}$ is the average of all the r_{ij}. The Kruskal-Wallis H test can be used in cases where the assumptions for parametric tests such as the t-test or the F-test are not met. For example, to compare the popularity of micro-targeted ads of more than two parties, which are not normally distributed.

The above statistical tests are just a subset of the ones existing for investigating hypotheses. Depending a specific research question and the nature of available data, a research scientist should elect the appropriate test for their analysis.

Machine learning

Machine learning lies in the epicenter of this dissertation. Not only do social media platforms and ADM systems deploy machine learning algorithms to provide services and influence human behavior, but so does machine learning emerge as a valuable tool for analyzing and understanding the digital traces generated at the human-algorithm interactions. In this thesis, both aspects of machine learning are taken into consideration. In the following, I provide a brief overview of machine learning methods and their properties.

Definition: Machine learning

Machine learning is concerned with algorithms solving optimization problems based on a dataset of observations. By minimizing a cost function, models learn to make inferences related to the data provided.

Categories of machine learning models

Machine learning models can be supervised, unsupervised, or semi-supervised.

Supervised algorithms take as input and output N observations containing a set of X_M features and Y_K labels respectively, with $M, K \geq 1$. They try to capture the real relationship f between X_M and Y_K by minimizing a cost function $L(Y_K, \hat{f}(X_M))$ and hence learning the approximation $\hat{f}$. The ability of a machine learning model to approximate f is dependent on the input data and the nature of the algorithm used. A supervised learning model can perform two tasks: regression and classification, depending if an output feature is numerical or categorical/ordinal respectively. Widely used supervised learning algorithms are linear and logistic regression [203, 204], random forests [205], support vector machines [206] and neural networks [207].

Unsupervised learning models take as input N observations containing a set of features X_M and optimize a cost function $L(\hat{g}(X_M))$. In contrary to supervised learning, there is no real function f mapping a set of input and output features that tries to be approximated by the model. An unsupervised learning algorithm learns a function $\hat{g}$, which gives a mathematical representation of each observation that was not explicit in the

initial dataset. Unsupervised learning algorithms might be used among others for clustering, anomaly detection, and latent variable modeling. Clustering techniques group observations into a number of clusters based on their similarity. Important clustering techniques include k-means [208] and hierarchical clustering [209]. Anomaly detection algorithms try to locate observations that are dissimilar to the rest, i.e. detect statistical outliers. An exemplary algorithm is the local outlier factor, while models used for clustering or latent variable modeling are also able to detect data anomalies. Latent variable modeling comprises a set of algorithms that try to project input variables to new mathematical spaces, reducing noise in the data and uncovering properties that were not visible before. Factorization models such as singular value decomposition [210], or non-negative matrix factorization [211], or other transformations such as principal component analysis are widely used for latent variable modeling [212].

At the intersection of supervised and unsupervised models, researchers have developed methodologies that perform *semi-supervised learning*. Semi-supervised learning algorithms draw inferences under partial supervision. That is, a data scientist feeds the model with a set of features X_M containing N observations, while the corresponding set of labels Y_K contains $L < N$ observations [213]. The models try to learn the relationship f between X_M and Y_M by also exploiting information existing in the unlabeled data. Prominent unsupervised learning architectures exist in model classes such as generative models (e.g. semi-supervised GANs [214]), graph-based methods (e.g. label propagation [215]) and semi-supervised SVMs (e.g. transductive support vector machines [216]).

Regardless whether a model is supervised, unsupervised or semi-supervised, it might be *generative* or *discriminative*. The generative and discriminative approaches characterize models in terms of their statistical assumptions about the DGP [217]. Discriminative models make no assumptions about the DGP. In a supervised learning setting, they try to calculate f by the information existing solely on X_M and Y_K by inferring the empirical probability $p(Y_j|X_j)$. Similarly, unsupervised discriminative models try to infer $\hat{g}$ as dictated by the given cost function. Typical discriminative models are linear regression, support vector machines, and non-negative matrix factorization. Contrary to discriminative models, generative models make explicit assumptions about the DGP that guide the exploitation of available data. In supervised generative models a scientist calculates f by making assumptions about the joint probability distribution $P(X, Y)$, then calculates $p(Y|X)$. The same applies for unsupervised generative models, where the models calculate g by assuming properties of $p(X)$. Prominent generative models are the naive Bayes classifier [218], hidden Markov Models [219] and the Latent Dirichlet Allocation [220].

Model development and selection

In order to understand properties of political machines or to deploy machine learning models in political machines, a data scientist has to find the ideal machine learning algorithm to fulfill their scope. To do so, they have to follow a well structured model development and selection process that differs between supervised and unsupervised models.

For supervised learning, a data scientist has to follow a four step process:

1. First, the scientist divides the available data in two subsets of randomly selected training and test observations. To avoid the problem of underfitting, that is the

development of a model that cannot capture the true DGP, the number of observations in the training set is much higher that of the test set.

2. In the learning stage, the training set is split into two random sets of training and validation observations. The scientist trains a model on the training set in order to fit the data in an optimal way. That is achieved by minimizing the model's cost function on the training data.

3. In the validation stage, the scientist validates the fitted model on the validation set given a performance metric. That could be the model's cumulative loss, its accuracy, or other metrics such as Recall or F-score. Because a model usually takes multiple hyperparameters as input, the scientist performs sensitivity analysis by repeating steps 2 and 3 until they reach an optimal model.

4. Finally, the scientist evaluates the performance of the chosen model on the test set. If the model performs satisfactorily on new data, then the development and selection phase is successfully concluded. Otherwise, the scientist has to repeat the process, by adapting or completely changing the model's architecture until the desired result is achieved.

For unsupervised learning, a data scientist performs following steps:

1. In the learning stage, the scientist fits multiple models on the dataset by varying the chosen architecture's hyperparameters.

2. In the selection stage, the scientist performs an external or internal evaluation and chooses the model that performs best given a performance metric.

   - The external evaluation takes place when the scientist knows the true labels in a dataset. They compare the predicted labels with the actual labels, usually by applying an information based criterion. Mostly, the well-known Kullback-leibler (KL) divergence is applied [221], which calculates the similarity between the predicted empirical probability distribution $P(x)$ of the labels to the actual one $Q(x)$, by the function

     $$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right).$$

     The KL-divergence might be transformed and applied in other forms, such as the mutual information index [222]. Other metrics can also be used for external evaluation such as the mean Square error [223] or the models' purity [224].

   - The internal evaluation takes place when the scientist does not have any available labels about the observations. Then, they can apply an information based metric to quantify the within-cluster and between-cluster similarities or to use other perplexity based methods [225], which are again related to the KL-divergence. Another alternative is to evaluate the model's likelihood by applying either the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) [226]. They are calculated by following equations:

$$AIC = -2l(a_i) + 2k$$

$$BIC = -2l(a_i) + k + log(N)$$

, where $l(a_i)$ the likelihood of the data inserted to model i, k the number of model parameters and N the sample size.

In this thesis, I apply information based criteria to select optimal models for natural language processing (chapters three, four).

2.3 Machine learning for Natural Language Processing

The objective of this thesis is to understand political dimensions of algorithmic influence across political machines. Since a large amount of interactions on social media platforms are text related and many ADM applications take text as input, I provide an overview of natural language processing and an introduction into three machine learning algorithms; Topic Models, Word Embeddings, and Neural Networks; all three are architectures that I exploit later in the thesis to answer the research question posed in chapter one.

Foundations of natural language processing

The digital traces generated at the interactions taking place on political machines largely involve the generation or processing of textual data. Textual data is a representation of lingual propositions, in which concepts ordered by a specific syntax are able to transmit a specific semantic content. This occurs because they belong to a *natural language* used in a given society. Any piece of text data, a letter, word, sentence or even a book, carries with it a specific meaning, which is able not only to provide a description of the world, but also to influence the humans that read it and the algorithms that take it as input. Therefore, it is really important to analyze and understand properties existing in text in order to understand properties of political machines. To that end, I use machine learning methods for *Natural Language Processing*.

Definition: natural language processing

Natural language processing (NLP) is an interdisciplinary field at the intersection of Computer Science, Artificial Intelligence and Linguistics, which aims to structure, understand, generate and make sense of textual data in a valuable way [227].

NLP consists of multiple sub-fields such as machine translation, speech recognition, question-answering, contextual recognition, text summarization, and text categorization. Most NLP techniques rely on machine learning to derive meaning from human languages. My thesis focuses on machine learning for *text mining*. Text mining involves NLP techniques to extract valuable information and insights from textual data [228]. The main operations in text mining are the following:

- Classification: Identifying the category to which a textual observation belongs.
- Clustering: Grouping text data to clusters based on their similarity.
- Summarization: Summing up the content of a textual chunk.

- Sentiment analysis: Extracting affective states and subjective information from text.
- Entity recognition: Information extraction by locating specific types of text, e.g. person names.
- Similarity analysis: Quantifying the similarity between different text chunks.

Most of the above operations are a constituent part of the dissertation. First, in chapter three I use clustering and similarity analysis to uncover properties of data-driven microtargeting. Next, in chapter four, I apply entity recognition, text clustering and similarity analysis algorithms to analyze political communication on social media. Finally, in chapter five I employ classification and similarity analysis algorithms to illustrate bias in NLP models and further ADM systems.

Preliminaries

The application of machine learning on natural language requires the formatting and appropriate preprocessing of the available text. The totality of available text in a specific case-study is called a *coprus*. Depending scientist's goal, they segregate the corpus into smaller chunks of arbitrary size called *documents*. For example, in a book, someone can represent each chapter, each paragraph or each sentence as a separate document.

The segregation of a document into smaller parts is performed by *tokenization*. The scientist splits the text in smaller units, in order to create models' input features. Units might be words, syllables, or characters depending the model architecture. They might also be a contiguous sequence of n units, called n-grams. State of the art algorithms as Sentence Piece [229] or Byte Pair Encoding [230] split the documents into subword units depending on their frequencies of appearance, as that contributes to the more efficient machine learning model training.

To extract additional information from text to be added to the models' inputs, researchers often perform *part-of-speech (POS) tagging* and *chunking*. POS Tagging denotes the labeling of each word in a document depending its grammatical and syntactic properties (e.g. if a word is a noun, if it is on singular or plural). Chunking (also called shallow parsing) extends POS Tagging by labeling each word based on its relational position in a sentence. In this way words are embedded into hierarchies of grammatical meanings, information that might be useful for specific NLP tasks.

Because text is highly complex, there are specific techniques that help the data scientist simplify the available content. Two techniques that serve a similar purpose but follow different processes are *stemming* and *lemmatization*. For specific tasks, only the general meaning of words might be important and not their syntactic and grammatical properties. Thus, words that have the same root can be brought into the same form, so that the model treats them equally. To achieve that, a data scientist can cut the endings of words and keep only their significant root, a process called stemming. Another way is to transform words to their general base, a process called lemmatization. Both can reduce the size of corpora and provide better results in algorithms such as topic models.

Another way to reduce text complexity and preserve only important content can be done by corpus *pruning* and *stopwords* removal. In pruning, the data scientist removes terms that appear too frequently or very rarely in the corpus, such as dominant articles or misspelled words. In this way, they keep only information that contain significant value

for the model. The same applies in stopwords removal, where the data scientist generates a list of words that might exist frequently in the corpus, but are of no informational value to them. Both techniques result in corpus size and complexity reductions.

In cases where the syntactic and grammatical relations of the texts are not important, a data scientist usually adopts a *bag-of-words* approach. The approach treats each document as a multi-set of words, ignoring their order and grammar, but keeping their multiplicity. These multi-sets can be represented in linear algebraic terms as *document-term-matrices*. A document-term matrix is a matrix that has N rows and K columns representing the N documents and K words, or more generally terms, existing in a corpus. For a specific word-document combination, the matrix gives the number of times the word appears in the specific document.

Because the frequency of a word appearing in a document does not necessarily correspond to its importance, a data scientist can apply a *term frequency-inverse document frequency (tf-idf)* transformation to the document term matrix, which weights how often a word appears and in the number of documents it appears so as to infer its actual importance. tf-idf is mathematically given by

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$$

, with

$$\text{tf}(t, d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

and

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}.$$

The term-frequency term tf counts the number of times that term t appears in document d by the frequency $f_{t,d}$ divided by the maximum term appearance among the rest of the terms t' in the document. The inverse document frequency idf measures how much information a term provides, i.e. if its overall common or rare in the corpus. That is given by taking the logarithm of division between the total number of documents in a corpus N and the number of documents that term t appears in.

For machine learning algorithms that process one word at a time, the bag-of words-approach becomes infeasible. In such cases, scientists can linear-algebraically manipulate words by *one-hot encoding* them. One-hot encoded representations are vectors that have length equal to the vocabulary V of a corpus. For each word w_i its one-hot encoded vector has an i_{th} element of value one, while the rest of the elements are zero. E.g. For the corpus C = {Viva La Revolution}, there will exist three one-hot vectors mapped as following:

$$Viva : [1, 0, 0]$$

$$La : [0, 1, 0]$$

$$Revolution : [0, 0.1]$$

The above techniques are usually combined for transforming raw unstructured text data into structured and meaningful features for machine learning. This efficient preprocessing is necessary for creating machine learning models of high quality. Next, given the preprocessed data, I present three model types used extensively in my thesis: topic models, word embeddings, and neural networks.

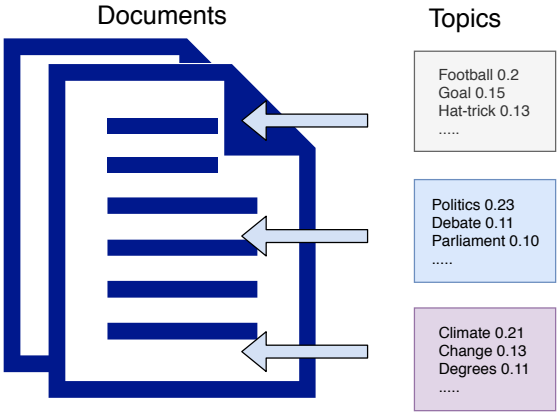

Figure 2.3: Topic models assign a probability to each word belonging to a topic. The words with the highest probabilities characterize a topic semantically. These topics are latent structures existing in the documents of a corpus.

Topic models

In political machines, especially on OSN's, users generate huge amounts of text in their interactions. To uncover properties of political machines, it is valuable to understand and analyse available text data. Because of the huge size of available corpora, it is infeasible for a scientist to qualitatively evaluate them. Thus, it is necessary to find means for understanding the content in fast and efficient ways. Topic models are a set of statistical and mathematical modeling techniques that contribute towards that end.

Definition: Topic models

Topic models encompass a set of algorithms that aim to annotate and explain large collections of documents with thematic information. They analyze words of an original theme, discover themes existing within them, how themes are interconnected, and how the change over time [231]. Each theme or topic in a corpus is represented by a set of words. The words' relative importance for a topic is given by a probability $P(w|k)$, where w is an arbitrary word and k a specific topic (figure 2.3). Similarly, topic models also provide the relative importance of a topic for a document d by the probability $P(k|d)$. The empirical distributions of topics over documents and words over topics are integrated in two document-topic and topic-word matrices DT and TW. The algorithmic task of topic models is taking a document-word matrix DW as input and by the applying an algorithm f calculating DT and TW respectively.

Prominent topic models

Researchers have developed multiple techniques for topic modeling (e.g. see [232, 233, 234, 220]). Here I present two widely used models: *Non Negative Matrix Factorization* (NNMF) and *Latent Dirichlet Allocation* (LDA).

- **NNMF**

  NNMF is a discriminative topic model that tries to approximate the matrices DT and TW as a product of DW : $DW \approx DT * TW$. The most common way to solve

that problem is by minimizing the frobenius norm of the squared errors between the product and the actual matrix DT. That is given by

$$F(\mathbf{DT}, \mathbf{TW}) = \|\mathbf{DW} - \mathbf{DT} * \mathbf{TW}\|_F^2.$$

The problem is solvable by applying a multiplicative update rule [235], gradient descent optimization [236] or other techniques [237]. Besides NLP, further NNMF fields of application are astronomy, computer vision, and bioinformatics [238, 239, 240].

- **LDA**

  LDA is a generative statistical topic model. It assumes a specific latent probabilistic mechanism in the generation of the themes in text and calculates the probability matrices DT and TW. It assumes the probability distributions of topics over words β_k , of documents over topics θ_d and predicts the probability that a specific word in a specific document will belong to a specific topic. The Bayesian admixture is described by the following probability distributions:

$$\begin{aligned} \theta_d &\sim Dir_K(a) \\ \beta_k &\sim Dir_V(\eta) \\ z_w &\sim Multinom_K(\theta_d) \\ w \mid z_w &\sim Multinom_V(\beta_k) \end{aligned}$$

  , where V is the number of unique words in the corpus, and α and η are Dirichlet parameters. Multinomial distribution z_w gives the probability that a topic will be assigned to a word, given the distribution of topics over documents. Finally, multinomial distribution $w \mid z_w$ gives the probability that the model generates a specific word in a specific document given a topic. Assuming the above generative mechanism, someone can optimize admixtures' likelihood on the data and calculate empirical probabilities $P(w|k)$ and $P(k|w)$ respectively. The model outputs them in the form of the matrices DT and TW. The optimization problem is usually solved by a collapsed gibbs sampling method [241], while other alternatives also exist [242]. LDA is probably the most famous topic modeling technique, and I use it in chapter three to uncover the topics of political interest of German social media users.

Model selection

Most topic models come with a drawback: they cannot optimize the number of topics K. Therefore, a data scientist should perform sensitivity analysis over the number of topics and compare models based on a performance metric. For that purpose, researchers have developed multiple techniques to achieve that [243, 244, 245]. The most general method for optimizing the topic number is by finding the model that has the minimum perplexity. The perplexity of a model is given by the equation

$$\text{perplexity}(M) = \exp\left\{-\frac{\mathcal{L}(\boldsymbol{w})}{\text{V}}\right\},$$

where

$$\mathcal{L}(\boldsymbol{w}) = \sum_{d} \log p(\boldsymbol{w}_d|\boldsymbol{\kappa}).$$

M is the model, V is the number of words in a corpus and $L(w)$ is the log-likelihood of the given documents over the topics, given the empirical probabilities of each document w_d over the topics κ. The problem is intractable and numerous solutions have been proposed to approximate the actual value [246].

In chapter three, I apply the method proposed by Deveaud et al. to optimize the topic number [247]. The method proposes to calculate the Jensen-Shannon divergence between topics for multiple topic models through the equation:

$$D(k_i, k_j) = \frac{1}{2}\sum_{w=1}^{V} \beta_{i,w} log(\frac{\beta_{i,w}}{\beta_{j,w}}) + \frac{1}{2}\sum_{w=1}^{V} \beta_{j,w} log(\frac{\beta_{j,w}}{\beta_{i,w}})$$

, where i, j are two different topics in a model and $\beta_{i,w}, \beta_{j,w}$ the probability density values of the distribution β_k for a word w in the corpus V and each topic respectively. Then it selects the model that maximizes the sum of the Jensen-Shannon divergence for all topic combinations by using the formula:

$$K_{opt} = argmax \frac{1}{K(K-1)} \sum_{k_i,k_j=1}^{K} D(k_i, k_j).$$

Word embeddings

The second technique that I exploit to understand properties of political machines are word embeddings.

Definition: Word Embeddings

Word embeddings is the name for a set of models that map words or documents to vectors of real numbers. These vectors can either be used for understanding properties of the corpus they were trained on, or to improve the predictive and generative abilities of further machine learning architectures.

Word embeddings are successful because they exploit a property that neither the document-term matrix nor one-hot encoding are able to do: By projecting lingual concepts to mathematical spaces based on their co-occurrence, they capture semantic properties of the initial corpus. Therefore, many NLP architectures take words as input in the form of word embeddings, and not in one-hot encoding form, in order to improve their inferences. These embeddings might be pretrained [248], or their training might be part of the model architecture itself [249]. Researchers have found that word embeddings are able to model semantic, syntactic, grammatical and contextual properties of text (figure 2.4), while they avoid the issue of sparsity that accompanies document-term matrices and word embeddings. Because of that, multiple techniques for word embeddings have been developed.

Prominent models

I present two prominent architectures that can be used to train word-embeddings: word2vec [248] and GloVe [250].

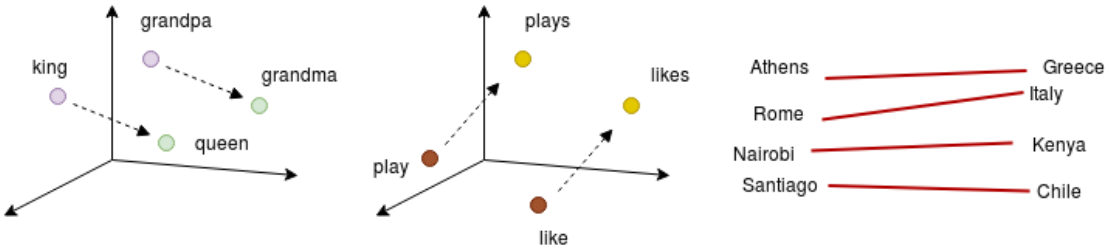

Figure 2.4: Word embeddings preserve semantic, syntactic and grammatical properties.

- **Word2vec**

  Word2vec encompasses two generalized logistic regression architectures that either try to predict the context words based on a target word (skip-gram), or to predict a word based on its context (CBOW). Both architectures calculate the values of the contextual and target word embeddings by optimizing the respective cost function. Given the vocabulary of a corpus, the word indexed as i is represented by vector $\mathbf{v}_i \in \mathbb{R}^d$ when it is a target word and by vector $\mathbf{u}_i \in \mathbb{R}^d$ when it is a context word. For all target and context words w_c and w_o, the skip-gram model predicts the conditional probability:

$$P(w_o \mid w_c) = \frac{\exp(\mathbf{u}_o^\top \mathbf{v}_c)}{\sum_{i \in \mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)}.$$

  The model is optimized by minimizing the cost function:

$$L = -\sum_{t=1}^{T} \sum_{-m \le j \le m,\ j \neq 0} \log P(w^{(t+j)} \mid w^{(t)}).$$

  , where T a text sequence, $w^{(t)}$ the word at time-step t and m the context window. Similarly, the CBOW calculates the probability that target word w_c appears, given context words $w_{o_1}, ..., w_{o_{2m}}$. That is given by the equation

$$P(w_c \mid w_{o_1}, \ldots, w_{o_{2m}}) = \frac{\exp\left(\frac{1}{2m} \mathbf{u}_c^\top (\mathbf{v}_{o_1} + \ldots + \mathbf{v}_{o_{2m}})\right)}{\sum_{i \in \mathcal{V}} \exp\left(\frac{1}{2m} \mathbf{u}_i^\top (\mathbf{v}_{o_1} + \ldots + \mathbf{v}_{o_{2m}})\right)}$$

  , which is optimized by minimizing the cost function:

$$L = -\sum_{t=1}^{T} \log P(w^{(t)} \mid w^{(t-m)}, \ldots, w^{(t-1)}, w^{(t+1)}, \ldots, w^{(t+m)}).$$

- **GloVe**

  Another widely used technique for word embeddings generation is *GloVe*. GloVe is a bilinear log regression model that calculates word vectors based on the co-occurence frequencies of the words in the dataset. This is done by optimizing the cost function:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j + logX_{ij})^2,$$

where V is the vocabulary, i and j are two words, w_i is the word vector of word i, $\tilde{w}_j$ is the context vector of word j, b_i and $\tilde{b}_j$ two added biases, and X_{ij} is the co-occurrence number of the words for a selected window. $f(x) = (x/x_{max})^a$ if x is lower than an elected x_{max}, otherwise $f(x) = 1$, with a being a hyper-parameter. In the thesis I apply GloVe to illustrate how data biases can result to biased algorithmic inferences (chapter five).

In resent years, researchers have developed more complex word embeddings architectures. For example Peters et. developed word embeddings associations based on bidirectional LSTM neural networks [251]. Similarly, general purpose machine learning architectures (e.g. [249, 252, 253]) include contextual word embeddings in their structure, which are document specific can be used in similar purposes as standard word embeddings.

Model selection

Word embeddings as machine learning architectures include the optimization of parameters based on a set of multiple hyperparameters. For example, a data scientist should elect the context window size, the length of the embeddings' vectors, how to split the corpus into documents and to set special hyperparameters depending an embeddings' model, in order to create specific vector representations of words. To that end, there are multiple techniques they can employ to decide the optimal model parameters. The standard probabilistic technique for electing the appropriate hyperparameters is through perplexity optimization. Similar to topic models, the scientist can calculate the likelihood of the generated vectors on a dataset and decide which set of vectors maximize it. Other techniques used to evaluate word embeddings include word similarity, word analogy, concept categorization, and a comparison to manually constructed word embeddings [254, 255]. For each of the above methodologies, scientists have developed various benchmark tests by which they investigate the efficiency of word embeddings [256, 254, 255, 257, 258].

Neural networks

The analysis of contents in political machines often includes the use of machine learning techniques for regression or classification. For example, a scientist might want to develop models that classify user contents on OSNs such as hate speech, or to predict user popularity based on social graph and textual features. The most straightforward models for prediction for regression and classification are linear and logistic regression respectively. Both can be depicted by graphs, as figure 2.5 illustrates. The models multiply inputs by a respective parameter and then calculate their summation. The sum is then transformed by a function $f(x)$ in order to generate prediction $\hat{y}$. For linear regression function f corresponds to the identity function $f(x) = x$, while for logistic regression f corresponds to the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$.

However, both models usually underfit datasets on complex phenomena because of their linear assumptions and the curse of dimensionality. To overcome these issues, a data

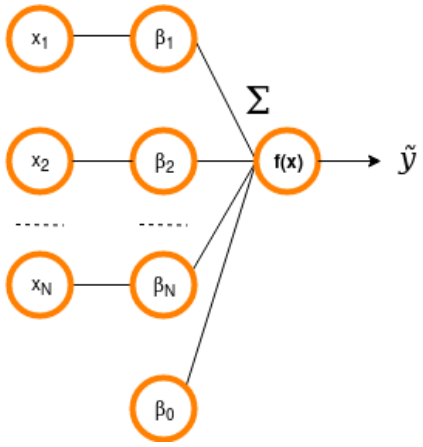

Figure 2.5: A graphical depiction of linear and logistic regression. Function f corresponds to the identity function and to the sigmoid function in the case of linear and logistic models respectively.

scientist can generalize the above graphs by adding more complex functions, connections, and parameters. These generalized architectures are called *neural networks*.

Definition: Neural networks

Neural networks are machine learning models, whose mathematical architecture can be represented by graphs. These graphs contain layers of nodes (or neurons), which are coupled to each other by weighted connections, usually referred to as *weights*. At each node, incoming inputs are always summed and transformed by an activation function f.

The simplest form of neural network is the multi-layered perceptron (MLP), illustrated in figure 2.6. The model consists of an *input layer* of $N features$, and an *output layer* of m nodes. All layers of an arbitrary number of nodes in between are the *hidden layers* of the model. Each node takes the sum of the previous layers' weighted output as input and transforms it by an activation function. It then passes it to the next layer through the weights. This process can be described by the functions

$$Z_i^l = w_i^T * a^{l-1} + b_i$$

and

$$a_i^l = f^l(Z_i^l)$$

, where Z_i^l is the summed input of node i at layer l, w_i is the matrix containing the weights of the layer, a^{l-1} is the matrix containing the output values of the previous layer and b_i is a bias term. The summed input output Z_i^l is transformed by an activation function f to the output a_i^l.

As any machine learning model, neural networks are optimized given a cost function. Typical cost functions for regression and classification is the mean squared error and the cross-entropy cost given by the following equations:

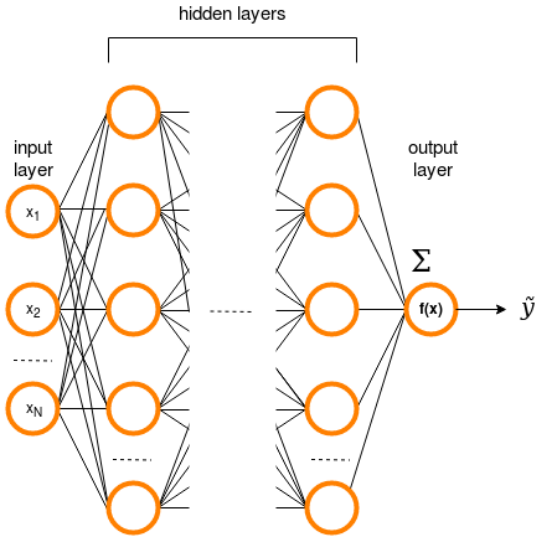

Figure 2.6: Graphical depiction of the multi-layered perceptron. It consists of an input, an output, and an arbitrary number of hidden layers.

$$\text{MSE: } L = \sum_i (Y_i - \hat{Y}_i)^2$$

$$\text{Cross-entropy: } L = -\sum_i [Y_i \ln \hat{Y}_i + (1 - Y) \ln (1 - \hat{Y}_i)].$$

Because of architectural complexity, no analytical solution exists for calculating the models' weights. To that end, the optimization of the model follows an iterative process of *forward propagation* and *backward propagation*. In forward propagation the scientist computes a prediction based on the current values of the weights. The prediction is compared to the true labels in the dataset and the model computes the actual loss. In backward propagation, the scientist applies a gradient descent optimization step for each weight in the architecture. They calculate the gradient of the loss function in respect to each weight and then alter each weight's value to the direction of the loss functions' minimum. The process of forward and backward propagation is repeated until the model provides the best predictions on the test set.

A data scientist can generalize the architecture of the model and adapt it to their needs. They can alter and introduce more complex weighted connections, decide if the neural network architecture will be supervised or unsupervised, change the cost or activation functions. For example, typical activation functions that scientists use, besides the sigmoid, are the rectangular linear, the hyperbolic tangent, or the leaky rectangular linear.

Specialized architectures

Because of lingual complexity, the deployment of simple neural architectures usually does not suffice to create adequate inferences on textual data. Towards that end, researchers have developed more complex neural sequences that are able to capture properties of language and thus provide models with improved accuracy. I present three important

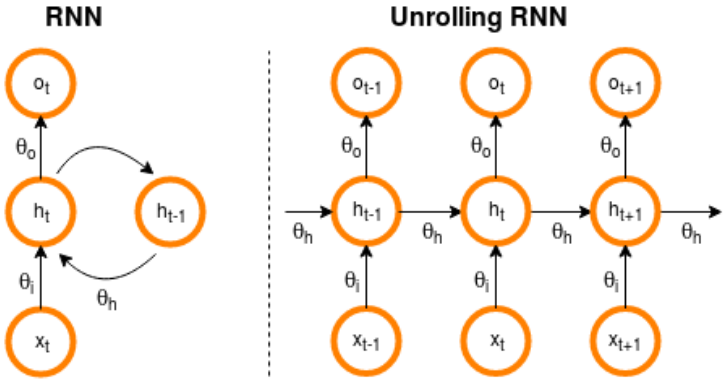

Figure 2.7: Left: The recurrent neural neuron takes as input at time t the new incoming information together with its last output at step $t-1$ Right: Unrolling a recurrent neural neuron through time.

neural architectures used for NLP: recurrent neural networks (RNNs), 1-dimensional convolutional neural networks (1-D CNNs), and self-attention.

- Recurrent neural networks (RNNs)

  RNNs is the basic architecture for dealing with sequences of data. In contrary to feedforward neural networks, they have the ability to store values based on the history of inputs in a sequence, and use them to evaluate each new sequence element. This is done using the internal state of a hidden layer, which functions as a neural network memory and leads the architecture to exhibit temporal dynamic behaviour. This is illustrated in figure 2.7 and described by the following equations:

$$o_t = f(h_T; \theta_o)$$
$$h_t = g(h_{t-1}, x_t, \theta_i, \theta_h)$$

  , where o_t is the output of the RNN at time t, x_t is the input of the node at time t, and h_t is the state of the hidden layer at time t. θ_o, θ_i, and θ_h are the weights of the output, input and hidden state of the RNN. Function h could be any linear or non-linear combination of the above parameters for preserving properties of the input history. Qualitatively, this means that at each step t the output of the neuron is kept and at step $t+1$ reinserted into the architecture together with the new input. Given the functions f and g, these values can be transformed, weighted and combined, in order for the network to integrate all necessary information existing in a sequence. In the past few decades, researchers have developed complex architectures of RNNs, such as LSTMs [259] or GRUs [260], which are able to exploit sequential information to the highest degree.

- 1 - dimensional convolutional neural networks (1-D CNNs)

  Researchers first developed CNNs for image recognition tasks. Inspired from the structure of the visual cortex [261], they developed an architecture that is able to take 2-dimensional inputs and combine neighbor values to improve models' inferences. The idea was that recognizing objects must be done by a model that

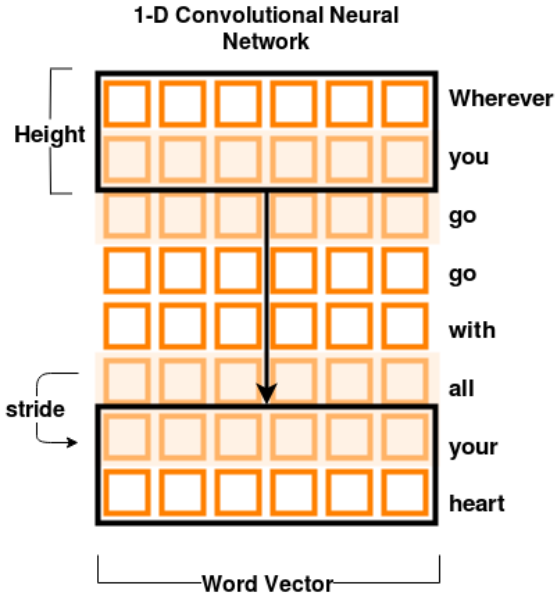

Figure 2.8: The 1-Dimensional CNN scans a sequence of embeddings height-wise, being able to capture semantic associations.

is able to capture shapes as spatial entities, and not as independent pieces of information. The simplified version of CNNs, which take one-dimensional data as input but nevertheless take into consideration neighboring values are called 1 - Dimensional CNNs. 1 - D CNNs decide how many sequences will be taken into consideration at each step by a stride of arbitrary height, multiplying all inputs with their respective weights and sending them to the next layer. The stride scans the totality of the sequence and hence is able to send information about each input based on their context. This can be formalised by the following equation:

$$a_i = \sum_{t=0}^{n-1} w_t x_{(i+t)} + b$$

, where i is the starting position of the stride, n is the height of the stride, t is the relative position of an element in the stride, $x_i + t$ is the input value of an element at position $i + t$, W_t the weight corresponding to the relative element position t and b is a bias term. 1 - D CNNs are able to successfully capture properties of text because they evaluate each word based on their context, which functions as a proxy for the meaning of a word.

- Self-attention

  Self-attention is probably the most efficient architecture for NLP, as it is able to capture the interplay and relations of words and sequences, while providing inferences with reduced computational cost. In contrast to RNNs and 1-D CNNs, that try to scan sequences iteratively and choose what information should be stored, Self-attention deals with text modeling as an information retrieval task. In information retrieval a scientist asks a query, searches over the keys (the existing

information) for the answer, and then returns a set of values (the results of the query). In self-attention, queries, keys, and values are all a transformed version of the input text sequence. The scientists asks each part of a sequence what other parts it should be associated with, and also to what magnitude. This is done by using following set of equations:

$$Q = W_q I, \quad K = W_k I, \quad V = W_g I$$

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{n}})V$$

, where I is the matrix version of the input sequence, Q, K, and V are the matrices for the Queries, Keys and Values respectively. W_q, W_k, W_g are the weights multiplied by the inputs in order to calculate the Queries, Keys and Values. The function $A(Q, K, V)$ associates the Queries to the Keys, and returns the resulting weighted values. Self-attention is a fundamental part of state-of-the-art complex neural architectures for NLP [262, 249]. Researchers have developed various differentiations of self-attention to improve models' inferences.

In chapter five, I employ neural networks both with self-attention and LSTM cells to investigate how biased machine learning models could be used for the detection of bias in new content.

2.4 Machine learning for recommender Systems

Political Machines largely include recommender systems for content suggestion. Social Media services employ algorithms for their news feed, automatically promoting, suggesting, and filtering content in a way that maximizes user engagement [263, 127, 141]. Similarly, online shops, product placement services, advertisement systems, all use recommendation algorithms for strategically influencing individuals [264, 265]. Therefore, the analysis and understanding of how recommender systems process inputs and generate suggestions is crucial for uncovering how influence processes take place on political machines. Depending the available data and scope of the application, recommendation systems vary both in terms of their recommendation technique, i.e. the qualitative way a suggestion is made, and their modelling technique, i.e. the quantitative way a suggestion is calculated. In the following, I present the scientific consensus of the existing basic recommendation and modeling techniques.

Recommendation techniques

The aim of recommendation systems is to suggest content that will influence an individual to perform a specific action. This might be buying a product, commenting on a post, clicking on an advertisement, changing their political opinion. As recommendation algorithms are bound to a specific platform, website, or mobile application, their training is dependent on the data available on the host service, thus the recommendation systems might have different structures. There are three main recommendation techniques: *collaborative filtering*, *content-based filtering*, and *hybrid recommendation systems* (figure 2.9).

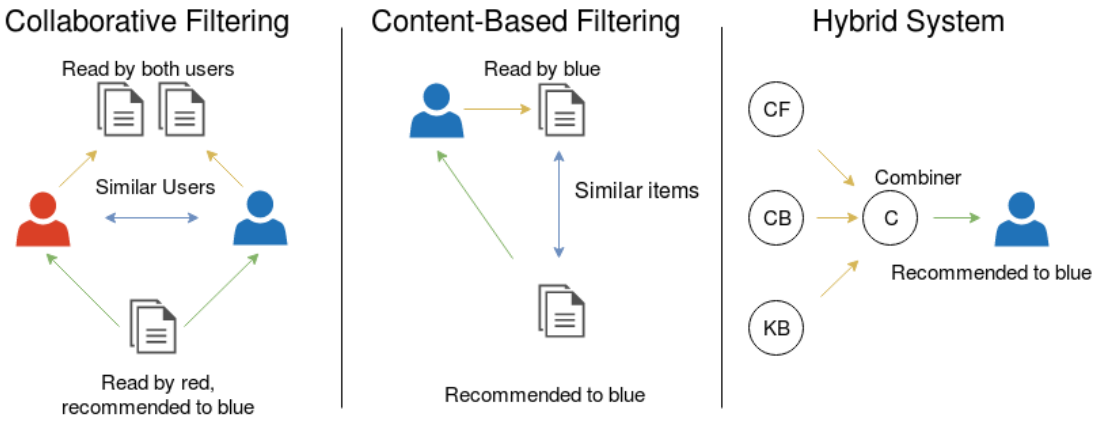

Figure 2.9: Left: Collaborative Filtering identifies similar users and recommends items based on that similarity. Center: Content-based Filtering identifies similar items that the user interacted with and suggests them to her. Right: Hybrid recommender systems combine collaborative, content-based and knowledge-based filtering to provide suggestions to the user.

In collaborative filtering, a data scientist has available data about the total user-item interactions and combines them to create personalized suggestions. This is done by searching for similar users. These are users that have interacted with almost the same items. Given two similar users A and B, there will be specific items that user A has interacted with but user B not. The algorithm will suggest these non-overlapping items to user B, when they are going to use the service the next time. Therefore, it necessary for collaborative filtering is to combine user data on a service, which should be adequate to describe each user's behavior. In case there are only few non-zero user observations, the algorithm suffers from the *cold start* problem. Because the input data is very sparse, any algorithm trained on it will not be able to consistently locate additional user preferences.

In content-based filtering, a data scientist has available information about available items and is able to rank them in terms of similarity. Given two similar items A and B, if a user will show interest in item A, then the algorithm will propose item B. In contrary to collaborative filtering, the algorithm does not require the totality and history of user-item interactions. On the contrary, it presupposes the existence of item features, based on which the algorithm can calculate a similarity metric. In this way, content-based filtering overcomes the cold start problem, because the algorithmic suggestion is based on the available items' properties, which are fully known by the data scientist.

Hybrid recommender systems combine all available information to calculate personalized suggestions. That might be the total and historical user-item interactions like in collaborative filtering, any item feature that can be used for comparing item similarity like in content-based filtering, or any other available information on the individuals. That might be any available profile information, including demographic and other features, that can contribute to the personalization of recommendations. The systems that exclusively use personal information for personalization are called *knowledge-based*. Because of the vast amount of available data on political machines, the most effective way to shape recommendation is by combining all available information and training hybrid recommender systems.

Modeling techniques

Regardless of the available data and the recommendation technique, a data scientist decides upon the mathematical architecture for processing the related information. In the following, I describe three basic modeling techniques: similarity-based models, factorization-based models, and deep neural models. I also describe the accompanying model selection criteria, which guide the data scientist in determining to the final model.

Similarity-based models

In collaborative and content-based filtering, a data scientist needs either to quantify the similarity between users' behavior or items. The most simple way to do so is to apply a metric that compares the user or item vectors. Given two vectors u and v, a function $sim(u, v)$ can map their similarity into R. By calculating the pairwise similarity of an item or user to the rest, the data scientist can create a ranking of similar users or items, from which they can sample the recommendations. For that task, a scientist can deploy multiple similarity functions, some examples of which are below:

- Pearson correlation

  Pearson correlation is one of the simplest distance metrics. It measures the linear correlation of two items or users by the equation

  $$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  , where x,y two users or items, n the number of available observations, and $\bar{x}$ and $\bar{y}$ their mean values.

- Jaccard index of similarity

  The Jaccard index of similarity quantifies how similarly two users engage to a set of items by using the equation

  $$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

  , where X, Y are the sets of items that two users engaged in respectively. The higher the overlap of engagements, the higher the Jaccard index and similarity of user behavior.

- Spearman's ranked correlation

  Spearman's ranked correlation quantifies the correlation of item rankings. It uses the same formula as the Pearson correlation but this time x_i, y_i provide the ranking position of the item values. It is valuable in cases, where the input features are ordinal or at significantly different scales in relation to each other.

- Kendal Tau correlation

  Kendal Tau is another ranked correlation metric of two items or users, calculated by

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}.$$

For any pair of observations $(x_i, y_i), (x_j, y_j)$ with $i < j$, the pair is considered concordant if both $x_i > x_j$ and $y_i > y_j$ or if $x_i < x_j$ and $y_i < y_j$. If these conditions are not fulfilled, the pair is considered discordant. Kendal Tau is usually more robust than Spearman's row, because it is mathematically more tractable [266] and it provides more reliable confidence intervals [267].

- PSS similarity

  PSS (Proximity-Significance-Singularity) similarity [268], weighs the proximity, significance, and singularity of an item in relation to a set of users. It investigates how similarly a set of users interact with the item, how much their ratings deviate from the median, and how special in general that item is. These are given by the following set of equations:

$$PSS\ (r_{u,p},r_{v,p}) = Proximity\ (r_{u,p},r_{v,p}) \times Significance\ (r_{u,p},r_{v,p}) \times Singularity\ (r_{u,p},r_{v,p})$$

, with

$$Proximity\ (r_{u,p},r_{v,p}) = 1 - \frac{1}{1+\exp(-|r_{u,p}-r_{v,p}|)}$$

$$Significance\ (r_{u,p},r_{v,p}) = \frac{1}{1+\exp(-|r_{u,p}-r_{med}|\cdot|r_{v,p}-r_{med}|)}$$

$$Singularity\ (r_{u,p},r_{v,p}) = 1 - \frac{1}{1+\exp\left(-\left|\frac{r_{u,p}+r_{v,p}}{2}-\mu_p\right|\right)}$$

, where r the value of engagement, p the item and u, v two users respectively.

Regardless of the chosen similarity metric, the data scientist ranks items and then generates personalized suggestions based on that ranking.

Factorization-based models

An alternative to applying similarity metrics on the database is the application of factorization techniques. Given either a matrix of user-item interactions (collaborative filtering), of user features (knowledge based-filtering) or of item features (content based-filtering), a data scientist can decompose them linear algebraically and conclude latent relations between users, items, or between users and items. The most common factorization techniques for that is the basic matrix factorization (MF), and the singular value decomposition (SVD).

MF takes a matrix M as input and calculates matrices U and V so that it holds $M \approx U * V$. In collaborative filtering, $\tilde{M} \in \mathbb{M}^{users \times items}$ represents the predicted matrix of user-item interactions, and $U \in \mathbb{R}^{users \times latentfactors}$ and $V \in \mathbb{R}^{latentfactors \times items}$ represent the projection of users and items in a latent space of k dimensions respectively. Specifically, the predicted action of user u on item i is computed as

$$\tilde{m}_{ui} = \sum_{f=0}^{kfactors} U_{u,f} V_{f,i}.$$

This problem is solvable by optimizing the following cost function:

$$\arg\min_{U,V} \|M - \tilde{M}\|_{\mathrm{F}} + \alpha\|U\| + \beta\|M\|$$

, where $||.||_F$ the frobenius norm and α, β regularization terms. Given the position of users and items in the latent space, the data scientist can infer about their similarity. The above optimization algorithm is called Funk MF [269]. For the same problem, a data scientist can deploy alternatives like the NNMF, or the SVD++ [270]. In knowledge based and in content-based filtering the scientist can apply the same process, but depending the case matrix M represents a user-features or an item-features matrix respectively, and the factorization provides information about the user and item similarity accordingly.

SVD calculates also two matrices U and V by the factorization $M = U\Sigma V^*$, where Σ is a diagonal matrix, and U, V are orthogonal matrices containing users, items, or features projected in the latent space, depending the recommendation technique. The problem is solvable by reducing M to its bidiagonal version and applying an iterative process of Householder transformation, QR factorization and QR decomposition [271]. As in MF, the data scientist can extract from U, V the related similarities and proceed to the sampling of recommendations.

Deep neural models

The complexity of available data, the extreme sparsity, feature variability, and the dynamics in the DGP often make standard similarity metrics and factorization techniques infeasible for recommendations. An alternative to that is the implementation of deep neural architectures, which are not only able to take any data combination into consideration, but also to exploit the ability to model highly non-linear processes to generate accurate suggestions. Deep neural models is the standard technique used by major tech companies when designing recommendation algorithms [263, 127, 272], with the deployed architectures varying not only in terms of inputs and outputs, but also in terms of hidden layers and cost functions. Depending on the exact recommendation technique the networks are modeling, as well as the nature of the existing data, architectures are adapted accordingly. For example, scientists have developed neural collaborative filtering models [273, 274], models that combine matrix factorization and neural architectures [275], pure deep learning models based on RNNS [263], or CNNs [276], and models that exploit social graph properties [277]. Regardless of the exact architecture, the data scientist optimizes models based on their ability to suggest content that can successfully influence individual behaviour, such as any other hybrid recommendation system (figure 2.9, right).

Because recommender systems input data is often sparse and highly skewed, data scientists need to deploy cost functions that guide the models' training process in an efficient way. The simplest cost function that can be deployed is the root mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i \in R_u}(p_i - r_i^2)}{|R_u|}}$$

, where R_u is the list of recommendations and p_i is the models' prediction of item i and r_i the actual behavior of the user towards item i. Nevertheless, the specific cost function

faces issues when engagement values are zero-inflated or highly skewed. To that end, scientists deploy more sophisticated architectures that cope with these issues. A solution to data sparsity is negative sampling [278], which denotes the loss calculation of only K randomly sampled items, together with the actual positive label for a recommendation. That makes the model not only less computationally expensive, but also the balances the importance between the items a user engaged with to those they did not.

For dealing with data skewness, data scientist propose further techniques. A solution is the weighting and the logistic transformation of a cost function [263]. A prominent function that does that is the weighted cross entropy:

$$\text{WCE}\left(r,p\right) = -\left(\beta_{r_1}\log\left(p\right) + \beta_{r_2}(1-r)\log\left(1-p\right)\right)$$

, which uses a class-specific parameter β_{r_i} that weighs the cost for the class prediction. Another solution is to use rank based cost functions, which are highly robust towards skewed data [279].

In chapter five, I employ deep neural and factorization-based models in order to investigate the political user-algorithm interaction on social media platforms and I deploy different cost functions to investigate their impact on the generated recommendations.

Model selection

Regardless of the recommendation and modeling technique, the data scientist should evaluate models' performance for selecting the optimal architecture. For that purpose, multiple metrics are available. In the following, I present six properties a recommendation system should have and metrics that could be used for their evaluation. I use R_u for the list of recommendations for user u, C_u for list of items that the user engaged with, p_i as the models' prediction of item i and r_i as the actual behavior of the user towards item i.

- **Accuracy**

  The data scientist investigates the models' predictive accuracy by deploying metrics that can quantify the error. These might be the *Mean Absolute Error* or the *Root Mean Squared Error*, formulated as

  $$MAE = \frac{\sum_{i \in R_u} |p_i - r_i|}{|R_u|}$$

  and

  $$RMSE = \sqrt{\frac{\sum_{i \in R_u} (p_i - r_i^2)}{|R_u|}}$$

  respectively. More sophisticated methods are the *Precision* and Recall metrics, given by

  $$Precision = \frac{|C_u \cap R_u|}{|R_u|}$$

  and

  $$Recall = \frac{|C_u \cap R_u|}{|C_u|}$$

  respectively. Precision measures the rate of items in the recommendation list that the user actually engaged with. Recall quantifies the ratio of items in the

recommendation list compared to the total number of items the user engaged with. Because in practice, a user will not come into contact with all the potential recommendations of the list, but only with the first K, adapted versions of the above metrics exist that take into consideration only these K first elements, called *precision@K* or *Recall@K*.

The above accuracy metrics do not take into consideration the rank of the items, which is crucial for recommender systems. The models ideally should recommend the most important items for the users first, in order to optimize their performance. Towards that end, the *Discounted Cumulative Gain* (DCG) can quantify the ranked accuracy of a model by the equation

$$DGP@K = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)}$$

, where rel_i is the actual relevance of the result at position i. Because $DGP@K$ is dependent on K, the data scientist can calculate the normalized DCG (NDCG), which is K-invariant. This is given by

$$NDCG@K = \frac{DCG@K}{IDCG@K}$$

, where

$$IDCG@K = \sum_{i=1}^{K} \frac{2^{rel@K_i} - 1}{\log_2(i+1)}.$$

$REL@K$ is the list of relevant items in the database up to position K. In chapter four I deploy MAE, RECALL@K, and NDCG@K to evaluate multiple recommendation systems performance.

- **Utility**

  Beside models' accuracy on test data, the data scientist can deploy user experiments that quantify how the users react to system recommendations. Based on them, the scientist can calculate utility metrics. The most simple metric is the *Interaction Rate* (IR), which gives the ratio of interacted items to the total items suggested by the systems. This is given by

$$IR = \frac{C_u}{R_u}.$$

  In order to evaluate how the system keeps users engaged dynamically, a data scientist can measure the retention of the system, i.e. the impact of the recommender systems in keeping users engaged. This is done by using A/B user testing, and by calculating the retention Δ, given by

$$\Delta_{retention} = MIR_T - MIR_C$$

  , where MIR_T and MIR_C are the mean interaction rate of the treatment and control group respectively. The data scientist can replace MIR with any other measure meaningful for quantifying the system's utility.

- **Novelty**

  Novelty quantifies how many of the suggested items suggested are new to the users. New might denote items or classes of items that the user has never seen or interacted with before, or items or classes of items that the recommender system itself has not proposed to the user before. Depending on the definition, a different rate can be calculated. A simple novelty metric is

  $$Nov = \sum i \in R_u.$$

- **Diversity**

  Diversity describes how different the items in a recommendation list are. That can be how many different items or classes of items were included in the list given the total available items or classes or items. It might also be how dissimilar items given specific item features are. A general function for diversity is

  $$Div = \sum_{i \in R_u} \sum_{j \in R_u, j \neq i} d(i, j)$$

  , where i, j two arbitrary items and $d(.)$ any distance function that quantifies the dissimilarity of items.

- **Serendipity**

  Serendipity describes the property of recommendations to be novel, unexpected and useful [280]. Serendipity is similar to diversity in the sense that a system must propose non redundant classes of items to the user, and it is similar to novelty since a system should also propose unexpected and in a sense unknown items. Serendipity thus can be generally defined by any function of the form

  $$SER = f(utility, unexpectedness)$$

  , which takes into consideration the utility of the proposed suggestions and how unexpected they are to the user[281].

- **Coverage**

  Coverage denotes the general capability of a system to provide efficient recommendations for all possible items from all possible classes, to all possible users. In its simplest form a coverage metric can be

  $$COV = \frac{|I_P|}{I}$$

  , which gives the rate of predicted items or classes of items I_P to the total available number in the dataset I.

The evaluation of the above systems' metrics is not only important for reassuring the models' mathematical rigor, but also in order to maximize the models' ability to influence human behavior.

2.5 Methodological contributions

In the previous sections, I provided an overview of existing machine learning techniques used to understand political machines. I described the scientific method applied in this thesis based on the KDD process (2.1). I provided the statistical and probability foundations used to compare properties of individuals and their interactions on political machines. I described the tools for formally modeling behavioral distributions and the fundamental statistical tests for comparing and testing population and sample properties. Then, I gave a general overview of machine learning models' properties (2.2).

Because text is a fruitful source of political information, this thesis deploys data-intensive NLP algorithms for extracting political knowledge. In 2.3 I described the foundations of NLP, and introduced the scientific consensus on three prominent NLP models: topic models, word embeddings, and text-based neural networks.

Finally, in 2.4 I analyzed the structure of recommender systems, which are common techniques for social influence appearing on social media platforms and ADM systems. I illustrated different architectures, model validation techniques, and described issues faced by the specific systems.

The thesis builds on the above statistical, probability and machine learning foundations in order to uncover properties of political machines. I adopt and extend existing machine learning techniques to analyze three case studies of algorithmic applications: A. data-driven microtargeting, B. political communication and recommender systems, and C. text-based algorithms for ADM systems. The methodological contributions of the thesis are following:

A. Data-driven microtargeting

- I illustrate that topic modeling algorithms are able to identify concrete user preferences in an efficient way. By analyzing a dataset of political user interactions on Facebook, I recognize partisanship tendencies and the contents that are of interest to individual users.

B. Political communication and recommender systems

- On similar social media political datasets I analyze the user activity distributions, by fitting probability distributions and performing statistical tests. I conclude that the log-normal is the formal distribution that describes the specific user behavior the best. This result complies with previous research stating that social behaviour tends to be highly skewed, asymmetric, and heavy-tailed distributed.

- I train standard and deep neural recommendation systems on simulated networks of social media political data to investigate how they influence the political discourse. I illustrate that standard recommender systems such as collaborative filtering are unable to correctly model user behaviour. I show that deep neural networks and rank-based functions are the best way to deal with the input data skeweness on the social media setting.

- I attack the trained deep neural recommender systems by graph-poisoning. I show that few adversarial examples are more than enough to manipulate the models'

suggestions in the network, raising questions of robustness about undisclosed algorithms deployed by social media platforms.

C. Word embeddings and unfair algorithms

- To understand how word embeddings influence ADM systems, I train vector representations of words based on German Wikipedia and social media data. To quantify immanent sexist, xenophobic, and homophobic biases, I develop a new methodology for bias detection in gendered languages.
- By training a sentiment classifier on the word embeddings I prove that models taking the embeddings as input carry the same biases as the vectors. I also illustrate that the standard debiasing technique at the embeddings level does not suffice for eliminating biases in the end product of the models.
- I illustrate that deep neural architectures trained on biased word embeddings are able to locate similar biases on new data.

3 Social media and data-driven microtargeting

3.1 Social media and microtargeting in Germany

Authors

Orestis Papakyriakopoulos, Morteza Shahrezaye, Andree Thieltges, Juan Carlos Medina Serrano, Simon Hegelich

In

Informatik-Spektrum, volume 40, pages 327–335, Springer-Verlag Berlin Heidelberg 2017. DOI: 10.1007/s00287-017-1051-4. Published as "Social Media und Microtargeting in Deutschland." The following is the author's version of the article. The final publication is available at link.springer.com

Abstract

Citizens in Germany increasingly use social media platforms for political debating. The data they produce in their interactions can be really valuable for political Microtargeting. The application of machine learning algorithms on them can cluster users with similar behaviour and attitudes. In this way, someone can identify social groups and individuals, as well as the political content their interested in. Political parties in USA already use Microtargeting for their puroposes, as they possess detailed information about the american electorate. Similar data do not exist in Germany, as the german privacy law forbids the collection, processing and evaluation of personal data. In this article, we demonstrate how someone can perform Microtargeting in Germany, in a way that complies with the existing privacy law. This is achieved through the collection of public user data from the social network Facebook and their algorithmic processing. Given the technological possibilities, we discuss the ethical and political consequences for the political system.

Contribution of thesis author

Theoretical design, model design and analysis, manuscript writing, revision and editing

Social Media und Microtargeting in Deutschland

Orestis Papakyriakopoulos[1], Morteza Shahrezaye[1], Andree Thieltges[1],
Juan Carlos Medina Serrano[1], Simon Hegelich[1]
[1]Bavarian School of Public Policy, Technical University of Munich, Munich, Germany
orestis.papakyriakopoulos@tum.de, morteza.shahrezaye@hfp.tum.de, andree.thieltges@hfp.tum.de,
juan.medina@tum.de, simon.hegelich@hfp.tum.de

Zusammenfassung:

Politische Debatten werden in Deutschland zunehmend über soziale Medien geführt. Die dabei produzierten Daten können mit geeigneten „machine learning“ - Verfahren für politisches Microtargeting genutzt werden. Die Anwendung von maschinellem Lernen auf diesen Daten ermöglicht das Zusammenfassen von Nutzern mit ähnlichem Verhalten oder Präferenzen. Dadurch können Gruppen identifiziert werden, die für bestimmte politische Inhalte besonders interessant sind. In den USA werden diese Verfahren bereits intensiv genutzt. Allerdings verfügen die dortigen politischen Akteure über Zugriff auf detaillierte Informationen über die Wähler. Solche Daten stehen in Deutschland nicht zur Verfügung, da die deutschen Datenschutzrichtlinien deren Sammlung, Verarbeitung und Auswertung verbieten. Im folgenden Artikel zeigen wir, wie es im Einklang mit den deutschen Datenschutzgesetzen möglich ist, Daten aus dem sozialen Netzwerk Facebook zu extrahieren und damit Microtargeting zu betreiben. Vor diesem Hintergrund werden abschließend die ethischen und politischen Konsequenzen für das politische System diskutiert.

Einleitung

Die aktuell stattfindenden Digitalisierungsprozesse verändern den politischen Diskurs und die Art und Weise, wie politische Akteure agieren. Big Data [23] bedeutet im politischen Bereich, dass immer mehr Bereiche des menschlichen Verhaltens kategorisiert, quantifiziert und aggregiert werden. Insbesondere durch die Nutzung des Internets und von sozialen Netzwerken werden in bisher undenkbaren Ausmaß Daten von Bürgern erhoben, die Aufschluss über private und politische Einstellungen geben können. Zudem entsteht eine neue Kommunikation zwischen Politik und Bürger, gerade über die sozialen Medien, die von vorneherein datengetrieben ist. Vervollständigt wird dieses Bild von Big Political Data durch moderne Algorithmen, insbesondere aus dem Bereich des *„machine learning“ (s. Artikel „Algorithmen und Meinungsbildung“ von Zweig und Krafft in diesem Heft)*, mit deren Hilfe man diese Datenmengen strukturieren und im Prinzip auf die Ebene des einzelnen Bürgers herunterbrechen kann. Vor diesem Hintergrund beginnen politische Akteure diese neu entwickelten Tools zur Analyse von Wählerverhalten und zur Einflussnahme der Wählerschaft zu nutzen. Eine Methode, die dafür eingesetzt wird, ist das sog. Microtargeting: Datenanalysen fließen in Wahlkampfstrategien ein, um einzelne Wähler oder Wählergruppen gezielt anzusprechen [1].

Der vorliegende Artikel zeigt exemplarisch in einem „proof of concept“, dass Microtargeting auf Basis von Daten aus sozialen Netzwerken auch in Deutschland möglich ist. Ziel der Untersuchung ist dabei, das Potenzial und die Gefahren dieser neuen Methode des Wahlkampfs vor dem Hintergrund des technischen „state of the art“ zu diskutieren. Dafür wurde eine Analyse von Facebook-Daten durchgeführt, die in einer tatsächlichen Wahlkampagne eingesetzt werden könnten. Im Folgenden wird zunächst das Konzept des Microtargeting erläutert und auf bestehende Hindernisse für dessen Einsatz eingegangen. Daran anschließend wird die Analysemethode beschrieben und die beispielhaften Ergebnisse werden vorgestellt. Abschließend werden diese Befunde unter besonderer Berücksichtigung ethischer Fragen diskutiert.

Microtargeting aus theoretischer Perspektive

Man kann politisches Microtargeting als strategischen Prozess beschreiben, der auf die Beeinflussung von Wählern durch die direkte Übertragung von Reizen oder Informationen abzielt. Dafür werden die Präferenzen und Verhaltensweisen, die aus Datensätzen der jeweiligen Individuen abgeleitet werden können, benutzt. Um dies möglichst genau abbilden zu können, benötigt man zunächst große Datenmengen, die sowohl Aufschluss über die politischen Präferenzen, als auch über das für politische Entscheidungen nicht relevante Verhalten der Wählerschaft geben. Dabei ist es zunächst egal, ob die Daten manuell oder durch „data-mining" gesammelt und aggregiert werden. Relevante Daten für das politische Microtargeting umfassen sowohl die Namen und Adressen von Wählern sowie deren zurückliegende Wahlentscheidungen, als auch abstraktere Parameter, wie bspw. persönliche Meinungsäußerungen über politische und unpolitische Themen, soziale Interaktionen und kulturelle Interessen oder soziodemografische Faktoren. Diese aggregierten Informationen und Daten über weitere Verhaltensweisen der Wähler werden mithilfe von geeigneten „machine learning"- Algorithmen strukturiert: Dabei geht es entweder darum, Vorhersagen über bestimmte Variablen wie z. B. die Wahlentscheidung zu treffen (sogenanntes „supervised learning") oder aber eine Ordnung in Form von Gruppen und Clustern in die Daten zu bringen („unsupervised learning"). Das heißt, die verwendeten Algorithmen sortieren die vorliegenden Informationen und bilden Untergruppen von Wählern mit gemeinsamen Charakteristiken und Eigenschaften (bspw. demografische Eigenschaften oder übereinstimmende/sich ähnelnde Ansichten und Verhaltensweisen) [26]. Auf Grundlage dieser „algorithmischen Sortierung" wird es möglich, Wahlwerbung individuell auf die Wähler der unterschiedlichen Subgruppen abzustimmen oder diversifizierte Wahlkampfstrategien zu entwickeln (sog. Nanotargeting) [15].

Der Einsatz von politischem Microtargeting reicht zurück bis zu den US-Präsidentschaftswahlen im Jahr 2000 [26]: Dort setzte die republikanische Partei Microtargeting in recht überschaubarem Umfang, aber mit großem Erfolg erstmals ein. Begünstigt durch die zunehmende und inzwischen massenhafte Sammlung von Daten auf fast allen gesellschaftlichen Ebenen wird dem Einsatz von Microtargeting als politischer Strategie in den USA seither großes Potenzial beigemessen: In den US-Präsidentschaftswahlen 2008 beruhte die sehr erfolgreiche Wahlkampfstrategie der demokratischen Partei sehr stark auf Microtargeting[17]. Dementsprechend kann Microtargeting inzwischen als durchgesetzte Methode bei Wahlkampagnen in den USA betrachtet werden [26], auch weil es Lösungen für die „klassischen" Wahlkampfproblematiken bietet: Die Prädispositionen und allgemeinen Interessen der Wähler werden sichtbar [16], was bspw. ein bessere Kandidatenausrichtung, beziehungsweise exaktere Anpassungen an die Wählerinteressen ermöglicht [8, 9]. Mit der Möglichkeit, individualisierte und passgenaue Wahlwerbung für bestimmte Einzelwähler oder Wählergruppen zu erstellen, sinkt das Risiko andere Wählergruppen mit bestimmten Themen in ihrer Wahlentscheidung zu verunsichern [34]. Durch den Einsatz von Microtargeting können sich politische Akteure zudem Zugang zum gesamten Wählerspektrum verschaffen, da ihre Wahlkampfstrategie nicht länger auf den Charakteristiken des „Medianwählers" beruht [13]. Die Basis für den durchschlagenden Erfolg von politischem Microtargeting in den USA stellen die vergleichsweise laschen Datenschutzregeln dar: Diese verschaffen den politischen Akteuren fast ungehindert Zugriff auf Datenbanken mit persönlichen Informationen der Wähler [32]. Restriktiver Umgang oder Zugriffsverbote auf „sensible" Daten sind in der US-amerikanischen Gesetzgebung nicht oder nur indirekt (bspw. durch die Festlegung der Datenverarbeitung auf bestimmte Zwecke oder zeitlich begrenzte Nutzung) verankert [7]. Im Vergleich dazu wird in Deutschland die Sammlung und Aufbereitung von persönlichen Daten durch Datenschutzrichtlinien viel stärker begrenzt [11, 14, 27].

Hindernis 1: Datenschutz

Es ließe sich daher argumentieren, dass Microtargeting in Deutschland keine Zukunft hat. Gerade im Bereich der sozialen Netzwerke liegen aber sehr viele Daten vor, die auch unter Berücksichtigung der deutschen Datenschutzlage für Microtargeting verwendet werden können[1]: Solange die Nutzer eigenständig Informationen über ihre politischen und sonstigen Präferenzen veröffentlichen und ihr politisches Verhalten offenbaren, ist eine Analyse dieser Daten legitim. Darüber hinaus sind soziale Netzwerke inzwischen einer

[1] Solange der Gebrauch der Daten den Interessen des Individuums nicht entgegensteht, erlaubt das deutsche Datenschutzgesetzt das Sammeln und Verarbeiten von veröffentlichten Personaldaten (bspw. aus sozialen Netzwerken) [12].

der Hauptkanäle auf dem Politiker mit der Wählerschaft kommunizieren [4, 20, 25] oder versuchen, politische Kampagnen zu bestimmten Themen anzustoßen[2]. Die Daten, die auf diesen öffentlichen Seiten anfallen, werden bereits von den Parteien erfasst und ausgewertet. Sofern nicht – wie in den USA üblich – verschiedene Datenquellen miteinander kombiniert werden, sondern die Datensammlung auf diese Social-Media-Daten beschränkt wird, besteht also in der derzeitigen Rechtslage kein Hindernisgrund für Microtargeting.

Hindernis 2: Bias in den Daten

Daten aus sozialen Netzwerken geben kein realistisches Bild der echten Welt und der echten Wählerpräferenzen. Die Gruppe der Leute, die sich auf Facebook zu politischen Themen äußert, ist nicht repräsentativ für die Gesamtbevölkerung [30], die Art der Äußerungen online ist nicht einfach mit einer politischen Meinungsäußerung gleichzusetzen (ein Like ist keine Wählerstimme) [20] und die Auswertung von Social Media Daten ist mit vielen ernsten methodischen Herausforderungen verbunden [19]. Trotz dieser – zum Teil ungelösten – Probleme gehen wir davon aus, dass Microtargeting dennoch benutzt werden wird. In den USA lässt sich zeigen, dass die Parteien ihre Kampagnen an dem Bild des Wählers orientieren, das sich aus den Daten ergibt, auch wenn dieses Bild unvollständig oder fehlerhaft ist [21]. Dieses *Perceived Voter Model* vereinfacht Entscheidungsprozesse in der Kampagnensteuerung. Da es beinahe unmöglich ist, die Wahlentscheidungen letztlich auf den Einsatz eines bestimmten Werkzeugs zurückzuführen, werden Wahlkampfmanager auf dieses Instrument setzen, solange sie von einer positiven Wirkung ausgehen – auch wenn diese vielleicht in Wirklichkeit gar nicht eintritt.

Methode

In diesem Artikel zeigen wir exemplarisch, wie es für Politiker möglich wird, Microtargeting im sozialen Netzwerk „Facebook“ einzusetzen und evaluieren anschließend einige der ethischen und politischen Konsequenzen, die sich daraus ergeben. Für unseren „proof of concept“ haben wir die Nutzeraktivitäten der öffentlichen Facebook Seiten deutscher Parteien und ihrer Wähler analysiert: Unser Sample umfasst folgende Parteien: CDU, CSU, SPD, FDP, Bündnis 90/Die Grünen, AfD. Konkret wurden dabei die „Likes“ der Nutzerinnen und Nutzer zu politischen Posts ausgewertet und auf Grundlage der unterschiedlichen Nutzerpräferenzen die Parteizugehörigkeit der Wähler ermittelt. Auf Basis von standardisierten Microtargeting-Auswertungsmethoden konzentrieren wir uns bei der Auswertung auf Nutzerinnen und Nutzer, die Inhalte auf Seiten von unterschiedlichen Parteien „gelikt“ haben. Dieser Fokus ergibt sich aus dem Potenzial der Gruppe der sog. Wechselwähler*:* Diese sind für potenzielle Beeinflussungen besonders empfänglich[3]. Nachdem so die unterschiedlichen Gruppen der Wechselwähler identifiziert worden sind, haben wir mit Machine-Learning-Algorithmen alle „Posts“ auf den Facebook-Seiten der verschiedenen Parteien zu 100 unterschiedlichen Themenclustern zusammengefasst. Dabei wird *Topic Modelling* basierend auf einer Latent Dirichlet Allocation (LDA) in Kombination mit einer Principal Component Analyse (PCA) verwandt. Mit dieser Methode können wir zeigen, wie es möglich wird, für die Gruppe der Wechselwähler Wahlwerbung mit individuell abgestimmten Themen zu erstellen.

Die Grundvoraussetzung für politisches Microtargeting stellt eine ausreichende Datenmenge über das Verhalten und die Präferenzen von Wählern dar. Dementsprechend haben wir über die „Facebook Graph API“ auf 438 Facebook-Seiten zugegriffen, die den o.g. Parteien zuzuordnen sind und die dortigen Posts, Comments und Likes analysiert. Das Auswahlkriterium für die Facebook-Seiten war die Suche nach dem Parteinamen in den Namen von Facebook Seiten. Die Suchergebnisse wurden dann durch eine manuelle Zuordnung überprüft[4]. Auf den Seiten wurden alle Posts extrahiert, die von den Betreibern seit dem Zeitpunkt der Seitenerstellung veröffentlicht wurden, sowie die Likes zu jedem Post, die von Facebook festgelegte ID der Nutzer die geliket haben und deren Profilname: In vielen Fällen ist der Profilname des Facebook Accounts deckungsgleich mit dem „Klarnamen“ der Person oder dieser ist durch relativ geringen Aufwand heraus zu finden. Dies war jedoch nicht Gegenstand unsere Studie. Insgesamt wurden 235.135

[2] Teilweise geschieht dies ebenfalls unter Nichtbeachtung des Datenschutzes.

[3] Wechselwähler besitzen ein hohes Interesse an Politik und sind bei ihrer Wahlentscheidung zumeist nicht festgelegt [11] [21].

[4] Die Facebook-Seiten der CDU und der CSU wurden dabei unter dem Begriff „Union“ zusammengeführt.

Posts und damit verbunden 6.696.954 Likes von uns gesammelt und darüber die Facebook-Aktivitäten von 1.399.510 Nutzern identifiziert. Die Zahl der Nutzerinnen und Nutzer bezieht sich ausschließlich auf die identifizierten Facebook-Seiten. Bei entsprechender Ausweitung des Samples kann die Zahl der rückverfolgbaren User also durchaus erhöht werden. Unsere Definition eines Anhängers einer bestimmten Partei leitet sich aus dessen Verhalten im sozialen Netzwerk ab: Er muss mindestens einen Post auf der Facebook-Seite der jeweiligen Partei „gelikt" haben. Wir unterscheiden davon den „Wechselwähler", der Posts von mindestens zwei oder mehr Parteien „gelikt" hat. Selbstverständlich kann man aus den „Likes" einer Person auf Facebook nicht automatisch auf eine Parteilichkeit schließen, allerdings hat es sich in unserer Studie als plausible Klassifikationsmethode erwiesen, da das „Liken" immerhin eine gezielte Interaktion mit den Inhalten der Partei unterstellt[5]. Abb. 1[6] zeigt, dass 50% der User auf den untersuchten Seiten der Parteien nur einmal einen Post „gelikt" haben. Wie häufig in Big-Data-Anwendungen, liegen also über die Mehrzahl der Nutzer nur sehr wenige Informationen vor.

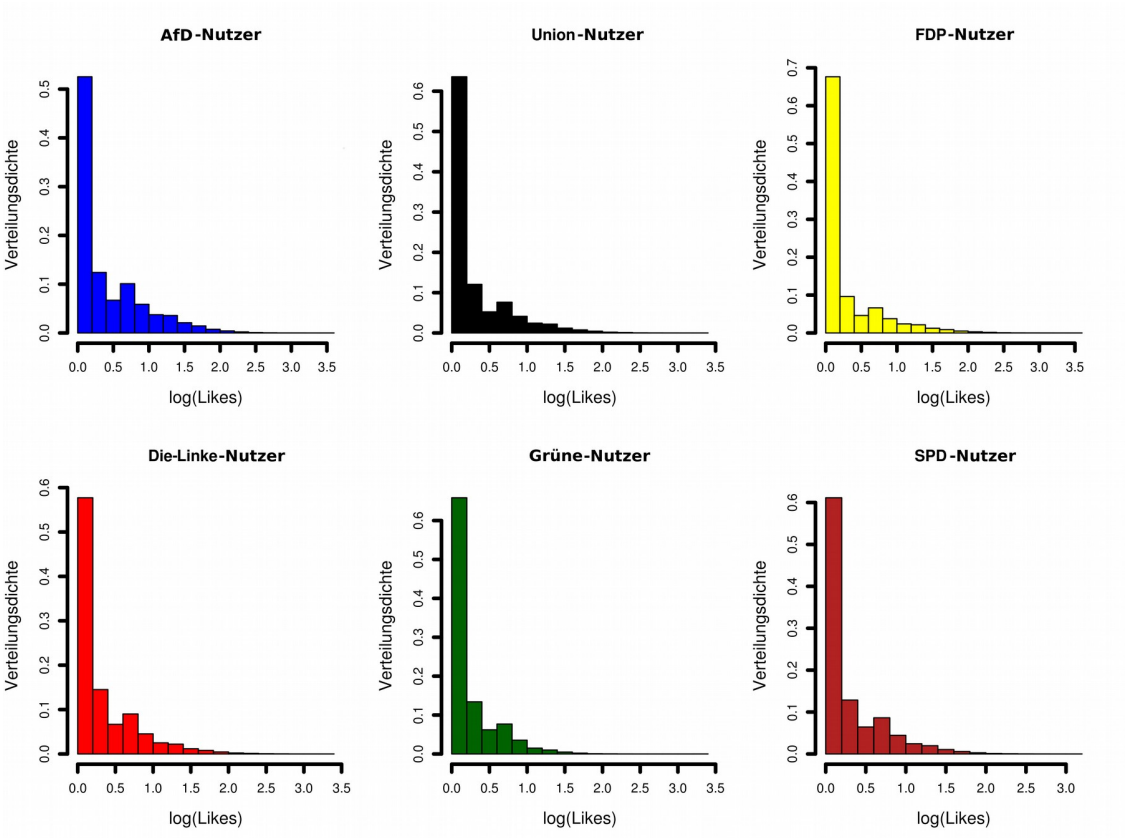

Abb. 1. „Likes pro Nutzer"-Verteilung auf AfD- und Union- Seiten.

Neben der Identifizierung möglicher Wechselwähler haben wir analysiert, für welche Inhalte diese Nutzer sich interessiert haben. Dafür wurde Topic Modelling mit dem LDA-Algorithmus über die 235.135 Posts angewandt. Ziel dabei ist es, für jeden Post die Wahrscheinlichkeit zu ermitteln, dass er einem allgemeinen Topic zuzuordnen ist. Dabei werden in einem rechenintensiven iterativen Verfahren die Wahrscheinlichkeiten ermittelt, dass ein bestimmtes Wort zu einem Topic gehört und dass ein bestimmter Post, der sich aus einem Teil der möglichen Worte zusammensetzt, zu einem Topic zugeordnet werden kann [6]. Dabei muss die Anzahl der Topics im Vorfeld bestimmt werden. Für unseren „proof of concept"

[5] Politische Microtargeting Ansätze zielen zumeist darauf ab, die Prädisposition des Wählers zu identifizieren. Eine genaue Überprüfung, ob jemand tatsächlich eine bestimmte Partei unterstützt, liegt zumeist nicht im Untersuchungsinteresse.

[6] Die Verteilungsdichte gibt prozentual an, wie oft die Nutzer auf den jeweiligen Seiten Posts gelikt haben (0,6 sind hier bspw. 60% der Nutzer).

haben wir 100 Topics festgelegt, so dass im Durchschnitt 2351 Posts in jedem Topic vorkommen. Der LDA Algorithmus gilt im Gegensatz zu Standardverfahren des Text-Mining als exakter [18] und ist in der Lage, komplexe Beziehungen im Datensatz zu erkennen. Zum Zweck der Visualisierung der Ergebnisse haben wir mithilfe des *Principal-Component-Analysis*-Algorithmus (PCA) [22] die entstandenen Cluster als zweidimensionalen Raum dargestellt (s.Abb.2). Die Distanz zwischen zwei Themenclustern zeigt hierbei ihren potenziellen Zusammenhang, wobei sich jedes Thema aus ähnlichen oder identischen Schlüsselbegriffen zusammensetzt, die in den untersuchten Posts häufig vorkommen[7].

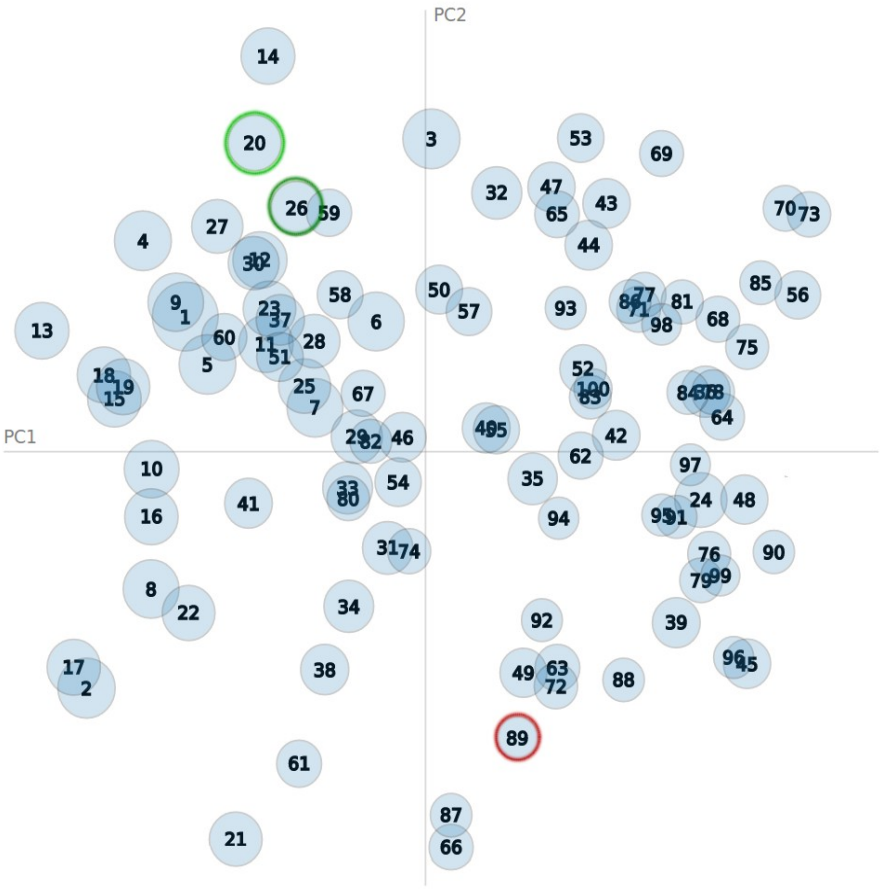

Abb. 2. Visualisierung der Topics mithilfe von PCA.

Ergebnisse

Die Anzahl der mit unserer Methode zu identifizierenden Wechselwähler beträgt insgesamt 149.292. Abb. 2 zeigt, dass die Wechselwähler die Tendenz haben, häufiger Posts zu „liken“ als der gewöhnliche User. Dies bedeutet allerdings nicht notwendigerweise, dass Wechselwähler in sozialen Netzwerken „aktiver“ sind, sondern ergibt sich zunächst dadurch, dass wir ja nur Aktivitäten auf den Seiten der Parteien verfolgen und Wechselwähler per Definition auf mindestens zwei Seiten aktiv waren. Vor dem Hintergrund des Perceived Voter Models kann es jedoch sein, dass die Wechselwähler auch deshalb als Zielgruppe

[7] Die beiden Topics 20 und 26 liegen bspw. sehr nah bei einander: Topic 20 beinhaltet die Posts in denen die Schlüsselbegriffe „Bürger, Demokratie, Volksentscheid, direkt, Bürgerbeteiligung“ vorkommen. In Topic 26 sind alle Posts mit den Schlüsselbegriffe „sozial, Recht, Reich, Arm, Gesellschaft“ geclustert. Weit davon entfernt ist bspw. das Topic 89, das die Schlüsselworte „Wunsch, Druck, Baum, Weihnachten, Advent“ beinhaltet und alle Posts zum Thema „Weihnachten“ zusammenfasst.

lohnend erscheinen, weil man sich einen Multiplikatoreffekt von diesen „aktiven“ Facebook-Nutzern erhofft.

Abb. 3 zeigt das Verhältnis von Wechselwählern zu der Gesamtanzahl der beobachteten Nutzer, die eine Partei gelikt haben. Dabei liegt der durchschnittliche Anteil der erkennbaren Wechselwähler bei ca. 20%. Der erhöhte Anteil der Wechselwähler unter den Anhängern der FDP kann damit erklärt werden, dass scheinbar viele Wähler von der FDP zur AfD abwandern beziehungsweise in den zurückliegenden Wahlkämpfen abgewandert sind [24].

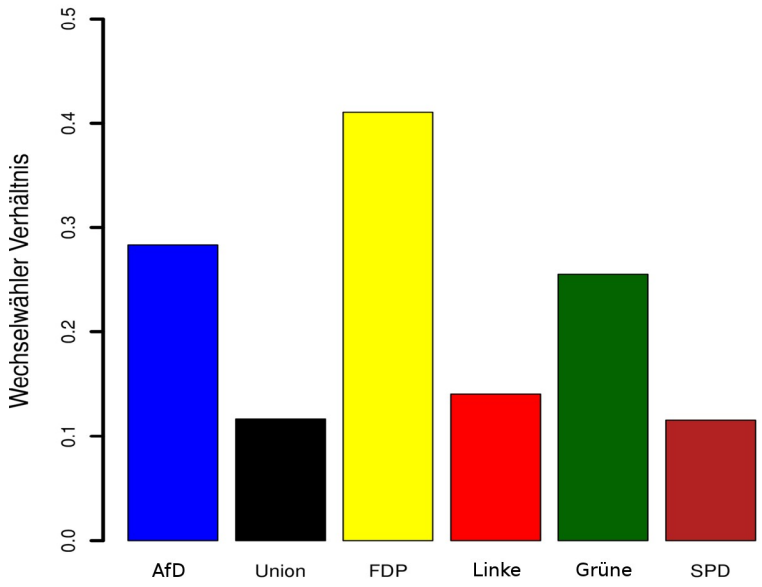

Abb. 3. Verhältnis von Wechselwählern auf den Parteiseiten

Unsere Analyse kann in zweifacher Weise für Microtargeting genutzt werden: Wir können von einzelnen Nutzern ausgehen und dann die Themen identifizieren, die sie gelikt haben. Oder wir gehen von prominenten Themen aus und schauen dann auf die Nutzer, die sich dafür interessieren. Um zu demonstrieren, wie politisches Microtargeting funktioniert, wählen wir ein beliebiges Themencluster aus: Das Cluster Nr.62 umfasst bspw. die Begriffe „TTIP, CETA, Handel, Europa, USA …“, die alle mit dem Thema „Freihandelsabkommen“ verbunden sind. Nun ist es mit Microtargeting möglich, herauszufinden welche Parteien dieses Thema auf ihrer Facebook-Seite behandeln, welche Posts sich konkret auf dieses Thema beziehen, und welche Nutzer diesen Post gelikt haben. Bei unserem Beispiel finden wir Posts aus der CDU/CSU-Koalition, die dem Topic „Freihandelsabkommen“ zugeordnet sind, und identifizieren die Wechselwähler, die diese Posts geliket haben. Aus dieser Nutzergruppe wählen wir stichprobenartig einen User aus und identifizieren alle Themencluster, in denen dieser User ebenfalls Posts geliket hat (vgl. Tab. 1). Im Beispiel handelt es sich um einen Nutzer, der auch auf Seiten der AfD geliket hat. Neben dem Thema Freihandel interessiert sich der Nutzer offenbar für Mecklenburg-Vorpommern und konservative Werte. Diese Informationen reichen für einen geübten Wahlkämpfer bereits, um eine sehr persönlich wirkende Ansprache zu verfassen und dem Nutzer personalisierte Wahlwerbung zu senden.

Tab. 1. Themencluster für Beispielnutzer und CDU/SPD- Wechselwähler

Target	Topic	Wort 1	Wort 2	Wort 3	Wort 4	Wort 5
Beispielnutzergruppe	62	Handel	CETA	Europa	TTIP	EU
	61	Alternativ	Wählen	Deutschland	AfD	Petry
	54	Mecklenburg	Vorpommern	Land	CDU	Schwerin
	10	Konservativ	Bürger	Demokratie	Politik	AfD
CDU/SPD Wechselwähler	3	Islam	Muslim	Kirche	Christlich	Glaube

Ebenso ist es möglich, Themen für relevante strategische Gruppen zu identifizieren. Die Wechselwähler, die zwischen CDU und SPD stehen, interessieren sich beispielsweise besonders für Topic 3, in dem es um Islam und Christentum geht. Diese Information kann von strategischer Bedeutung sein, wenn die eine Partei gezielt ein Thema besetzen will, das auch Wähler der anderen Partei anspricht.

Die Liste der Worte pro Topic ist im Übrigen nicht beschränkt. Darüber hinaus lassen sich alle Posts eines Topics – oder auch ein entsprechendes Sample – natürlich auch als Originaltext analysieren, um zu genaueren Einschätzungen zu kommen.

Auf Grundlage von Daten, die wir aus dem sozialen Netzwerk Facebook extrahiert haben, ist es also möglich politisches Microtargeting für deutsche Parteien zu betreiben. Aus unserem Experiment wird außerdem klar, dass die Daten, die in sozialen Netzwerken gesammelt werden, für den Einsatz von Microtargeting geeignet sind und das, obwohl die deutschen Datenschutzrichtlinien die Rückverfolgung, Aufbereitung und Benutzung von persönlichen Daten verbietet. Uns war es mit den oben beschriebenen „Machine-Learning“ Methoden und der Anwendung der benannten Algorithmen möglich, auf über 400 Facebook-Seiten zuzugreifen und 149.000 Wechselwähler zu identifizieren. Wir konnten über die Likes, die die Wechselwählern zu bestimmten Posts gesetzt haben, Themenbereiche herausfiltern, die sich für eine personalisierte Ansprache (bspw. in Form von digitaler Wahlwerbung) eignen, da sie die politischen Interessen und Präferenzen der einzelnen User widerspiegeln.

Microtargeting im Einsatz

Die hier präsentierten Ergebnisse sind als *„proof of concept“* zu verstehen. Wir haben gezeigt, dass Microtargeting auch in Deutschland eine Option ist, die künftig aller Wahrscheinlichkeit nach von politischen Parteien genutzt werden wird. Der Einsatz in einer realen Kampagne kann dabei weit über das hier vorgestellte Verfahren hinausgehen. Dabei sind zwei Entwicklungen erkennbar: Erstens sind die Parteien bereits jetzt bemüht, Onlinewahlkampf und Haustürwahlkampf zu verbinden. Die CDU hat beispielsweise mit ihrer App „connect17“ in die Verknüpfung von digitalem und analogem Wahlkampf investiert. Analysen wie die hier vorgestellte könnten also dafür genutzt werden, Wahlhelfern Hinweise zu geben, welche Themen sie bei wem ansprechen sollen. Zusätzlich ließen sich die Daten, die z. B. über Apps wie „connect17“ erhoben werden, mit Social Media Daten kombinieren – zumindest, wenn es dafür eine Zustimmung der Betroffenen gibt. Zweitens sind wir in unserem Beispiel davon ausgegangen, dass die Datenanalyse nur die Präferenzen der Wähler einbezieht. In Zukunft wird aber auch die Wirkung der personalisierten Werbung immer besser erfasst werden. So lässt sich feststellen, welche Inhalte für Online-Werbung tatsächlich funktionieren, indem Klickstatistiken und ähnliches festgehalten werden. Es ist sogar möglich, die Gestaltung der Inhalte teilweise – und perspektivisch sogar vollständig – zu automatisieren. Dadurch entsteht die Frage, ob die politischen Inhalte überhaupt noch von den Politikern erstellt werden, oder zunehmend eine algorithmische Antwort auf die Reaktionen der Wähler sind. Gleichzeitig ist aber

auch abzusehen, dass politische Parteien den Eindruck, sie stünden gar nicht wirklich hinter den von ihnen verbreiteten Themen, verhindern wollen.

Insgesamt sind der weiteren Auswertung der Daten im Prinzip keine Grenzen gesetzt. So ließe sich z. B. auch vorhersagen, welche anderen Topics für einen Nutzer interessant sind, auch wenn er sie noch gar nicht gelikt hat. Zudem weisen die Daten natürlich eine Menge zusätzlicher Dimensionen auf, die in unserem Beispiel gar nicht beachtet wurden, wie die Verteilung von Topics nach soziodemografischen Merkmalen oder die Entwicklung der Topics über die Zeit.

Fazit und Schlussfolgerung

Das „Datafication" viele bisher private Bereiche von Menschen durchdrungen hat, wird durch unser Experiment, das auf der Sammlung und Auswertung von persönlichen Daten der Nutzerinnen und Nutzer von Facebook beruht, noch einmal unterstrichen[8]. Dementsprechend ist es für uns wichtig, an dieser Stelle die ethischen Auswirkungen von Technologien, die auf Algorithmen beruhen, zu beleuchten. Neue Technologien sind immer mit neuen Herausforderungen für die sozialen Zusammenhänge und die politischen Entscheider in einer Gesellschaft verbunden: In anderen Bereichen und Einsatzfeldern von Algorithmen in sozialen Netzwerken (bspw. Social Bots) ist die Diskussion der ethischen Folgen bereits angestoßen worden [33]. Wir wollen im Folgenden erörtern, welche ethischen Fragen in Bezug auf die politische Einflussnahme durch Microtargeting aufgeworfen werden und welche potenziellen Veränderungen der politischen Landschaft in Deutschland durch diese neue Technologie vorangetrieben werden könnten.

Unsere Studie zeigt, wie „einfach" es ist, personalisierte Wahlwerbung zu erstellen, die zur Beeinflussung von Usern sozialer Netzwerke genutzt werden kann. Dementsprechend stellt sich die Frage, ob es mit Microtargeting möglich wird, die Wählerschaft in ihrer Wahlentscheidung zu manipulieren? Die Versendung von personalisierten Nachrichten oder Informationen führt erst einmal nicht direkt zu einer Beeinflussung des Wählers, da er ja frei in der Entscheidung ist, wen sie/er letztendlich wählt. Allerdings führt die o.g. Preisgabe von persönlichen Informationen, bspw. um bestimmte Serviceleistungen oder Apps nutzen zu können, mehr und mehr dazu, dass Algorithmen in der Lage sind, exaktere Prognosen über menschliches Verhalten und deren Präferenzen zu liefern. Wenn durch Microtargeting adressierte Nachrichten an konkrete Personen versendet werden, kann man von sog. „instant influence" sprechen: dem Versuch, durch einen bestimmten Reiz die Person zu einer gewünschten Reaktion zu veranlassen [29]. Dies wird noch verstärkt, wenn die Reize in kurzen Abständen auf die Person „einströmen", denn dann ist es nicht möglich, diese Reize rational zu verarbeiten [31]. Information, die so adressiert werden, führen bei den Adressaten zu einer intuitiven Verknüpfung: In unserem Anwendungsbeispiel zwischen der versandten Nachricht und dem politischen Akteur oder der Partei [28]. Eine systematische und dauerhafte Anwendung von politischem Microtargeting könnte also zu einem Zustand führen, der sich mit „progression from thought to action artificially"[11] beschreiben lässt. Instant influence lässt sich demnach nur erkennen, wenn Nutzerinnen und Nutzer von sozialen Netzwerken zwischen „normalen" Nachrichten und aus Microtargeting-Prozessen abgeleiteten Nachrichten unterscheiden können. Dementsprechend ist die Fähigkeit, Nachrichten „bewusst" in neutrale oder manipulative Inhalte zu unterscheiden, unerlässlich für eine selbstständige Meinungsbildung und die damit zusammenhängende individuelle Wahlentscheidung.

Microtargeting hat aus ethischer Sicht allerdings auch Vorteile: Die Diskussion um so-genannte Filterblasen oder Echokammern zeigt, dass die Struktur der sozialen Netzwerke nicht zu einem möglichst breit gefächerten Austausch der Meinungen führt, sondern sich politische Cluster bilden, die die Tendenz haben, sich vornehmlich in ihrer Meinung zu bestärken. Welche Rolle dabei das Verhalten der Nutzer (Echokammer) oder die Algorithmen der Plattformen (Filterblase) spielen, sei dahingestellt. Das empirisch sehr gut bestätigte Phänomen der Homophilie, also der Präferenz gegenüber Ähnlichgesinnten, stellt ein großes Problem dar, wenn die politische Meinungsbildung zunehmend über soziale Netzwerke stattfindet. Microtargeting könnte ein Möglichkeit sein, solche „starke politischen Cluster" zu durchbrechen: Mit der

[8] Gerade das Netzwerk Facebook ist ein „Ort", an dem Menschen sehr viele persönliche Informationen und Daten über sich preisgeben.

personalisierte Wahlwerbung können ggf. Aspekte und Themen betont werden, die zu einer Diversifizierung und stärkeren individuellen Abgrenzung den Nutzer führen. Gerade für die Gruppe der Wechselwähler bietet die zielgerichtet und auf die Präferenzen abgestimmte Versendung von Wahlwerbung zudem die Möglichkeit, die User für (partei-)politische Alternativen „jenseits des eigenen politischen Clusters“ zu interessieren. Eine somit erreichte, unmittelbare Beeinflussung nimmt also nicht den Umweg, „nur ein Angebot unter vielen“ zu sein, dass „ausschließlich“ für einen bestimmten Personenkreis im Netzwerk sichtbar ist [2]. Durch den grundgesetzlich festgehaltenen Auftrag, dass die Parteien an der politischen Willensbildung mitwirken, ließe sich daher sogar eine Art Verpflichtung zu Microtargeting ableiten, zumindest wenn man unterstellt, dass Meinungsbildung sonst nicht mehr funktioniert.

Ein großes Problem ist auf jeden Fall durch das Perceived Voter Model gegeben. Gerade weil eine Mehrheit der Nutzer in den sozialen Netzwerken eigentlich sehr wenig aktiv ist, besteht die Gefahr, dass sich die Politik auf die Gruppe konzentriert, die die meisten Daten hinterlässt, auch wenn das nicht repräsentativ ist [3]. Je weniger Daten man über die einzelne Person sammeln kann, umso ungenauer gerät die Einschätzung ihrer Präferenzen und umso größer ist das Risiko, eine Politik zu propagieren, die sich an einem falsch verstandenen Wählerwillen ausrichtet [29]. Wenn die Ausrichtung eines Wahlkampfes in hohem Maße oder ausschließlich auf der Auswertung von großen Datenmengen beruht, führt dies oftmals zum Perceived Voter Phänomen [21]: Alle Entscheidungen, die während einer Kampagne getroffen werden, beruhen auf Annahmen über die Wählerschaft, die von Algorithmen berechnet worden sind. Allerdings handelt es sich bei diesen Prognosen nicht um exakte Vorhersagen und die „Verzerrung“ der Wirklichkeit wird ggf. noch verstärkt, wenn die für das Vorhersagemodell benutzten Daten auf Informationen aus sozialen Netzwerken beruhen [30]. Mit der Methode des politischen Microtargeting kann es also passieren, dass sich Parteien oder politische Akteure „ihre“ Wirklichkeit und die darauf beruhende Wählerschaft „konstruieren“.

Würde man allerdings einfordern, die Parteien sollten die Wählerpräferenzen möglichst genau erfassen, so entsteht ein Zielkonflikt mit dem Grundsatz der Datensparsamkeit. Betrachtet man die Situation in den USA, so lässt sich fragen, ob Wahlen dort eigentlich noch frei, gleich und geheim sind: Die Kampagnen versuchen gezielt, bestimmte Wählergruppen zu demobilisieren, bestimmte demografische Schichten oder Wählergruppen in bestimmten Gebieten sind für den Wahlsieg wesentlich wichtiger als andere und durch die umfassende Datenerhebung – insbesondere bei der Wählerregistrierung – ist ziemlich klar, wer eigentlich was wählt. Genau diese Diskussionen kommen auch in Deutschland auf uns zu und sollten daher bereits heute diskutiert werden. Unser „proof of concept“ dient hoffentlich dazu, diese Debatte anzustoßen und zu versachlichen.

Literatur

1. Agan T (2007) Silent Marketing: Micro-targeting. Penn, Schoen and Berland Associates, New York

2. Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. Science 348:1130-2

3. Barberá P, Rivero G (2015) Understanding the political representativeness of Twitter users. Social Science Computer Review 33:712-29

4. Barberá, P, Zeitzoff T (2016) The new public address system:
Why do world leaders adopt social media?. http://pablobarbera.com/static/world_leaders_paper.pdf, 21.06.17

5. Barbu O (2014) Advertising, microtargeting and social media. Procedia-Social and Behavioral Sciences 169: 44-9

6. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research 3:993-1022

7. Boehm F (2015) A comparison between US and EU data protection legislation for law enforcement purposes. Directorate-general for internal policies. European Parliament. http://www.europarl.europa.eu/RegData/etudes/STUD/2015/536459/IPOL_STU%282015%29536459_EN.pdf, 21.06.17

8. Bond R, Messing S (2015) Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. American Political Science Review 109:62-78

9. Capara GV, Barbaranelli C, Zimbardo PG (1999) Personality profiles and political parties. Political psychology 20:175-97

10. Cialdini RB (1993) Influence: the psychology of persuasion. HarperCollins, New York

11. Däubler W, Klebe T, Wedde P, Weichert T (2010) Bundesdatenschutzgesetz. Bund, Frankfurt a. M., Aufl. 5

12. Dorschel J (2015) Praxishandbuch Big Data. Wirtschaft-Recht-Technik. Gabler (SpringerLink: Bücher), Wiesbaden

13. Downs A (1957) An economic theory of political action in a democracy. Journal of Political Economy 65:135-50

14. Directive EU (1995) 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Union L 281:31-50

15. Edsal, T (2012) Let the nanotargeting begin. New York Times Online. http://nyti.ms/QfX792, 21.06.17

16. Ellul J (1965) Propaganda: the formation of men's attitudes. Knopf, New York

17. Franz MM, Ridout TN (2010) Political advertising and persuasion in the 2004 and 2008 presidential elections. American Politics Research 38:303-29

18. Grimmer J, Stewart BM (2013) Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Political analysis 21:267-297

19. Hegelich S (2017) R for social media analysis. In: Sloan L, Quan-Haase A (Hrsg) The SAGE handbook of social media research methods. Sage, London, 486-99

20. Hegelich S, Shahrezaye M (2015) The communication behavior of German MPs on Twitter: reaching to the converted and attacking opponents. European Policy Analysis 1

21. Hersh ED (2015) Hacking the electorate: How campaigns perceive voters. Cambridge University Press, New York

22. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. Journal of educational psychology 24:417-441

23. Mayer-Schönberger V, Cukier K (2013) Big data: A revolution that will transform how we live, work, and think. Hachette UK, London

24. Niedermayer O, Hofrichter J (2016) Die Wählerschaft der AfD: Wer ist sie, woher kommt sie und wie weit rechts steht sie?. ZParl Zeitschrift für Parlamentsfragen 47:267-85

25. Nulty, P, Theocharis Y, Popa SA, Parnet O, Benoit K: (2016) Social media and political communication in the 2014 elections to the European Parliament. Electoral Studies 31:429-44

26. Panagopoulos C (2016) All about that base: changing campaign strategies in US presidential elections. Party Politics 22:,179-90

27. Parliament E. Directive (2002) 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector. Off. Official Journal of the European Communities. Official Journal of the European Union L 201:37-47

28. Piaget J (1950) The psychology of intelligence. Routledge & Paul, London

29. Persily, N (2017) Can democracy survive the internet? Journal of Democracy 28:63-76

30.Ruths D, Pfeffer J (2014) Social media for large studies of behavior. Science 326:1063-4

31. Simon HA (1996) The sciences of the artificial. MIT press, Cambridge, Massachusetts

32. Sotto LJ, Simpson AP (2014) United States. In: Data Protection & Privacy (s. 191-–204), Law Business Research Ltd

33. Thieltges A, Schmidt F, Hegelich S (2016) The devil's Ttriangle: ethical considerations on developing bot detection methods. 2016 AAAI Spring Symposium Series, AAAI, North America

34. Woo HY (2015) Strategic communication with verifiable messages. PhD Thesis. University of California, Davis

3.2 Social media and microtargeting: Political data processing and the consequences for Germany

Authors

Orestis Papakyriakopoulos , Simon Hegelich, Morteza Shahrezaye, Juan Carlos Medina Serrano

In

Big Data & Society, July–December 2018: 1–15, DOI: 10.1177/2053951718811844

Abstract

Amongst other methods, political campaigns employ microtargeting, a specific technique used to address the individual voter. In the US, microtargeting relies on a broad set of collected data about the individual. However, due to the unavailability of comparable data in Germany, the practice of microtargeting is far more challenging. Citizens in Germany widely treat social media platforms as a means for political debate. The digital traces they leave through their interactions provide a rich information pool, which can create the necessary conditions for political microtargeting following appropriate algorithmic processing. More specifically, data mining techniques enable information gathering about a people's general opinion, party preferences and other non-political characteristics. Through the application of data-intensive algorithms, it is possible to cluster users in respect of common attributes, and through profiling identify whom and how to influence. Applying machine learning algorithms, this paper explores the possibility to identify micro groups of users, which can potentially be targeted with special campaign messages, and how this approach can be expanded to large parts of the electorate. Lastly, based on these technical capabilities, we discuss the ethical and political implications for the German political system.

Contribution of thesis author

Theoretical design, model design and analysis, manuscript writing, revision and editing

Social media and microtargeting: Political data processing and the consequences for Germany

Orestis Papakyriakopoulos, Simon Hegelich, Morteza Shahrezaye and Juan Carlos Medina Serrano

Abstract

Amongst other methods, political campaigns employ microtargeting, a specific technique used to address the individual voter. In the US, microtargeting relies on a broad set of collected data about the individual. However, due to the unavailability of comparable data in Germany, the practice of microtargeting is far more challenging. Citizens in Germany widely treat social media platforms as a means for political debate. The digital traces they leave through their interactions provide a rich information pool, which can create the necessary conditions for political microtargeting following appropriate algorithmic processing. More specifically, data mining techniques enable information gathering about a people's general opinion, party preferences and other non-political characteristics. Through the application of data-intensive algorithms, it is possible to cluster users in respect of common attributes, and through profiling identify whom and how to influence. Applying machine learning algorithms, this paper explores the possibility to identify micro groups of users, which can potentially be targeted with special campaign messages, and how this approach can be expanded to large parts of the electorate. Lastly, based on these technical capabilities, we discuss the ethical and political implications for the German political system.

Keywords

Microtargeting, social media, Germany, influence, datafication, electorate

Introduction

The contemporary digital revolution is constantly transforming the political world. Datafication (Mayer-Schönberger and Cukier, 2013), i.e. the categorization, quantification and aggregation of phenomena into databases, and their further algorithmic processing, have opened new opportunities in understanding and evaluating complex social phenomena. More specifically the use of social media and the internet has resulted in the creation of enormous databases that contain information about citizens' personal and political preferences. Based on these *Big Political Data* a new type of data-driven interaction between politics and citizens emerges through social media. In its core lies the application of advanced statistical and machine learning algorithms, the possibilities of which enable the development of new political strategies. Consequently, political actors have started using newly developed tools in order to analyse citizens' behaviour and to influence the electoral body. One of these methods is microtargeting, which allows the formulation of personalized messages and their direct delivery to groups and individuals (Agan, 2007), hence creating a promising tool for electoral campaigning and opinion formation.

In this paper, we demonstrate a *proof of concept* regarding the ways political actors could establish the conditions for political microtargeting in Germany, through the utilization of social media platforms.

Bavarian School of Public Policy, Technical University Munich, Germany

Corresponding author:
Orestis Papakyriakopoulos, Bavarian School of Public Policy, Technical University Munich, Richard-Wagner-Strasse 1, 80333 Munich, Germany.
Email: orestis.papakyriakopulos@tum.de

The scope of our analysis is to identify the possibilities and dangers of microtargeting in electoral campaigning, taking into consideration 'state of the art' technology. Therefore, we apply our method to Facebook data that could actually be used in political campaigning. Initially, we explain the theory behind microtargeting and discuss existing obstacles that prevent its application. Second, we illustrate our methodology and present our results. Lastly, we evaluate data-driven microtargeting ethically and comment on its political consequences.

Microtargeting in theory

Microtargeting is a strategic process intended to influence voters through the direct transmission of stimuli, which are formed based on the preferences and characteristics of an individual. First of all, microtargeting presupposes the collection of large amounts of data able to depict the political preferences and other non-political characteristics of voters. This data can be either manually collected or acquired through data-mining and can include information ranging from a person's name, address, and voting history to more abstract properties such as a person's opinion about political and non-political topics, their social activity and cultural background. The gathered data are then processed with the aid of appropriate *machine learning* algorithms, while the acquired results depend on the type of algorithm used. It is then possible to make predictions about specific variables, for example, the outcome of a political decision (supervised learning) or identification of patterns in the data through clustering (unsupervised learning). Implementing the latter, political actors are in a position to detect sub-groups of voters that share common demographic and attitudinal traits (Barbu, 2014). Based on the algorithmic results, they can then generate messages or plan actions aimed at influencing each specific sub-group or person (often called nanotargeting (Edsall, 2012)), leading to their potential mobilization or de-mobilization.

Microtargeting was first applied to a limited extent in the US 2000 Federal Elections by the Republican Party (Panagopoulos, 2015). Since then, the increasing datafication of societies has provided fertile ground for its expansion as a political strategy. A milestone for its application was the 2008 Federal Elections (Franz and Ridout, 2010), when the Democratic Party campaign applied the strategy at full scale. Today, microtargeting is a standard online and offline (Panagopoulos, 2015) campaigning method in the US as it overcomes problems of classical political campaigning. First of all, it has the potential to partly track the predispositions or general interests of a voter (Ellul, 1966), and based on them, to modify the candidates' public images in a way that complies with the voters' opinions (Bond and Messing, 2015; Capara et al., 1999). Furthermore, by directly communicating individual- or group-specific messages, candidates are able to reduce the risk of alienating other voters that might disagree on a topic (Woo, 2015). Another advantage is that microtargeting allows political actors to target voters from the entire political spectrum, rather than exclusively developing their campaign on the characteristics of the *median voter* (Downs, 1957), as was the case in the past. Finally, given that opinion polls in the 2016 US and 2017 German elections failed to make plausible forecasts of election results, microtargeting provides a methodology to overcome political decisions based solely on survey polls. Despite the above advantages, it is important to note that there is no comprehensive study that proves the effectiveness of microtargeting (Jungherr, 2017; Karpf, 2016); to date it remains a promise emerging from the technological state of the art.

One of the main reasons behind the success of microtargeting in the US is the loose legal framework, which allows political actors to almost freely create, acquire and use databases that contain personal information. It is characteristic that there is no dedicated data protection law or a concept of 'sensitive' personal data in the US legislation. Hence, there is no general legislative framework exclusively dealing with the protection of a person's privacy rights (Sotto and Simpson, 2015). Although legal frameworks, as the FTC, ECPA, HIPAA, etc., indeed aim to regulate the monitoring of personal data and their protection in their respective fields, the administration of data policies takes place usually only indirectly, by laws that might impose purpose limitations or time limits on the data retention (Boehm, 2015). Furthermore, the US law presents significant gaps concerning the protection of individual privacy (Ohm, 2014): e.g. the datafication or reuse of information acquired as a by-product of providing services is largely unregulated (Strandburg, 2014: 22). Consequently, such legal inconsistencies facilitate the development of huge political databases, which can then be used for political campaigning (Bennett, 2016).

Contrary to the US, the legal framework applicable in Germany significantly limits the potential of microtargeting. Germany's privacy law complies with the EU-directive on the processing of personal data. The General Data Protection Regulation (EU-Directive, 2016) provides an extensive regulatory framework for

the protection of privacy and personal data, their acquisition, use and exchange. The GDPR thoroughly describes the limits and responsibilities of data controllers and processors, supports the subjects' rights to privacy and consent, and stipulates the exact regulating role of public authorities. Furthermore, the German data protection law explicitly defines the conditions and cases in which someone is able to access and use personal data (Däubler et al., 2016) and lays down the rights of persons affected (Broy, 2017), strongly limiting data exploitation.

Barrier 1: Privacy and data protection policy

Some authors[1] have argued therefore that microtargeting cannot be applied in German politics. However, despite the legal restrictions, there is ample leeway for it on social media platforms (Papakyriakopoulos et al., 2017). The reason is that the German privacy law permits the collection and processing of public personal data stemming from social media, as long as the individuals' interests are not challenged (Dorschel, 2015). The GDPR clearly states that given the appropriate safeguards, personal data on political opinion can be used for electoral activities (EU-Directive, 2016: 11). In addition, users on social media services consent to companies using their personal data for commercial and other activities, by opting in. Hence, the legal requirements for using social media data as basis for political microtargeting are met. Given the fact that users agree to publish on social media a huge amount of data about their political and non-political preferences and behaviour, these platforms are an ideal source for political knowledge extraction. Social media have become a key environment for political campaigns, as the majority of politicians can use them to communicate directly with the electoral body (Barberá and Zeitzoff, 2017; Hegelich and Shahrezaye, 2015; Medina Serrano et al., 2019; Nulty et al., 2016; Stier et al., 2017). That aside, political actors often perform organized influencing strategies on social media, frequently trespassing the legal limits set (Weedon et al., 2017).

Barrier 2: Data bias

The legal framework is not the only obstacle for successful microtargeting. The type of data subjected to algorithmic process and their entailed results can sometimes lead to spurious political action. In our case, the world of social media is not identical to the offline world. Hence, political preferences appearing on social media platforms cannot be assumed to be the same for the actual electorate. The politically active user population on Facebook is in no way representative of the whole population of a country (Ruths and Pfeffer, 2014), while the expression of an opinion online does not fully correspond to a coherent political statement (a like is not a vote; Hegelich and Shahrezaye, 2015). Furthermore, the evaluation of social media data is bound with multiple methodological issues (Hegelich, 2017). Still, the case of the United States has shown that political campaigning is more than ever based on data, from which an electorate's image is derived, also known as *perceived voter model* (Hersh, 2015). This model may be misleading but nevertheless used, as it reduces the complexity in campaign decision-making. Due to the fact that it is almost impossible to causally link a campaigning tool to election results, microtargeting is used as long as it is assumed to have a successful influence – even if in reality it might not. The difficulties in causal inference arise – amongst others – from potential self-fulfilling prophecies: should a campaigning tool identify a target group, the campaign will increase interaction with this group. This special attention might yield positive results; but these results could have also been the same for a totally different group, as well. Despite the above, microtargeting is applied, even if it might be epistemologically impossible to evaluate its exact impact.

Data and method

In this paper, we demonstrate how politicians in Germany can create the conditions for microtargeting based on data from the social media platform *Facebook* and we evaluate its ethical and political consequences. Facebook was chosen as a data source for three reasons: (1) the German Facebook population is larger and less selective than that of Twitter. (2) It is part of the company's business model to offer targeted advertisement services for political campaigning, the possibilities of which we are exploring. (3) Contrary to the US, where there are extensive political databases with personal identifiers (Bennett, 2016), in Germany this is not the case. Hence, social media provide a straightforward way to acquire knowledge for microtargeting.

For our *proof of concept*, we analysed the public Facebook pages of the German political parties and their supporters: Our sample includes the following parties: Christlich Demokratische Union (CDU), Christlich Soziale Union (CSU), Sozialdemokratische Partei Deutschlands (SPD), Bündnis 89/ Die Grünen, Die Linke, and Alternative für Deutschland (AfD).

CDU is the main conservative party of Germany, while CSU is the conservative party active in Bavaria. SPD represents the main German social-democratic party, and Die Linke the radical left. AfD has a nationalist, anti-immigrant and neo-liberal agenda, while FDP is a conservative, neo-liberal party. Finally, Bündnis 90/Die Grünen is the German green party.

For each political page, we evaluated user "Likes" on political posts and assigned a partisanship to each user according to their preferences (Figure 1). Following a standard microtargeting technique, we focused our study on users who have liked content on pages of more than one political party. The reason behind this decision is that the specific group of voters, also named as cross-pressured partisans, has the highest likelihood to be influenced, as they are both undecided and engaged in politics (Ellul, 1966; Hersh, 2015). After identifying the relevant groups, we applied machine learning algorithms to cluster the various pages' posts and created a mapping of 55 different topics, to which each of the posts might be assigned. To achieve this, we performed topic modelling analysis by applying a *Latent Dirichlet Allocation* algorithm (Blei et al., 2003). In this way, we demonstrate how someone can detect individual political topics of interest and how these can be later used to shape targeted messages for each micro group of users.

Prerequisite for the application of microtargeting is the existence of a rich database containing voters' characteristics and preferences. Therefore, we mined data from 570 public pages related to the major political parties in Germany through the Facebook Graph API, and analysed posts and Likes. We selected the pages by searching the respective party names in the name field of the Facebook pages. We then classified manually our results, and removed irrelevant pages.[2] We mined every post generated by the administrators of the pages since their creation, the Likes each post got, and the unique IDs and profile names of the users liking them. Usually, the profile name of a Facebook account tends to be the same with the real name of the account holder, as Facebook maintains a real name policy.[3] In total, we collected 251,947 posts with 6,347,448 Likes related to them and identified the activity of 1,208,740 unique users. This is only data related to the pages mined, hence the actual size of trackable users is even larger. We define a user who has liked at least one post of a party as partisan, and a user who has liked posts on pages of two or more parties as a cross-pressured partisan. Of course, the act of liking per se does not make someone a party partisan, but in this case it provides a plausible classification method for the users. Furthermore, it does not distort the microtargeting process, as microtargeting targets the identification of voter's predispositions and not to definitely certify someone's exclusive support to a party. As shown in Figure 1, around 50% of the active users per party have made only one Like. This is typical of Big Data applications on social media phenomena, where the information for the majority of users is low.

Along with the identification of potential cross-pressured partisans, we wanted to identify the specific content that they find interesting. Therefore, we applied the LDA topic modelling algorithm (Blei et al., 2003) to classify 251.947 posts. LDA has many advantages over other standard text-mining algorithms (Grimmer and Stewart, 2013), as it can recognize complex relations in text-datasets. The algorithm has the ability to cluster posts in a certain number of topics, where each topic is a set of words that characterize different contents. Hence, someone can evaluate all the posts without having to investigate them one by one. LDA assigns a probability for each post belonging to a specific topic. Then, by ascribing to each post the topic with the highest probability and by detecting the users who liked it, we can explicitly track the topics that each user is interested in.

The LDA algorithm is a three-level hierarchical Bayesian model that predicts the probabilities of words and documents belonging to a number of topics K given the empirical distribution of words (or n-grams) in a corpus (Blei et al., 2003, 2002). In our case, the corpus consists of the total number of posts M under investigation, while each post corresponds to a document d, which is a sequence of N_d words. LDA is a generative model, i.e. it assumes the probability distributions of topics over words β_k, of documents over topics θ_d and predicts the probability that a specific word in a specific document will belong to a specific topic. This Bayesian admixture can be described by the following probability distributions

$$\theta_d \sim Dir_K(a)$$

$$\beta_k \sim Dir_V(\eta)$$

$$z_w \sim Multinom_K(\theta_d)$$

$$w \mid z_w \sim Multinom_V(\beta_k)$$

where V is the number of unique words existing in the corpus, and α and η are Dirichlet parameters. Multinomial distribution z_w gives the probability that

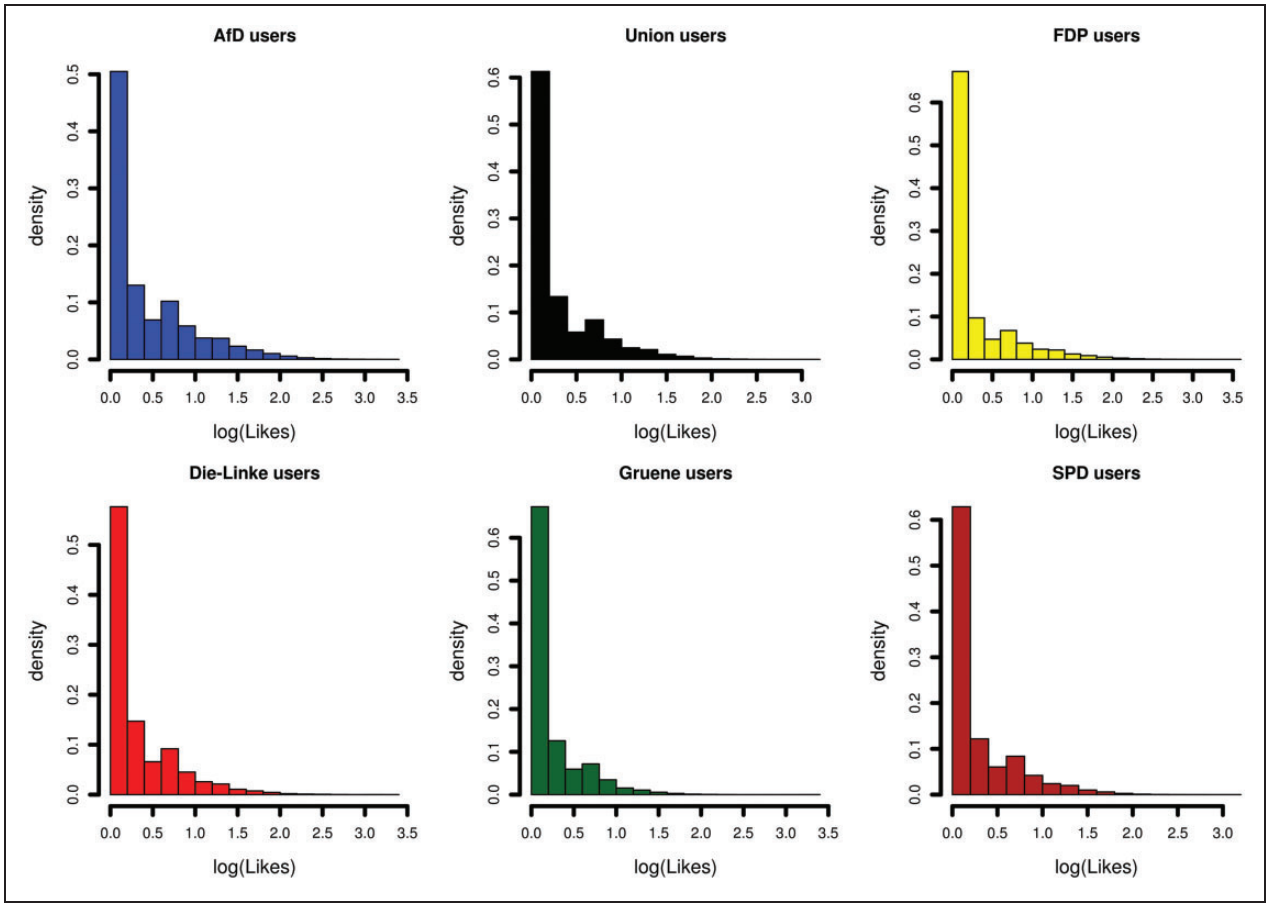

Figure 1. Likes distribution for the users on parties' pages.

a topic will be assigned to a word, given the distribution of topics over documents. Finally, multinomial distribution $w \mid z_w$ gives the probability that the model will generate a specific word in a specific document given a topic (Figure 2).

In our case, we want to create topics about the content of our corpus based on the empirical distribution of words over documents. Given the complexity of the model and the fact that the initial distributions are assumed and not empirically provided, we randomly assign topics to words and documents and we follow a Markov chain Monte Carlo procedure to update their values (Griffiths, 2002). By iteratively applying a Markov chain, we can converge to the assumed distributions and hence sample from them (Gilks et al., 1995; Roberts and Smith, 1994) the probability $P(z_w|w)$ that a word in a document belongs to a specific topic. More specifically, we used a collapsed Gibbs sampling Marcov chain Monte Carlo (MCMC) (Geman and Geman, 1984) method to identify the relevant topics. The specific algorithm comes with the advantage of integrating out the probability distributions β_k, θ_d (Darling, 2011). Thus as part of the iterative Markov chain, one can calculate the targeted probabilities through the process

$$\text{for each document} : d_i = (1 \ldots M)$$
$$\text{for each term in a document } i = (1 \ldots N_{d_i})$$
$$P(z_i = j \,|\, z_{-i}, V) = \frac{v_{-i,j}^{w=i} + \eta}{\sum_{w=1}^{V} v_{-i,j} + V\eta} \frac{n_{-i,j}^{d_i} + \alpha}{\sum_{k=1}^{K} n_{-i}^{d_i} + K\alpha}$$

where i is the concrete appearance of a word, $-i$ denotes its exclusion and j is a topic. $v_{-i,j}^{w=i}$ corresponds to the number of times word i is assigned to topic j, without its current appearance and index $\sum_{w=1}^{V} v_{-i,j}$ gives the total number of words in the corpus assigned to topic j excluding i. Furthermore, $n_{-i,j}^{d_i}$ contains the total number of words in document d_i that are assigned to topic j without i. Finally, $n_{-i}^{d_i}$ corresponds to the

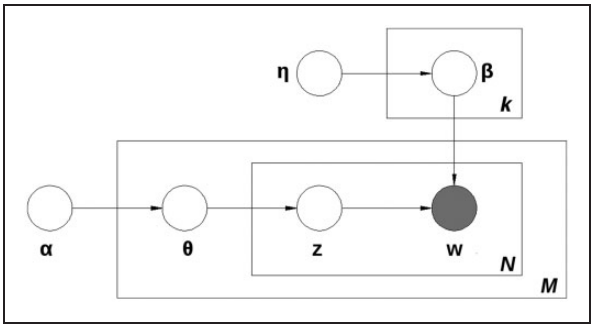

Figure 2. Plate notation for the Latent Dirichlet Allocation algorithm.

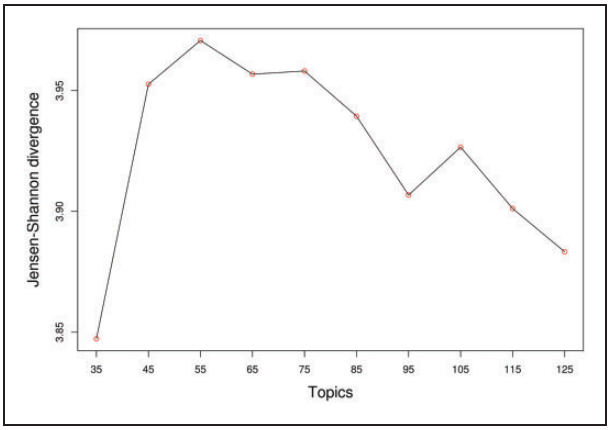

Figure 3. Topic optimization process. The model with the highest Jensen–Shannon convergence contained 55 topics.

total number of words in the document, again not including *i*.

Necessary for the creation of a useful LDA model is the election of an appropriate number of topics, in order to split the content into interpretable subgroups. Electing a small number of topics results in a clustering of posts, from which one cannot identify concrete political topics of interest. On the contrary, if the number of topics is too large, the algorithm selects many words as topic-important that actually have no political value. To overcome this issue, we applied a topic optimization algorithm proposed by Deveaud et al. (2014). More specifically, we calculated the Jensen–Shannon divergence between topics for multiple LDA models through the equation

$$D(k_i, k_j) = \frac{1}{2}\sum_{w=1}^{V} \beta_{i,w}\log\left(\frac{\beta_{i,w}}{\beta_{j,w}}\right) + \frac{1}{2}\sum_{w=1}^{V} \beta_{j,w}\log\left(\frac{\beta_{j,w}}{\beta_{i,w}}\right)$$

where *i, j* are two different topics in a model and $\beta_{i,w}, \beta_{j,w}$ the probability density values of the distribution β_k for a word w in the corpus V and each topic, respectively, then selected the model that maximizes the sum of the Jensen–Shannon divergence for all topic combinations given the expression

$$K_{opt} = \text{argmax}\frac{1}{K(K-1)}\sum_{k_i, k_j=1}^{K} D(k_i, k_j)$$

Based on the optimization process (Figure 3), we concluded on an LDA model with 55 topics. In order to sort and visualize topics according to their similarity, we used the method proposed by Sievert and Shirley (2014). We used the already calculated Jensen–Shannon divergences for the unique 1485 topic combinations and created a distance matrix. On it, we applied a principle component analysis algorithm (Hotelling, 1933) and we plotted the first three components.

Results

The first result of our analysis was the specification of the political content of the investigated posts. The LDA algorithm clustered the posts in 55 topics that can be split into three main categories. These categories were chosen manually, and do not denote that they are the optimal ones; still their election makes the results much more interpretable.[4] The first category includes topics related to general political issues, such as social involvement (topic 1), education (topics 2, 15), national economy (topic 4) and homeland security (topic 32). Some topics do not only illustrate the relevance of posts to a political issue, but also the exact opinion underlying them. For example, topics 10 and 12 are both migration related, but topic 10 includes posts that are refugee-friendly, while topic 12 contains posts that demand a stricter migration policy. In addition, there are topics that analyse political parties (topic 39) or persons (topic 38). In the same category, also exists a set of topics (9, 27, 14) that contain posts that do not make concrete political statements, but declare uncertainty and reflection.[5] The second category includes topics that are related to political actors and candidates, but not as part of a political discussion. They summarize posts about political events, media appearances and electoral campaigning. Finally, the third category contains topics that are location related and discuss political problems about regions. For example, topic 54 includes

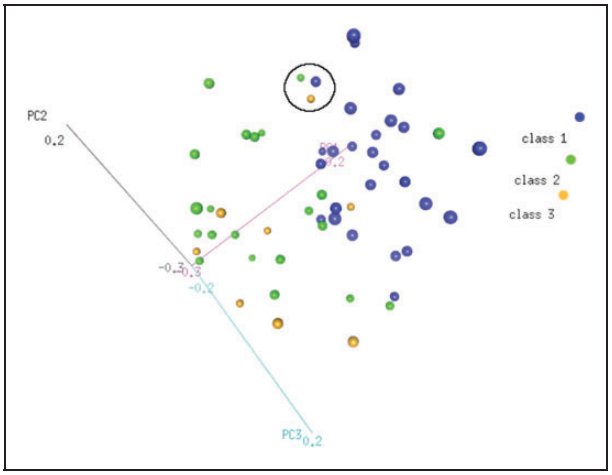

Figure 4. Topic distance visualization with the help of PCA. Circled are topics 21, 43, 38.

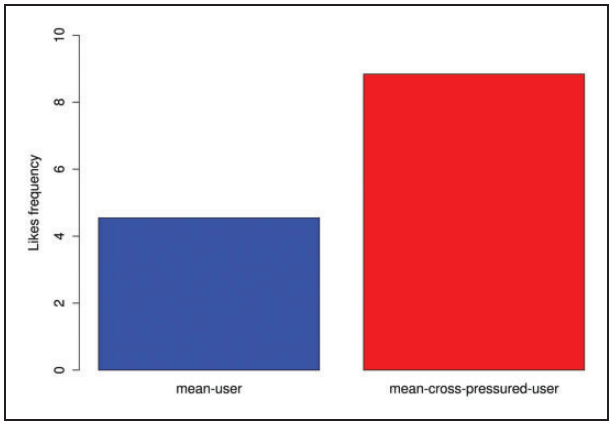

Figure 5. Average Likes frequency for the mean and the cross-pressured user.

posts about Berlin, topic 31 about Hamburg and topic 43 about Bavaria.

In order to evaluate and verify our topic classification, we visualized the relationship between the developed topics in a three-dimensional space with the help of PCA (Figure 4). Each sphere corresponds to a different topic, while their size is proportional to the number of posts they contain. Their distance in 3D-space functions as a measure of their content similarity. It is visible that three categories classify topics into unique clusters. As expected though, there is some overlapping between categories, as a topic might contain keywords belonging to more than one categories. For example topics 21, 43, 38 appear very close, even though we classified them differently (Table 1). This occurs because they all include a combination of posts of all classes. Topic 21 is about AfD, including both posts about its political background and the elections. Topic 38 is about Angela Merkel and her political activity, as well as her party structure. Finally, topic 43 is about Bavaria, including a number of posts about the regional CSU party and its candidates.

In our analysis, we identified a total of 58,532 cross-pressured users. Figure 5 shows that cross-pressured users tend to like more frequently than the average Facebook partisan. This however does not mean that cross-pressured partisans tend to be more active; on the contrary, it denotes that we can only trace cross-pressured partisans, when the users are more active online. This has an important implication for the perceived voter's model: The selection of cross-pressured partisans as targeted population comes with the advantage that they behave as multiplicators, and thus their potential influence will contribute to the motivation of other users as well.

Figure 6 shows the ratio of cross-pressured partisans between parties. In the given dataset, more or less 10% of the page users for each party are cross-pressured. This does not mean though, that this number corresponds to the actual electorate, as the descriptive results are biased through our statistical sample and the structure of the social media platform. Nevertheless, it is possible to recognize certain predispositions of the electorate, as for example an increased interaction of Union and FDP users and the almost non-existent overlap of users that are interested in both Die Grünen and AfD.

After the concretization of the topics of interest, microtargeting can be performed in two ways: one can either initially focus on single users and then track afterwards the topics they are interested in, or select specific topics and then identify users interested in them. To demonstrate how further steps of the microtarging process could be realised, we choose randomly topic 4 as an example. Topic 4 includes, amongst others, the words: Euro, Steuergeld, Milliarde, Zuschuss, Kosten, i.e. it is linked to German economic policy. It is possible to analyse the relevance of this topic for each party, as well as to identify users who like the topic. In this case, we find Union coalition posts that talk about the German economy and identify the relevant cross-pressured partisans. Then, we randomly pick one of the users to investigate all the other topics that are of interest to her. Our random cross-pressured user has also liked FDP posts, and as Table 2 shows, she has also expressed interest in political issues of Schleswig Holstein and homeland security. Hence, we can identify significant political topics of

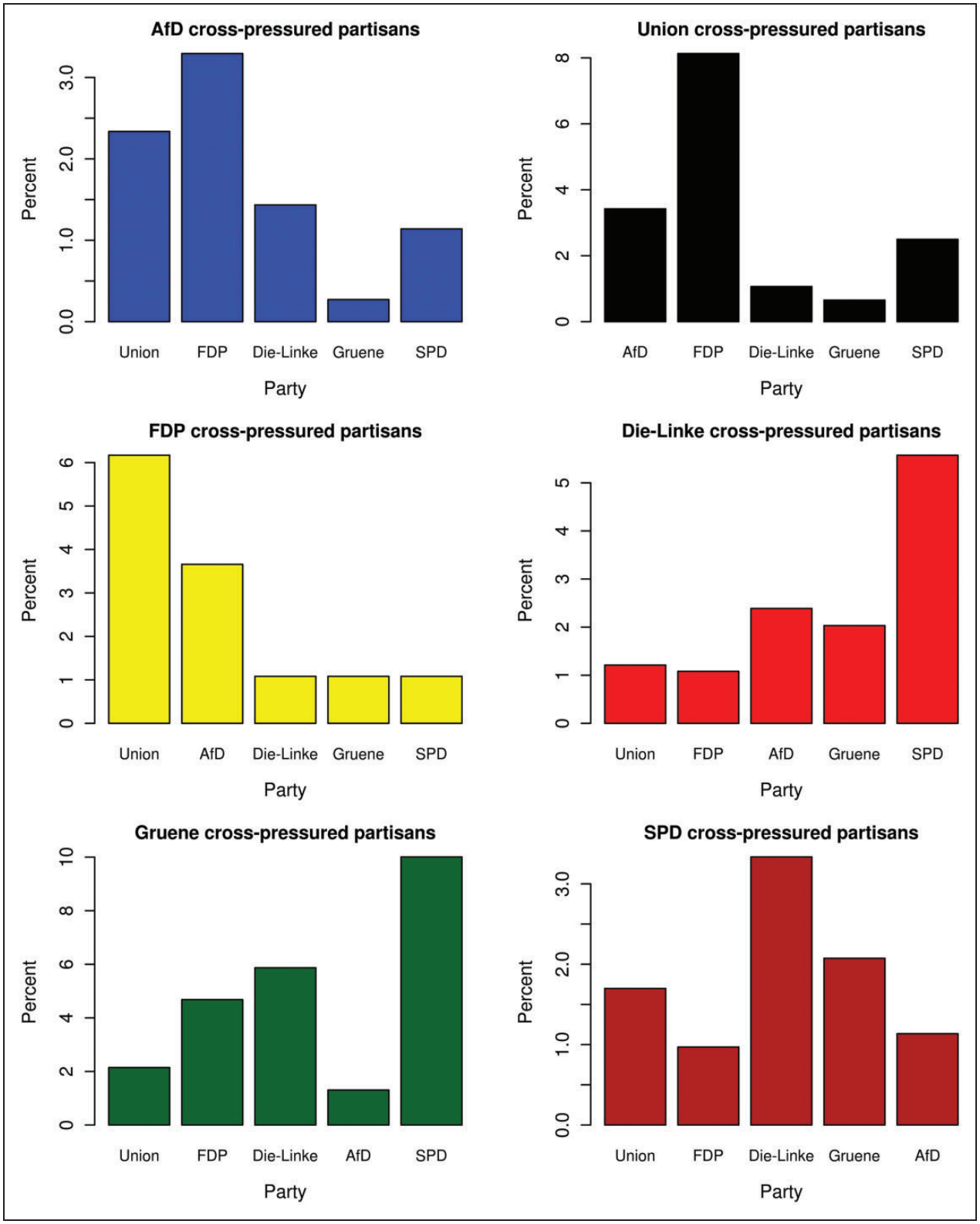

Figure 6. Percentage of cross-pressured partisans per party.

Table 1. Extended keywords for topics 21, 38, 43.

Topics	Extended keywords
21	AfD, Partei, rechtpopulist, Position, Altparteien, Wahlen, Argument, Stimmen, vertreten, Gegner
38	Merkel, Angela, Kanzlerin, CDU, Union, CSU, Seehofer,Volk, Fluchtlingspolitik, Terroranschlag
43	Bayern, München, Freistaat, Wahlprogramm, CSU, muss, Regierung, Generalsekretarin, Schalzwedel, Grüne

Table 2. Topics of interest for an example-user and for Union-SPD cross-pressured users.

Target	Topic	keywords				
Example user	4	Euro	Steuergerl	Milliar	Zuschuss	kosten
	32	Innenminister	Polizei	ermittelt	Justitz	Kriminalität
	51	Schleswig	Holstein	Kiel	Rostock	Schwerin
Union-SPD cross-pressured users	8	Islam	Muslim	Christlich	Religion	Kirche

interest for the user, as well as political parties to which, the user is positively inclined.

The topic modelling algorithm, however, does not illustrate if the user thinks positively or negatively of a political topic, i.e. it does not trace their exact political attitudes to the issues. To do this it would be necessary to apply a sentiment analysis algorithm to the parties' posts, or a qualitative analysis thereof. In the current research, we did not perform a sentiment analysis. Given the results of the sentiment analysis, the person's political evaluations of political topics and party sympathy, a campaign-maker has adequate information to create personalized messages and communicate them through micro-targeted advertisement.

Similarly, it is possible to identify topics that are important for groups of strategic importance. For example, partisans that are cross-pressured by the Union and SPD are highly interested in topic 8, which is related to Islam and Christianism. Thus, after the combination with a sentiment analysis, the creation of an advertisement specifically related to this topic can provide additional advantage to a political party, as it might mobilize an important part of the electorate towards its ends. Of course, the content of a personal message can be further specialised, as it is always possible to access recursively the full post that a user liked, and locate exactly its content in relation to the topic it belongs to.

Given the mined Facebook data, we proved that there is an extensive dataset for potential microtargeting in German politics available in social media services. Although national privacy regulations usually forbid the direct acquirement and use of personal data, data existing on social media platforms provide a fruitful source for microtargeting. By mining and structuring the content of 570 German political pages, we managed to detect over 58,000 cross-pressured users through their Likes. The selection of this sub-population was based on the idea that they are people both active in politics and potentially undecided on their exact party preference. Hence, communicating a message to them is of greater value than to people who are strict supporters of one party or are not interested in politics at all. In order to track topics of interest of cross-pressured users, we applied simple machine learning algorithms on the pages' content and found the most common issues discussed. Finally, we connected the topics with the users through their posts' Likes, finding out valuable political information about them. Accompanied with a sentiment analysis algorithm, the necessary knowledge can be gathered for the creation of personalized messages. Last step is to contact the users, a process that should be adapted to and compliant with the legal frameworks.

The communication of the message could theoretically be performed in two ways: One could cluster users sharing common characteristics and directly target them through the platform's advertisement service, which allows campaigners to define custom target audiences. This comes with the advantage that there is no need for manual matching of users to their real world identities, as it suffices to communicate the message to them through the platform. The second way is to manually look at a person's further public activity on Facebook, and given additional sociodemographic data available, try to find another communication path (e.g. email, mail, phone number, etc.). Although the second way is time-

consuming, complicated, and sometimes inadequate, gathering socio-demographic data about individuals and then targeting them offline is actually what is intensively done in US campaigns (Hersh, 2015: 77). Still, in EU the feasibility of the strategy is much lower, due to the existing privacy laws. For the second way to be applicable, political actors should develop platforms, applications, or services, through which they would get the person's consent to target them with the related messages.

The processing of the social media political dataset also comes with specific limitations. The inferences drawn reveal only part of a person's political characteristics, and only if indeed someone's online behaviour matches their actual political preferences. Furthermore, the users detected online might not have a voting right in Germany, making the sampling process biased and distorting the advertisement process.

The presented results serve as a proof of concept. We have thoroughly described how microtargeting based on social media data could be performed. The analysis was focused on Germany, where the acquisition of relevant data is usually problematic. The described method can be extended through further actions in both online and offline campaigning. For example, parties have already started promoting apps to connect the digital and analogue campaigning.[6] These apps help to analyse the reactions of people, giving feedback to the campaign-managers about their campaigning tactics. Furthermore, the combination of the app data with data coming from social media can provide even more insights on the relevant issues. The processed social media data can also be used to complement standard opinion prediction techniques. Existing census data about demographic characteristics and public record data about past voting behaviour can be combined with results from the topic modelling and sentiment analysis algorithms and hence explain the features of political behaviour.

In our study, we focused only on the detection of voters' political topics of interest, however part of the microtargeting process is also the evaluation of the personalized advertisement's success. This can be done after the first application of microtargeting, through analysis of click-statistics, performance of surveys and the actual election results. Furthermore, after the calibration of the process, the generation of microtargeting data can be highly automated. This of course raises the question of whether politicians' positions would still be a result of their actual opinions or just an algorithmic creation for attracting voters. Finally, machine learning algorithms can predict the users' interest in further topics or parties, even if they have not liked them on the platform. Further data would be required for this, which in this case were not taken into consideration, but are still publicly available online (Kosinski et al., 2013). By collecting data from other social media interactions, e.g. likes on news media or other non-political pages, one can train models and assign probabilities of someone being interested in a political issue or party. In this way, political knowledge can be extracted about users that actually did not actually interact with any party-related content on the platform and hence be included as audience of political microtargeting.

Discussion

The penetration of datafication into people's privacy is once more proven through our investigation, as we were able to gather and process a large amount of user data from the social media platform Facebook. Hence, from our perspective, it is important to evaluate the impact of the latest technological advances on the ethical and political life of our society. The discussion that has already started regarding the application of data-intensive algorithms to social networks (e.g. social bots (Thieltges et al., 2016), using algorithms for social engineering (Strohmaier and Wagner, 2014)), must now be also extended to the effect of microtargeting as a technology driven campaigning method. As the new technological capabilities raise questions regarding the limits of ethical political influence and the potential transformation of political behaviour in contemporary society, our task is to identify and reflect on the newly emerged issues.

The study showed, that through machine learning, it is possible to track someone's interests and subsequently develop personalized political advertisement that can be used to influence social media users. Hence, the first question emerging is whether microtargeting might lead to the manipulation of voters. The transmission of a personalized message does not per se signify the manipulation of a person, as each individual possesses the freedom to decide whom to vote for. As the public is offering more and more voluntarily their information in exchange for online or offline services (Barbu, 2014) though, algorithms tend to become more precise in evaluating personal preferences and attitudes. As microtargeting could potentially contact the person directly with a very well adapted message, it might achieve what is called instant influence: trigger the person's mind to develop a conditioned response the way the political actors desire (Cialdini, 2007). This happens, because in cases of fast incoming information stimuli, the individual does not process them rationally (Simon, 1996). On the contrary, the

information is assimilated intuitively, creating a phenomenalist connection between the message and the political party (Piaget, 1947). Of course, framing a party successfully also presupposes other psychological, social and political preconditions to be present (Domke et al.. 1998; Schmitt-Beck, 2003), which cannot be formed by simply sending well-adapted personal messages. But given these conditions, a systematic application of microtargeting might lead to a 'progression from thought to action artificially' (Ellul, 1966). A reaction to this issue is the conscious understanding of the person that they are being microtargeted. In this way, they would be in position to evaluate a message totally differently, knowing that the incoming stimuli are already adapted to their own attitudes. The rule of the conscious over the unconscious is a precondition for the society to remain autonomous (Castoriadis, 1997).

This type of consciousness is not only needed at the moment of evaluating a political message, but must also exist at the level of privacy. It is common that through the use of apps and online platforms, people voluntarily provide their personal data and allow their further usage as a by-product of the service. It is important for users to become aware of what they are agreeing on, and what consequences their actions have. In this direction, certain normative and legal imperatives have already been formulated: Transparency of data collection, processing and application (Barocas et al., 2017), autonomy of the subject on having control of their own personal data (McDermott, 2017), and (in)visibility: the right of the subject to choose if and to know how personal data might be collected and used (Taylor, 2017), are stated as necessary for supporting someone's privacy. The EU General Data Protection Regulation makes also steps towards this direction, by explicitly incorporating transparency and consent in its regulatory claims.

Despite the regulatory efforts, the act of a user opting in, given a very long document of terms and conditions, where how personal data might be used is outlined in a short and general manner does not signify transparency, or actual consent (Strandburg, 2014). Especially regarding personal data for microtargeting, the information that should be presented to the subject in order to give their consent should clarify exactly what information is going to be collected, how, by whom and for what purpose. This is a prerequisite for the subjects' expectations about the collected data to coincide with the actual data usage (Barocas and Nissenbaum, 2014). At the same time, the individuals should be emancipated, by both getting to know through access to the history of their personal data used by services (Kennedy and Moss, 2015), and realizing how datafication has pragmatically altered the contemporary social structure.

Important for the ethical evaluation of microtargeting, as well as for data privacy, is also who acquired the related data, not only how. For us being able to gain access to the aforementioned dataset poses a dilemma: Should public data, for which users have provided their consent to be used and further processed, become openly available, or should they remain only under the control of the initial gatherer? The question is relevant more than ever to the present discussion, given the contemporary Facebook data scandal (Facebook, 2018a, 2018b), as well as the platform's decision to significantly limit the data available through its application programming interface (API). On the one hand, making data broadly open might result to an uncontrolled data mining phenomenon (Pasquale, 2015), with private data becoming a part of the public sphere. On the other hand, the possession of these public data only by the original gatherer might result in the problem of a knowledge monopoly, making the data holder much more powerful in economic and political terms than other social actors.

The specific case study would have a different form, if the data were collected under the new API rules of the platform. Important public data for microtargeting, as user likes, cannot be downloaded in an automated way. If public online data are accessible only to the extent platforms decide, and political actors can target users exclusively through the targeting services provided, then the political system itself becomes contingent to technological companies. Electing microtargeting as a political campaigning strategy thus presupposes the constant compliance of political actors with the existing political and legal conditions (Kruschinski and Haller, 2017), as well as with the market structures and the dominant online platform decisions.

Another issue regarding microtargeting is related to the perceived voter model. Given that the majority of users in social networks are relatively inactive, the danger exists that politicians will concentrate on the analysis of data provided by the more active users, even if that sample is not representative of the population (Barberá and Rivero, 2015). The less data one can gather about a person, the more inexact can their attitude-prediction be. Thus, a campaign might be developed based on falsely assessed voters' attitudes. If political campaigns are highly or exclusively data-driven, it leads to the perceived voter phenomenon (Hersh, 2015): All campaigning decisions are based to an algorithmically calculated electorate and thus, any forecasts are dependent on the nature of the collected data. Given that social media data always possess a

certain rate of bias (Ruths and Pfeffer, 2014), it is possible that political actors might perform a campaigning on a 'constructed' reality and not on an actual one. Of course, gathering of even more data is not a solution. If someone observes campaigning in the US, they might question the independency of the electorate: US parties' campaigns aim for the mobilization or de-mobilization of specific social groups, demographic layers and geographic populations in order to strategically achieve their goals (Hersh, 2015; Kreiss, 2016; Persily, 2017). Furthermore huge public databases contain extensive data about the majority of the electorate and their voting history. The discussion about microtargeting and data privacy is already under way in Europe and the newly emerged issues should be assessed.

This study demonstrates through its 'proof of concept' certain possibilities and dangers of microtargeting, in order to initiate an important debate for the political system. To expand this discussion, further qualitative and quantitative research is needed, in order to uncover: (1) How political communication on social media influences the formation of political attitudes in terms of polarization, political mobilization and opinion formation? (2) What is the effect of political campaigning services offered by social media and other internet platforms? (3) At which level current privacy policies protect individuals and what else could be done? The answers to the aforementioned questions, if given, can redefine how the political discourse should be performed in the digital age.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: the German Research Foundation (DFG) and the Technical University of Munich within the funding programme Open Access Publishing.

ORCID iD

Orestis Papakyriakopoulos http://orcid.org/0000-0003-4680-0022

Notes

1. See e.g. Christl (2016) and Thiele (2017).
2. The pages of CDU and CSU were classified together under the term *Union*.
3. https://www.facebook.com/help/112146705538576 (accessed 21 March 2018).
4. Appendix 1 contains the full description of the topics created, as well as their important keywords.
5. The topics contain keywords as e.g. Vielleicht, aber, glaube, nachdenken.
6. E.g. CDU's app 'connect17'.

References

Agan T (2007) Silent marketing: Micro-targeting. Available at: http://gaia.adage.com/images/random/microtarget031207.pdf (accessed 27 March 2018).

Barberá P and Rivero G (2015) Understanding the political representativeness of twitter users. *Social Science Computer Review* 33(6): 712–729.

Barberá P and Zeitzoff T (2017) The new public address system: Why do world leaders adopt social media? *International Studies Quarterly* 62(1): 121–130.

Barbu O (2014) Advertising, microtargeting and social media. *Procedia – Social and Behavioral Sciences* 163: 44–49.

Barocas S, Bradley E, Honavar V, et al. (2017) Big data, data science, and civil rights. *ArXiv*. Available at: https://arxiv.org/abs/1706.03102 (accessed 5 November 2018).

Barocas S and Nissenbaum H (2014) Big data's end run around anonymity and consent. *Privacy, Big Data, and the Public Good: Frameworks for Engagement* 1: 44–75.

Bennett CJ (2016) Voter databases, micro-targeting, and data protection law: Can political parties campaign in Europe as they do in North America? *International Data Privacy Law* 6(4): 261–275.

Blei DM, Ng AY and Jordan MI (2002) Latent Dirichlet allocation. In: Dietterich TG, Becker S and Ghahramani Z (eds) *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press, pp. 601–608.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Boehm F (2015) A comparison between us and EU data protection legislation for law enforcement purposes. Available at: www.europarl.europa.eu/RegData/etudes/2015/536459/IPOL_ STU%282015%29536459_EN.pdf (accessed 28 March 2018).

Bond R and Messing S (2015) Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review* 109(1): 62–78.

Broy D (2017) Germany: Starting implementation of the GDPR-brief overview of the government bill for a new Federal Data Protection Act. *European Data Protection Law Review* 3: 93.

Capara GV, Barbaranelli C and Zimbardo PG (1999) Personality profiles and political parties. *Political Psychology* 20(1): 175–197.

Castoriadis C (1997) *The Imaginary Institution of Society*. Cambridge, MA: MIT Press.

Christl W (2016) Big data im wahlkampf: An ihren daten sollt ihr sie erkennen. Available at: http://www.faz.net/aktuell/feuilleton/medien/big-data-im-wahlkampf-ist-microtargeting-entscheidend-14582735.html (accessed 28 March 2018).

Cialdini RB (2007) *Influence: The Psychology of Persuasion*. New York, NY: Harper Collins.

Darling WM (2011) A theoretical and practical implementation tutorial on topic modeling and Gibbs sampling. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, Portland, Oregon, 19 – 24 June 2011. pp.642–647. Madison, WI: Omnipress Inc.

Däubler W, Klebe T, Wedde P, et al. (2016) *Bundesdatenschutzgesetz*. Frankfurt: Bund-Verlag.

Deveaud R, SanJuan E and Bellot P (2014) Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17(1): 61–84.

Domke D, Shah DV and Wackman DB (1998) Media priming effects: Accessibility, association, and activation. *International Journal of Public Opinion Research* 10(1): 51–74.

Dorschel J (2015) *Praxishandbuch Big Data: Wirtschaft–Recht–Technik*. Wiesbaden: Springer-Verlag.

Downs A (1957) An economic theory of political action in a democracy. *Journal of Political Economy* 65(2): 135–150.

Edsall TB (2012) Let the nanotargeting begin. Available at: https://campaignstops.blogs.nytimes.com/2012/04/15/let-the-nanotargeting-begin/ (accessed 28 March 2018).

Ellul J (1966) *Propaganda*. New York, NY: Knopf.

EU-Directive (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L119: 1–88.

Facebook (2018a) Hard questions: Update on Cambridge Analytica. Available at: https://newsroom.fb.com/news/2018/03/hard-questions-cambridge-analytica/ (accessed 20 September 2018).

Facebook (2018b) Suspending Cambridge analytica and SCL group from Facebook. Available at: hhttps://newsroom.fb.com/news/2018/03/suspending-cambridge-analytica/ (accessed 20 September 2018).

Franz MM and Ridout TN (2010) Political advertising and persuasion in the 2004 and 2008 presidential elections. *American Politics Research* 38(2): 303–329.

Geman S and Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 6. IEEE, pp.721–741.

Gilks W, Richardson S and Spiegelhalter D (1995) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Griffiths T (2002) Gibbs sampling in the generative model of latent Dirichlet allocation. Available at: https://people.cs.umass.edu/~wallach/courses/s11/cmpsci791s s/readings/griffiths02gibbs.pdf (accessed 28 March 2018).

Grimmer J and Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.

Hegelich S (2017) R for social media analysis. In: Luke Sloan AQH (ed.) *The SAGE Handbook of Social Media Research Methods*. London: SAGE Publications, Chapter 28.

Hegelich S and Shahrezaye M (2015) The communication behavior of German MPS on twitter: Preaching to the converted and attacking opponents. *European Policy Analysis* 1(2): 155–174.

Hersh ED (2015) *Hacking the Electorate: How Campaigns Perceive Voters*. New York, NY: Cambridge University Press.

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24(6): 417.

Jungherr A (2017) Einsatz Digitaler Technologie im Wahlkampf. *Schriftreihe Medienkompetenz* 10111: 92–101.

Karpf D (2016) The partisan technology gap. In: Gordon E and Mihailidis P (eds) *Civic Media: Technology, Design, Practice*. Cambridge, MA: MIT Press, pp. 199–216.

Kennedy H and Moss G (2015) Known or knowing publics? Social media data mining and the question of public agency. *Big Data & Society* 2(2): 2053951715611145.

Kosinski M, Stillwell D and Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110(15): 5802–5805.

Kreiss D (2016) *Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy*. New York, NY: Oxford University Press.

Kruschinski S and Haller A (2017) Restrictions on data-driven political micro-targeting in Germany. *Internet Policy Review* 6(4): 1–23.

McDermott Y (2017) Conceptualising the right to data protection in an era of big data. *Big Data & Society* 4(1): 2053951716686994.

Mayer-Schönberger V and Cukier K (2013) *Big Data – A Revolution that Will Transform how We Live, Work, and Think*. Orlando, FL: Houghton Mifflin Harcourt.

Medina Serrano JC, Hegelich S, Shahrezaye M, et al. (2019) *Social Media Report: The 2017 German Federal Elections*. Munich: TUM University Press.

Nulty P, Theocharis Y, Popa SA, et al. (2016) Social media and political communication in the 2014 elections to the European parliament. *Electoral Studies* 44: 429–444.

Ohm P (2014) Changing the rules: General principles for data use and analysis. *Privacy, Big Data, and the Public Good: Frameworks for Engagement* 1: 96–111.

Panagopoulos C (2015) All about that base. *Party Politics* 22(2): 22–190.

Papakyriakopoulos O, Shahrezaye M, Thieltges A, et al. (2017) Social media und microtargeting in Deutschland. *Informatik-Spektrum* 40(4): 327–335.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.

Persily N (2017) Can democracy survive the internet? *Journal of Democracy* 28(2): 63–76.
Piaget J (1947) *The Psychology of Intelligence*. London, New York: Routledge.
Roberts G and Smith A (1994) Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications* 49(2): 207–216.
Ruths D and Pfeffer J (2014) Social media for large studies of behavior. *Science* 346(6213): 1063–1064.
Schmitt-Beck R (2003) Mass communication, personal communication and vote choice: The filter hypothesis of media influence in comparative perspective. *British Journal of Political Science* 33(2): 233–259.
Sievert C and Shirley KE (2014) Ldavis: A method for visualizing and interpreting topics. In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, Baltimore, Maryland, 27 June 2014, pp. 63–70.
Simon HA (1996) *The Sciences of the Artificial*. Cambridge, MA: MIT Press.
Sotto LJ and Simpson AP (2015) Data protection & privacy: United States. Available at: https://www.huntonprivacyblog.com/wp-content/uploads/sites/18/2011/04/DDP2015_United_States.pdf (accessed 28 March 2018).
Stier S, Posch L, Bleier A, et al. (2017) When populists become popular: Comparing Facebook use by the right-wing movement Pegida and German political parties. *Information, Communication & Society* 20(9): 1–24.
Strandburg KJ (2014) Monitoring, datafication and consent: Legal approaches to privacy in the big data context. In: Lane J, Stodden V, Bender S, et al. (eds) *Privacy, Big Data and the Public Good*. New York, NY: Cambridge University Press, pp. 5–43.
Strohmaier M and Wagner C (2014) Computational social science for the world wide web. *IEEE Intelligent Systems* 29(5): 84–88.
Taylor L (2017) What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society* 4(2): 2053951717736335.
Thiele M (2017) Die wahlschlacht der datenbanken. Available at: http://www.tagesspiegel.de/themen/freie-universitaet-berlin/wahlkampf-mit-big-data-die-wahlschlacht-%der-datenb anken/19938576.html (accessed 28 March 2018).
Thieltges A, Schmidt F and Hegelich S (2016) The devils triangle: Ethical considerations on developing bot detection methods. In: *Proceedings of the 2016 AAAI spring symposium, Stanford University*, Palo Alto, California, 21 – 23 March 2016, Vol. 2123, pp.253–257.
Weedon J, Nuland W and Stamos A (2017) Information operations and Facebook. Technical report, Facebook. Available at: https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf (accessed 28 March 2018).
Woo HY (2015) *Strategic communication with verifiable messages*. PhD Thesis, University of California, USA.

Appendix 1

Table 3. Topics overview for category 1: general political issues.
Topic content keywords.

1	State / Citizens / Social involvement	Bürgerinnen	Gestalten	Zusammenhalten	Engagement	Landkreis
2	Education / Thoughts	Gymnasium	Lösung	Lernen	Klasse	Anforderung
3	Law	Bundesverfassungsgericht	Verfassung	Urteil	Grundgesetz	Bundesrepublik
4	Economy	Euro	SteuergerId	Milliard	Zuschuss	kosten
5	Transportation policy	Flughafen	Nahverkehr	Bahn	Mitarbeiter	Verkehrspolitik
6	Democracy / People / Germany	Demokratie	Volk	Elite	Freiheit	Bürger
7	Against left-wing radicalism	Linksextremisten	Antifa	Gewalttat	Straftat	Polizei
8	Religion / Islam / Christianism	Islam	Muslim	Christlich	Religion	Kirche
9	Thoughts	Vielleicht	Aber	Glaube	Eigentlich	Ich
10	Refugee policy / for	Unterkunft	Fluchtling	Asybewerber	Aufnahme	Geflüchtet
11	Energy policy	Energie	Umwelt	Klimaschutz	Landwirtschaft	Energiepolitik
12	Refugee policy / Against	Fluchtling	Asyl	Abschiebung	illegal	Asylverfahren
14	Austerity / Unemployment / Thoughts	Jobcenter	eigentlich	soll	Sparen	Rettung
15	Education	Schule	Kinder	Eltern	Bildung	Lehre
16	Foreign policy	EU	Russland	Ukrain	USA	Turkei
18	Social policy / Hartz 4 / Poverty	Hartz IV	Armut	Sozial	Gerecht	Rente
19	Greek crisis	Griechenland	Bank	Finanz	Steuerzahl	Schuld
20	Housing policy	Wohnung	Wohnraum	Miete	Verwaltung	Wohnungsbau

(continued)

Table 3. Continued

21	AfD (political discussion)	AfD	Partei	rechtspopulist	Position	Altparteien
23	Against Pegida	Demonstration	Pegida	Nazis	Rassismus	gegen
24	Against right-wing radicals, racism	Diskriminierung	Homophobie	Rechtsextremismus	Freiheit	Rassisten
25	Income / Workers unions	Mindestlohn	Arbeitsgeber	Arbeitsnehmer	Gewerkschaft	Arbeitsbedingung
26	German left-wing history	DDR	Rosa Luxemburg	NATO	Geschichte	Revolution
27	Thoughts	Nachdenken	Denkst	Wahrheit	Du	Einfach
28	Family	Frau	Mann	Mutter	Familie	Kinder
32	Homeland security	Innenminister	Polizei	Ermittelt	Justitz	Kriminalität
37	Against TTIP/CETA	TTIP	CETA	Stopp	unterschreiben	Aktion

Table 4. Topics overview for category 2: political actors' activity.

Topic content keywords						
13	Political events	Eingeladen	Veranstaltung	Lädt	Vortrag	Diskussion
22	Greetings	Gruss	Liebe	Freunde	Melden	Spenden
29	Die Grünen	Grünen	Bündnis	Landtag	Grün-linke	Sachsen
30	After political events	Danke	Besuch	toll	Fotos	Impression
33	Congratulations	Glückwunsch	Herzlich	Gratulieren	Wahlgang	Wiedergewählt
34	Schwesig (Politician)	Schwesig	Manuela	Andrea	Nahles	Frau
35	Political Coalitions	rot	grün	Schwarz	Gelb	Koalition
38	Merkel	Merkel	Angela	Kanzlerin	CDU	Union
39	Petry	Petry	Lucke	Alternative	Deutschland	AfD
40	Election campaign	Daum	Druck	Wahlkampf	Stimmen	Sonntag
41	Candidates	Wahlkreis	Kandidate	Landesliste	Nominiert	Listenplatz
42	Wagenknecht	Mannheim	Wagenknecht	sahra	Linksjugend	Freiburg
45	Twitter	Twitter	Schaut	Teilen	Mitmachen	Abstimmen
46	Various politicians	Gabriel	Schulz	Gauck	Bundespresident	Steinmeier
48	Debates/ TV	live	Aktuell	TV	gleich	Fernsehen
49	FDP/ Rheinland	Pfalz	Rheinland	Liberal	FDP	Liberte
50	Greetings/ Thank you	Wünsche	Spass	Gut	frohe	Feiertag
52	Lindner/ FDP	Lindner	Christian	NRW	Bundesvorsitzender	Kubicki
53	Die LINKE	die Linke	linksfraktion	Riexinger	Kipping	themen
55	German news media	Focus	Welt	Spiegel	Interview	Zeitung

Table 5. Topics overview for category 3: regional topics.

Topic content keywords						
117	NRW politicians	Münster	Bochum	Bezirksvertreter	Essen	Ruhr
31	Hamburg	Altona	Hamburg	Landesparteitag	Bezirkversammlung	Bürgerschaft
36	Leipzig AfD	Leipzig	Kreisvorsitzende	Kreisverband	Vorstand	Mitglied
43	Bayern	Bayern	München	Freistaat	Wahlprogramm	CSU
44	Baden-Württemberg	Baden	Württemberg	Ministerpresident	Stuttgart	bw
47	Bielefeld/ Koblenz	Bielefeld	Koblenz	Mainz	Rülke	Theurer
51	Hamburg/ Schleswig-Holstein	Schleswig	Holstein	Kiel	Rostock	Schwerin
54	Berlin	Berlin	Tempelhof	Lichtenberg	Schöneberg	Bezirk

4 Political communication and recommender systems

4.1 Distorting political communication: The effect of hyperactive users in online social networks

Authors

Orestis Papakyriakopoulos, Morteza Shahrezaye, Juan Carlos Medina Serrano, Simon Hegelich

In

IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 157-164. IEEE, 2019. DOI: 10.1109/INFCOMW.2019.8845094.

Abstract

Online social networks (OSNs) are used increasingly for political purposes. Among others, politicians externalize their views on issues, and users respond to them, initiating political discussions. Part of the discussions are shaped by hyperactive users. These are users that are over-proportionally active in relation to the mean. In this paper, we define the hyperactive user on the social media platform Facebook, both theoretically and mathematically. We apply a geometric topic modelling algorithm (GTM) on German political parties' posts and user comments to identify the topics discussed. We prove that hyperactive users have a significant role in the political discourse: They become opinion leaders, as well as set the content of discussions, thus creating an alternate picture of the public opinion. Given that, we discuss the dangers of replicating the specific bias by statistical and deep learning algorithms, which are used widely for recommendation systems and the profiling of OSN users.

Contribution of thesis author

Theoretical design, model design and analysis, manuscript writing, revision and editing

This is the accepted paper version. Full version: Papakyriakopoulos, Orestis, Morteza Shahrezaye, Juan Carlos Medina Serrano, and Simon Hegelich. "Distorting Political Communication: The Effect Of Hyperactive Users In Online Social Networks." In IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 157-164. IEEE, 2019. DOI: 10.1109/INFCOMW.2019.8845094 https://ieeexplore.ieee.org/document/8845094

Distorting Political Communication: The Effect Of Hyperactive Users In Online Social Networks

Orestis Papakyriakopoulos, Morteza Shahrezaye, Juan Carlos Medina Serrano, Simon Hegelich

Bavarian School of Public Policy
Technical University of Munich
D-80333 Munich, Germany
{orestis.papakyriakopoulos,morteza.shahrezaye, juan.medina}@tum.de, simon.hegelich@hfp.tum.de

***Abstract*— Online Social Networks (OSNs) are used increasingly for political purposes. Among others, politicians externalize their views on issues, and users respond to them, initiating political discussions. Part of the discussions are shaped by hyperactive users. These are users that are over-proportionally active in relation to the mean. In this paper, we define the hyperactive user on the social media platform Facebook, both theoretically and mathematically. We apply a geometric topic modelling algorithm (GTM) on German political parties' posts and user comments to identify the topics discussed. We prove that hyperactive users have a significant role in the political discourse: They become opinion leaders, as well as set the content of discussions, thus creating an alternate picture of the public opinion. Given that, we discuss the dangers of replicating the specific bias by statistical and deep learning algorithms, which are used widely for recommendation systems and the profiling of OSN users.**

I. INTRODUCTION

Today, internet prevails as a prominent communication and information medium for citizens. Instead of watching TV or reading newspapers, increasing numbers of people get politically informed through online websites, blogs, and social media services. The latest statistics demonstrate that internet as a news source has become as important as television, with its share increasing year by year [1]. Given this shift in the means of news broadcasting, politicians have altered their tactics of communication to the society. OSNs, such as Twitter, Facebook and Instagram, have become a cornerstone of their public profiles as they use OSNs to transmit their activities and opinions on important social issues [2], [3], [4].

The growth of online communities on social media platforms have created a public amenable to political campaigning. Political parties and actors have adapted to the new digital environment [5], and besides the application of new campaigning methods as microtargeting [6], have created microblogs through which they can inform citizens of their views and activities. In addition, OSNs have enabled to users to respond to or comment on the politicians' messages, giving birth to a new type of political interaction and transforming the very nature of political communication.

On OSNs, the flow of information from politicians to citizens and back follows a different broadcasting model than the classical one [7]. Instead of journalists monitoring the political activity, political actors themselves produce messages and make them publicly available on the platforms. Each platform provides its users with access to the generated content, as well as distributes it to them through recommendation algorithms [8], [9]. The received information is then evaluated both directly or indirectly [10], [11]: The political message is interpreted immediately, or subsequently through further social interactions among citizens on the related topics. On OSNs, not only can users respond to politicians in the traditional manner -i.e. through their political activity in the society-, but also respond to or comment on the politicians' views online. This creates a new type of interactivity, as users, who actively engage in online discussions sharing their views, are able to influence the way the initial political information will be assimilated by passive users as well as directly influencing political actors.

This new form of political interactivity transforms political communication. Given the possibility of users to directly respond to the political content set by political actors, and discuss online about political issues with other citizens, OSNs emerge as a fruitful space for agonistic pluralism. They provide the necessary channels for different interests and opinions to be expressed, heard and counterposed; elements that constitute the very essence of political communication. If the discussions held are legitimized within a democratic framework, they form the basis for reaching a conflictual consensus [12], based on which societal decisions can be made. Hence, political communication on OSNs opens new possibilities for citizens to participate in the political shaping of the society, providing them with additional space to address their interests.

Problem statement

Although the above type of political communication has the potential to improve the function of democracy, OSNs possess a structural property that obstructs the unbiased constructive interaction between political actors and citizens: The activities of users on OSNs follow an extreme value distribution [13], [14], [15], [16]. Practically, this means that users are not equally active when using a specific OSN. Among others, the majority of users remain passive, or participate with a very low frequency; they either simply read

the content or like, comment, tweet, etc. very rarely. On the contrary, a very small part of the users is hyperactive, as they over-proportionally interact with the platform they use. Thus, in political communication on OSNs, hyperactive users are citizens who over-proportionally externalize their political attitudes compared to the mean. This could be done by liking, commenting, tweeting or using any other interaction possibility provided by a platform to declare an attitude to a political issue.

The given activity asymmetry becomes a major issue, considering that a considerable part of the society is politically informed via OSNs. As hyperactive users externalize their political attitudes more than the others, they have the potential to distort political communication; political issues that are important to them become overrepresented on OSNs, while the views of normally active users become less visible. Hence, hyperactive users may influence the political discussions towards their ends, creating a deformed picture of the actual public opinion on OSNs. This fact violates the assumption of an equitable public political discourse as part of political communication [17], because the interests and views of normally active users appear as less important.

The above distortion of political communication is intensified by the business models of the OSN platforms. OSNs were not created to be arenas of political exchange. Their aim is to maximize the number of platform users, by keeping them satisfied [18], and to transform this social engagement to profits, i.a. through advertisement. Hence, on OSNs, users are both consumers and citizens [19]. In order to maximize their profits, OSN platforms adjust their recommendation algorithms to the content popularity, with a view to promoting information that most users will like. As hyperactive users influence asymmetrically the popularity of political content, these algorithms might replicate this asymmetry. Thus, a platform might recommend content, which is actually consistent with the political interests of hyperactive users. This phenomenon per se denotes a form of algorithmic manipulation of the political communication: The platform unintentionally magnifies hyperactive users' interests, thus posing the risk of political invisibility for the ones of normal users [20].

Last but not least, the aforementioned misrepresentation of public opinion has a direct impact on political campaigning. Contemporary political actors develop their influence strategies based on the perceived voter model [21], which presupposes the gathering of demographic and political data for the development of statistical models about the electorate's attitudes. As major part of these data is derived from social media, models that fail to take the effect of hyperactive users into account would face an important bias.

Considering the above, we want to answer following questions regarding the activity of hyperactive users:

RQ1: **How can we define hyperactive users mathematically?**

RQ2: **How can we compare and evaluate the political attitudes of hyperactive users in relation to the mean?**

Original Contribution

We mathematically define hyperactive users on OSN Facebook, and identify them on the public pages of the major German political parties. By applying a state-of-the-art topic modelling algorithm, we investigate whether they spread or like different messages on political issues other than normal users and politicians do. We prove that hyperactive users not only are responsible for a major part of online political discussions, but they also externalize different attitudes than the average user, changing the discourse taking place. We quantify their effect on content formation by measuring their popularity and showing that they adopt an opinion leader status. Finally, given the potential influence of hyperactive users on recommendation algorithms, we initiate an important discussion on OSNs as spaces of political communication.

II. Data & Method

A. Data Description

To investigate the effect of hyperactive users, we chose to analyse the public Facebook pages of the main German political parties. Our sample included CDU, CSU, SPD, FDP, Bündnis 90/Die Grünen, Die Linke, and AfD. CDU is the main conservative party of Germany, while CSU is the conservative party active in Bavaria. SPD represents the main German social-democratic party, and Die Linke the radical left. AfD has a nationalist, anti-immigrant, and neo-liberal agenda, while FDP is a conservative, neo-liberal party. Finally, Bündnis 90/Die Grünen is the German green party. We focused on Facebook, because the platform's api restrictions and its monitoring system largely prevent automated activities, as e.g. performed by social bots on other platforms [22], [23]. Therefore we could evaluate the natural behaviour of hyperactive users and not an algorithmically generated one.

We took into consideration all party posts and their reactions in the year 2016. This choice was made, because we wanted to investigate a full year of user activities. We preferred 2016 over 2017, because 2017 was an election year, with most content produced by the parties being campaign related. By contrast, 2016 was marked by the Refugee Crisis in Europe, and we were interested in evaluating the discussions on the topic. In total, by accessing the Facebook Graph API, we collected 3,261 Posts, 3,084,464 likes and 382,768 comments, made by 1,435,826 users. The sample included all posts and comments on the posts generated for the period under investigation.

B. Defining hyperactive users

We consider hyperactive users as people, whose behaviour deviates from the standard on an OSN platform. To obtain an understanding of the overall behaviour of the users, we fitted discrete power-law and extreme value distributions to mathematically describe the users' like and comment activities. Additionally, we ran bootstrapped and comparative goodness-of-fit tests based on the Kolmogorov-Smirnov [24] and the Vuong [25] statistic to evaluate the potential fits, as proposed by Clauset et al. [26]. The KS test examines

the null hypothesis that the empirical sample is drawn from the reference distribution, while the Vuong test measures the log-likelihood ratio of two distributions and, based on it, investigates whether both empirical distributions are equally far from a third unidentified theoretical one.

In order to mathematically describe the activities of hyperactive users, we selected to treat them as outliers of the standard OSN population. We adopt the definitions made by Barnett and Lewis [27], Johnson and Wichern [28] and Ben-Gal [29], and see outliers not as errors, or coming from a different generative process, but as data containing important information, which is inconsistent with and deviating from the remainder of the data-set. Therefore, given the extreme skewed distribution of the activities, we followed the method proposed by Hubert and Vam der Veeken [30] and Hubert and Vandervieren [31] for outlier detection. We calculated the quartiles of our data Q_1 and Q_3, the interquartile range $IQR = Q_3 - Q_1$ and the whiskers w_1 and w_2, which extend from the Q_1 and Q_3 respectively to the following limits:

$$[Q_1 - 1.5e^{-4MC} IQR, Q_3 + 1.5e^{3MC} IQR] \qquad (1)$$

where MC is the medcouple [32], a robust statistic of the distribution skewness. Data beyond the whiskers were marked as outliers.

C. Topic Modeling

After evaluating the likes and comments distributions, as well as identifying the existing hyperactive users, we prepared our data for the application of a topic modelling algorithm. As it has been shown that a noun-only topic modelling approach yields more coherent topic-bags [33], we cleaned our posts and comments from the remaining part-of-speech types. To do so, we applied the spaCy pretrained convolutional neural network (CNN) classifier [34] based on the Tiger [35] and WikiNER [36] corpuses, classified each word in our document collection, and kept only the nouns.

We wanted to investigate the various topics that users and parties discussed about but did not want to differentiate on the way they talked about them. Parties usually use a more formal language when posting on a topic than users. Therefore, there was the risk that the topic modelling algorithm would create different topics on the same issue, one for the parties and one for the users. To avoid this, we fitted our model on the user comments, and then classified the parties' posts through the trained model.

For our analysis, we applied a non-parametric Conic Scan-and-Cover (CoSAC) algorithm for geometric topic modeling [37]. Our decision was based on the fact that most topic modelling algorithms (e.g. LDA [38], NMF [39]) need a priori as input the number of topics to split the corpus. CoSAC has the advantage of electing itself the number of topics to find the most efficient topic estimates.

The algorithm presupposes that the optimal number of topics K are embedded in a V-1 dimensional probability simplex Δ^{V-1}, where V the number of words in the corpus. Each topic β_K corresponds to a set of probabilities in the word simplex. The totality of topics build hence a convex polytope $B = conv(\beta_1, ..., \beta_K)$. Each document corresponds to a point $p_m = (p_{m1}, ..., p_{mV})$ inside Polytope B, with $p_m = \sum_k \beta_k \theta_{mk}$. θ_{mk} denotes the proportion that topic k covers in document m. Finally each document is drawn from a multinomial distribution of words: $w_m \sim Multinomial(p_m, N_m)$, where N_m the number of words in document m.

The CoSAC algorithm iteratively scans the polytope B and finds the furthest point from its center Cp. It then constructs a conical region with angle ω, starting from Cp and embedding the specific point (Figure 1). All points within the cone are considered to belong in the same topic and are removed from the polytope. The procedure is iterated K-1 times, until almost no points remain in the polytope. A cone is considered sufficient if it covers at least a proportion of documents λ. After fitting the cones, CoSAC places a sphere with radius R to the polytope, to cover the remaining points. The K geometric objects and their respective points correspond to the K topics created by the algorithm. In our model, the hyperparameters were set to $\omega = 0.6, \lambda = 0.001$ and $R = 0.05$, as proposed by the authors.

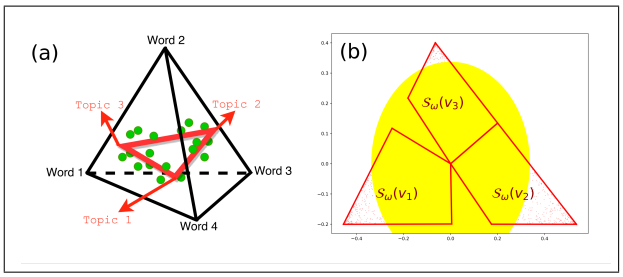

Fig. 1. (a) The topic polytope embedded in the word simplex. (b) Cones and sphere coverage of the polytope.

D. Comparison of activities

Given our topics, we wanted to evaluate the differences in the activity of normal and hyperactive users. Therefore, we calculated the empirical distributions $f(comment|topic)$ over all topics for the comments of normal and hyperactive users respectively. We pairwise compared the distributions for each topic, by applying a 2-Sample Anderson-Darling Test [40]. The test calculates the probability that the populations from which two groups of data were drawn are identical.

Besides testing the empirical comment-topic distributions, we assigned to each comment the topic with the highest probability and compared the most commented topics for normal and hyperactive users. Similarly, we assigned the classified party posts to their most probable topic and aggregated the likes of normal and hyperactive users. In this way, we were in the position to locate the concrete political interests of users.

III. Results

The results are split into three parts. First, we present our findings on the general user distribution on the investigated

TABLE I

VUONG TEST RESULTS

Log-normal vs	Likes LL-ratio (p-value)	Comments LL-ratio (p-value)
Power-law	15.1 (<0.01)	34.9 (<0.01)
Poisson	34.9 (<0.01)	12.7 (<0.01)
Exponential	12.7 (<0.01)	26.6 (<0.01)

pages. Based on that, we analyze the number and distribution of hyperactive users among the different pages. Then, we compare the behaviour between hyperactive and normal users by taking into consideration the topic modelling results and further statistical tests. Given that, we evaluate the importance and role of hyperactive users in the political discourse on OSNs.

A. Describing user activity

As a first result, we identified the log-normal distribution as the the best measure for describing the user activities (Figure 2). The bootstrapped KS-Tests (100 samples, two tailed) for both comments and likes failed to reject the null that our data come from a log-normal distribution (gof $<$ 0.01, p $>$ 0.05 and gof $<$ 0.01, p $>$ 0.2 respectively), while the comparative Vuong tests showed a better fit of the log-normal in comparison to the power-law, poisson and exponential distributions (Table I). Our results comply with the existing literature, which states that usually complex social network properties are log-normally distributed [15], [41], [42]. Figure 2 shows the empirical frequencies of user activities and their respective log-normal fits.

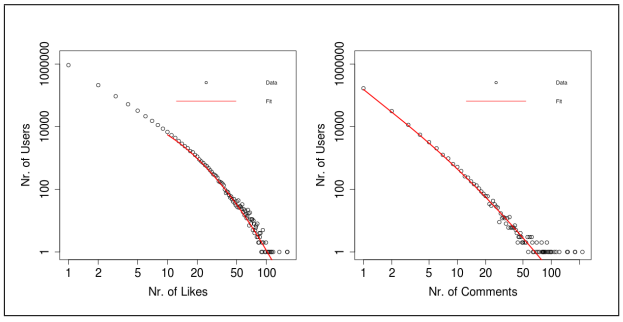

Fig. 2. (a) The topic polytope embedded in the word simplex. (b) Cones and sphere coverage of the polytope.

B. Detecting hyperactive users

Through our outlier detection methodology, we detected 12,295 hyperactive users on the comment section of pages, who correspond to 5.3% of the total users commenting on the pages. Due to the extreme skewness of the comments' distribution, a user was characterized as hyperactive if they made three or more comments. This is justified by the fact that actually 74% of the users under investigation made only one comment. Although hyperactive users represented 5.3% of the total commenting population, they accounted for 25.8% of the total comments generated on the parties' pages. Furthermore, 56% of these hyperactive users commented on two or more party pages, denoting that they generally interacted with users across the political spectrum. By evaluating the popularity of the users' comments, it was found that hyperactive users tend to get more support than the rest. Comments made by hyperactive users on average gained 3.52 likes, while normal users' comments only 3.07, a difference that was statistically significant (Mann-Whitney Test with continuity correction, one tailed: W = 1.4^{10}, p$<$ 0.01). This complies with previous research stating that highly active users tend to have the characteristics of opinion leaders [43].

TABLE II

HYPERACTIVE USERS PER PARTY - COMMENTS

Party	Comments by Hyperactive Users	Ratio
AfD	43,017	0.24
CDU	20,929	0.45
CSU	18,312	0.22
FDP	1,400	0.15
Die Grünen	8,946	0.36
Die Linke	2,343	0.16
SPD	3,926	0.13

Similarly, the evaluation of the pages' likes resulted in the characterization of 61,372 users as hyperactive, or 4.3% of the total users that liked the parties' posts. As before, the methodology labelled users as hyperactive if they made three or more likes, because the majority of the active Facebook population rarely interacted with the related pages. The likes of these hyperactive users accounted for 26.4% of total likes, hence having a major effect on the overall content liked. In addition, 29% of hyperactive users liked posts of more than one party, denoting again that their activities were spread over the entire parties' network. The overview of the hyperactive users' commenting and like activities for each party can be found in tables II and III. We also compared the like and comment distributions, by calculating their gini index. The measure provides a proxy for inequality, with 0 denoting perfect equality and 1 extreme inequality. In our case, we calculated a value of 0.35 and 0.45 for the comment and like distribution respectively. This denotes that like activities are more unequally distributed than the comment activities, i.e. hyperactive users play a bigger role in the formation of likes. In addition, the values denote a degree of inequality between normal and hyperactive users, but not an extreme one. Nevertheless this is misleading, because the measure does not take into consideration the inactive users. Given that information, the gini index would have been much higher in both cases.

C. Evaluating the political attitudes of hyperactive users

Based on the categorization of users as hyperactive or normal, we could then evaluate the results of the topic modelling algorithm. The model clustered the users' comments in 69 main topics. A major part of the topics concerned the refugee crisis of 2016 and the related discussions about Islam. A set of topics aggregated comments on German Chancellor Merkel, on the leaders of other parties, on female and male politicians and the German parties in general. There was one

TABLE III
HYPERACTIVE USERS PER PARTY - LIKES

Party	Likes by Hyperactive Users	Ratio
AfD	555,564	0.35
CDU	16,997	0.2
CSU	139,493	0.2
FDP	20,188	0.16
Die Grünen	28,777	0.19
Die Linke	24,546	0.14
SPD	29,057	0.12

topic summing up comments in English language, as well as a topic containing hyperlinks. Furthermore, the algorithm created policy related topics regarding foreign affairs, as well as the economy and labour market and the state in general. Other topics were related to the German national identity, society and solidarity, and the nature of democracy. Users also discussed about family and gender policy, homeland security, transportation and environmental policy. There were topics that included wishes, fear, ironic and aggressive speech, as well as topics aggregating user thoughts. Finally, a set of topics was about political events and communications and a number of topics included comments against mainstream media and the political system. An overview can be found in table IV. The geometric topic modelling algorithm was able to provide a broad picture of the discussion topics on the parties' pages, revealing numerous insights about the way Facebook users commented on the parties' posts. By splitting the comments into two categories, one for the ones generated by hyperactive users and one for the comments of normal users, and by assigning them to the topics to which they were mostly related, we created a stacked chart illustrating the share of hyperactive users' comments for every topic (Figure 3). It is evident that hyperactive users covered a major part of the comments, and although more active, they commented more or less similarly to the normal users among the various topics. Despite that, the Anderson-Darling tests rejected the null hypothesis that hyperactive and normal users' comments come from the same distribution for 54 out of the 69 topics. Practically, this means that the topic density distributions varied between the comments of normal and hyperactive users. This is caused when the comments contain different words in different proportions. Hence, hyperactive and normal users used different vocabularies when referring to a topic and, consequently, externalized overall different views and sentiment, or focused on different issues in each case.

Besides the fact that hyperactive users had a different behaviour on the posts' comments, our analysis showed that they also had different liking preferences. After classifying each party post to the most relevant topic, we counted the likes of the posts that belong to each topic. We took into consideration only topics that were based on either political vocabulary or politicians, and ignored topics that contained aggressive speech or sentiment, because the related vocabulary was rarely used by the parties. Figure 4 illustrates

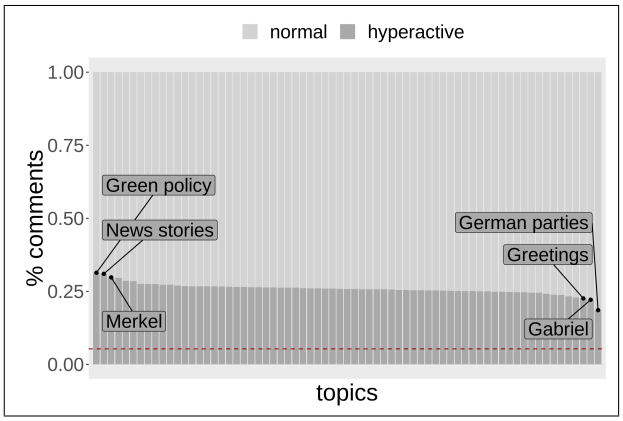

Fig. 3. Proportion of comments generated by normal and hyperactive users. The dotted red line gives the actual proportion of hyperactive users. The plot also illustrates the three most and least interesting topics for hyperactive users.

a stacked chart depicting the share of hyperactive users' likes. In contrast to the comments' chart, it is obvious that hyperactive users liked specific topics with different intensity than normal users. Even though hyperactive users performed on average 26.4% of the likes, they liked much more content related to EU politics and labour policy, while had less interest on the conservative party AfD, citizens' rights and the region of Bavaria. Therefore, it is clear that hyperactive users influence the like distribution of the public to party posts.

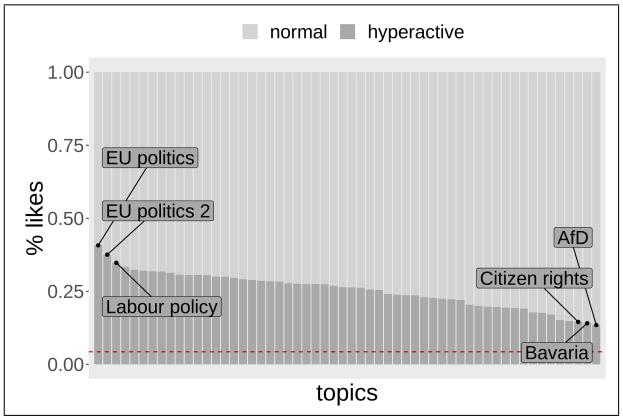

Fig. 4. Proportion of likes generated by normal and hyperactive users. The dotted red line gives the actual proportion of hyperactive users. The plot also illustrates the three most and least interesting topics for hyperactive users.

It must be noted that our analysis gives an overview of the content of posts. It cannot identify sentiment, or specific predispositions of users. For a firm understanding of the issues that were over- or under-represented by hyperactive users an additional extensive analysis is needed, which is beyond the scope of this paper. Our analysis demonstrated that, both on commenting and liking, hyperactive users have a different behaviour than the other users.

Taking the above into consideration, it was possible to show that political communication on Facebook is strongly

constituted by the behaviour of hyperactive users. By describing the user like and comment activities on the platform, we managed to characterize users as hyperactive or normal through outlier detection. We proved that hyperactive users account for a significant part of the total users' activities, they participate in discussions differently from the rest, and they like different content. Moreover, they become opinion leaders, as their comments become more popular than these of the normal users. Taking Facebook as an example, we showed that user activities on OSNs are neither equally nor evenly distributed.

IV. Discussion

Given that activity asymmetries are a feature of online social networks, it is important to evaluate the consequences for science and the society. Although our analysis was concentrated on Facebook, previous research has proven that hyperactive accounts, either human or automated, have the potential to equally influence political communication in other platforms [44], [45]. The specific formation and distribution of political interactions on OSNs rises various questions regarding the role and impact of OSNs on a political level, on an algorithmic level, as well as on the intersection of both.

In the political dimension, the OSN activity asymmetries are transformed into an asymmetry of disseminated political content, as the attitudes and interests of hyperactive users appear over-proportionally in the discussions taking place. Until now, research [46], [47] has stated that OSNs suffer from a population bias: The people using OSNs are not representative of the actual society. On top of that, a content bias is now added: The content disseminated on OSNs is not even representative of the mean users' attitudes on the platform. This poses a scientific problem, as it might lead to false research results. Equally importantly, it poses a political problem, because political discussions and opinion exchange is distorted by the effect of hyperactive accounts. This happens not because the diffused information in the network is transformed or changed, but because hyperactive users strongly contribute to the type of information diffused. Their attitudes fill the communication space, leading to a bias on the political feedback to politicians, and to a shift on the issues that shape the political agenda. Although OSNs provide a more open environment to express opinions than traditional media, it ends up being partly a gathering of political echoes [48] that struggle to be imposed on each other.

In the algorithmic dimension, the extreme skewness of the activity distributions raises specific issues regarding the recommendation algorithms used by OSN platforms. The first problem is related to algorithmic accuracy: skewed data are, imbalanced data, and their raw use, either as input features or as output labels, can yield weak classification results. The imbalanced learning problem applies to both standard statistical algorithms, collaborative filtering and neural networks [49], [50], [51], with algorithms overestimating the importance of outliers and under-estimating the importance of the rest. This also happens in the case of a poor selection of a cost function [52]. Furthermore, it is proven that statistical models as Markov-chains might fail to capture the signal immanent in highly skewed data, while deep learning methods might face the same issue given power-law distributions of data [53].

The second potential problem is that an algorithm might fail, not in the sense that it might be unable to learn from the data, but rather learn the wrong signal. Hyperactive users can be seen as physical adversaries [54] of the mean user attitudes. Algorithms trained in the full data will include the bias, tracking and predicting behavioural associations that correspond to hyperactive users rather than to the population majority. It is not coincidental that the detection of adversaries in machine learning can be done by sample distribution comparison [55], in the same way as we tracked the different preferences of hyperactive users.

Solutions to the aforementioned issues exist and are usually taken into consideration by data scientists, who develop recommendation algorithms. Nevertheless, in the case of political communication, an algorithmic issue automatically becomes a political one. Recommendation systems come with a social influence bias [56], [57], i.e. have the power to change users' opinion. Hence, OSNs promoting biased political content might result in the algorithmic manipulation of political communication.

In addition, social media platforms are not designed to foster political discourses [58], but rather aim at increasing active users, in order to sell advertisement and attract funding from venture capitalists [59]. Hence, the structure and impact of recommendation algorithms distorts human behaviour [60], having transformative effects that were not foreseen a priori [61].

It is evident from the above, that each algorithm mediates and redefines the importance of political interests [62], raising further questions about the opacity of the recommendation systems [63]. In a political context, it becomes important to know as citizens, how, why and with what impact algorithms change political communication. This presupposes awareness of the data processed and, the mathematical method applied, as well as knowledge of what exactly a machine learning cost function optimizes and to what extent recommendation systems alter human behaviour. Proposals for algorithmic transparency have already been made [64], [65], [66], and wait to be applied in practice.

The above issues need to be extensively analyzed, in order to evaluate and shape the structure of political communication in the digital era. In this paper we laid the foundations for this discussion, by defining, demonstrating and quantifying the effect of hyperactive users on OSNs, through the example of Facebook. We also illustrated and defined the risks of algorithmic manipulation by the OSN recommendation systems. Future research needs to focus on the aforementioned consequences, evaluate the structure of OSNs ethically, politically and normatively as political intermediators, as well as propose and apply solutions to the newly posed problems.

TABLE IV
TOPIC MODELING, AD-TEST RESULTS AND PROPORTION OF HYPERACTIVE USERS

Nr.	Topic	AD-test gof, (p-value)	Comments	Likes
1	Immigration	3.8, (0.0)	0.27	0.30
2	Merkel	104.2, (1.0)	0.28	0.24
3	AfD	15.9, (0.0)	0.25	0.30
4	News stories	17.4, (0.0)	0.31	0.29
5	English	8.8, (0.0)	0.26	-
6	Green policy	15.1, (0.0)	0.31	0.18
7	Islam	4.8, (0.0)	0.26	0.31
8	Integration immigrants	6.7, (0.0)	0.27	0.28
9	Female politicians	9.5, (0.0)	0.26	0.22
10	Deportion immigrants	9.2, (0.0)	0.26	0.20
11	EU politics	2.5, (0.0)	0.26	0.41
12	Economic policy	6.1, (0.0)	0.28	0.31
13	Greetings	17.7, (0.0)	0.23	-
14	Polls	16.3, (0.0)	0.25	0.26
15	Union	71.2, (1.0)	0.29	0.26
16	CSU	69.2, (1.0)	0.24	0.24
17	National identity	11.5, (0.1)	0.26	0.29
18	Human rights	1.5, (0.1)	0.26	0.24
19	Security	2.6, (0.0)	0.27	0.24
20	Democracy	32.3, (0.0)	0.25	0.27
21	Citizen rights	33.9, (0.0)	0.25	0.15
22	Congratulations	26.5, (0.0)	0.24	0.26
23	Gabriel	43.2, (1.0)	0.22	0.23
24	Foreign affairs	5.0, (0.0)	0.26	0.26
25	Homeland security	17.3, (0.0)	0.25	0.25
26	Interviews	23.9, (0.0)	0.25	0.18
27	Turkey affairs	11.0, (0.0)	0.26	0.19
28	Terrorism	7.1, (0.0)	0.26	0.19
29	Fear	1.6, (0.1)	0.26	-
30	Party system	4.3, (0.0)	0.27	0.29
31	The people	3.2, (0.0)	0.27	0.27
32	News media	1.3, (0.1)	0.27	0.31
33	Erdogan	7.1, (0.0)	0.27	0.23
34	German parties	25.4, (0.0)	0.19	0.19
35	Social policy	10.9, (0.0)	0.26	0.27
36	Reflection	14.5, (0.0)	0.26	-
37	TTIP/CETA	15.7, (0.0)	0.25	0.28
38	Syria	2.4, (0.0)	0.25	0.17
39	Labour policy	20.9, (0.0)	0.24	0.30
40	Party policies	0.2, (0.3)	0.26	0.27
41	Media	32.1, (0.0)	0.25	-
42	DDR	12.9, (0.0)	0.26	0.33
43	Male politicians	2.5, (0.0)	0.25	0.28
44	East Germany	5.0, (0.0)	0.26	0.32
45	Speeches	53.6, (1.0)	0.25	-
46	Bavaria	67.1, (1.0)	0.25	0.14
47	State media	21.4, (0.0)	0.25	-
48	Female politicians 2	12.0, (0.0)	0.30	0.20
49	Bundestag	10.4, (0.0)	0.25	0.32
50	Interviews 2	16.9, (0.0)	0.25	0.28
51	Irony	42.4, (1.0)	0.26	-
52	Trump	16.2, (0.0)	0.26	0.22
53	Welfare policy	12.3, (0.0)	0.26	0.32
54	Videos	13.0, (1.0)	0.25	-
55	Government	26.1, (0.0)	0.26	0.31
56	Transportation policy	37.0, (0.0)	0.23	0.15
57	Green policy 2	3.7, (0.0)	0.27	0.20
58	Politicians	12.1, (0.0)	0.23	-
59	Public services	18.4, (0.0)	0.25	0.20
60	Gender Equality	19.7, (0.0)	0.26	0.31
61	Insults	30.5, (0.0)	0.25	-
62	Boarder security	3.4, (0.0)	0.27	0.32
63	Media 2	13.5, (0.0)	0.27	-
64	EU politics 2	2.3, (0.0)	0.25	0.38
65	Merkel 2	39.9, (0.1)	0.30	0.15
66	AfD 2	2.6, (0.0)	0.26	0.13
67	Funny	23.9, (0.0)	0.25	-
68	Germans	0.5, (0.2)	0.27	0.22
69	Labour policy 2	8.5, (0.0)	0.27	0.35

REFERENCES

[1] J. Gottfried and E. Shearer, "Americans' online news use is closing in on tv news use," Sep 2017. [Online]. Available: http://www.pewresearch.org/fact-tank/2017/09/07/americans-online-news-use-vs-tv-news-use/

[2] S. Hegelich and M. Shahrezaye, "The communication behavior of german mps on twitter: preaching to the converted and attacking opponents," *European Policy Analysis*, vol. 1, no. 2, pp. 155–174, 2015.

[3] G. S. Enli and E. Skogerbø, "Personalized campaigns in party-centred politics: Twitter and facebook as arenas for political communication," *Information, Communication & Society*, vol. 16, no. 5, pp. 757–774, 2013.

[4] V. Arnaboldi, A. Passarella, M. Conti, and R. Dunbar, "Structure of ego-alter relationships of politicians in twitter," *Journal of Computer-Mediated Communication*, vol. 22, no. 5, pp. 231–247, 2017. [Online]. Available: http://dx.doi.org/10.1111/jcc4.12193

[5] J. C. M. Serrano, S. Hegelich, M. Shahrezaye, and O. Papakyriakopoulos, *Social Media Report: The 2017 German Federal Elections*, 1st ed. TUM.University Press, 2018.

[6] O. Papakyriakopoulos, S. Hegelich, M. Shahrezaye, and J. C. M. Serrano, "Social media and microtargeting: Political data processing and the consequences for germany," *Big Data & Society*, vol. 5, no. 2, p. 2053951718811844, 2018. [Online]. Available: https://doi.org/10.1177/2053951718811844

[7] M. E. McCombs and D. L. Shaw, "The agenda-setting function of mass media," *Public opinion quarterly*, vol. 36, no. 2, pp. 176–187, 1972.

[8] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[9] Twitter, https://help.twitter.com/en/using-twitter/twitter-trending-faqs, 2018, online; accessed 24 August 2018.

[10] M. Hilbert, J. Vásquez, D. Halpern, S. Valenzuela, and E. Arriagada, "One step, two step, network step? complementary perspectives on communication flows in twittered citizen protests," *Social Science Computer Review*, vol. 35, no. 4, pp. 444–461, 2017.

[11] S. Choi, "The two-step flow of communication in twitter-based public forums," *Social Science Computer Review*, vol. 33, no. 6, pp. 696–711, 2015.

[12] C. Mouffe, *The democratic paradox*. Verso, 2000.

[13] N. Blenn and P. Van Mieghem, "Are human interactivity times lognormal?" *arXiv preprint*, 2016.

[14] P. Van Mieghem, N. Blenn, and C. Doerr, "Lognormal distribution in the digg online social network," *The European Physical Journal B*, vol. 83, no. 2, p. 251, 2011.

[15] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks," in *Proceedings of the fourth International AAAI Conference on Web and Social Media*. AAAI, 2010, pp. 90–97.

[16] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. ACM, 2009, pp. 49–62.

[17] A. Schaap, "Agonism in divided societies," *Philosophy & Social Criticism*, vol. 32, no. 2, pp. 255–277, 2006.

[18] N. Shi, M. K. Lee, C. M. Cheung, and H. Chen, "The continuance of online social networks: how to keep people using facebook?" in *Fourty-third Hawaii international conference on System sciences*. IEEE, 2010, pp. 1–10.

[19] C. R. Sunstein, *# Republic: Divided democracy in the age of social media*. Princeton University Press, 2018.

[20] T. Bucher, "Want to be on the top? algorithmic power and the threat of invisibility on facebook," *new media & society*, vol. 14, no. 7, pp. 1164–1180, 2012.

[21] E. D. Hersh, *Hacking the electorate: How campaigns perceive voters*. Cambridge University Press, 2015.

[22] Facebook, "Using the graph api - documentation." [Online]. Available: https://developers.facebook.com/docs/graph-api/using-graph-api/

[23] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Eleventh international AAAI conference on web and social media*, 2017.

[24] T. B. Arnold and J. W. Emerson, "Nonparametric goodness-of-fit tests for discrete null distributions." *R Journal*, vol. 3, no. 2, 2011.

[25] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica: Journal of the Econometric Society*, pp. 307–333, 1989.

[26] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[27] V. Barnett and T. Lewis, *Outliers in statistical data*. Wiley, 1974.

[28] R. Johnson and D. Wichern, *Applied multivariate Statistical Analysis.* Pretice Hall, 1998.

[29] I. Ben-Gal, "Outlier detection," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 131–146.

[30] M. Hubert and S. Van der Veeken, "Outlier detection for skewed data," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 22, no. 3-4, pp. 235–246, 2008.

[31] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," *Computational statistics & data analysis*, vol. 52, no. 12, pp. 5186–5201, 2008.

[32] G. Brys, M. Hubert, and A. Struyf, "A robust measure of skewness," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 996–1017, 2004.

[33] F. Martin and M. Johnson, "More efficient topic modelling through a noun only approach," in *Proceedings of the Australasian Language Technology Association Workshop 2015*, 2015, pp. 111–115.

[34] M. Honnibal and M. Johnson, "An improved non-monotonic transition system for dependency parsing," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1373–1378.

[35] S. Brants, S. Dipper, P. Eisenberg, S. Hansen-Schirra, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit, "Tiger: Linguistic interpretation of a german corpus," *Research on language and computation*, vol. 2, no. 4, pp. 597–620, 2004.

[36] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from wikipedia," *Artificial Intelligence*, vol. 194, pp. 151–175, 2013.

[37] M. Yurochkin, A. Guha, and X. Nguyen, "Conic scan-and-cover algorithms for nonparametric topic modeling," in *Advances in Neural Information Processing Systems*, 2017, pp. 3878–3887.

[38] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[39] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[40] F. W. Scholz and M. A. Stephens, "K-sample anderson–darling tests," *Journal of the American Statistical Association*, vol. 82, no. 399, pp. 918–924, 1987.

[41] K. Sun, "Explanation of log-normal distributions and power-law distributions in biology and social science," Department of Physics, University of Illinois at Urbana-Champaign, Tech. Rep., 2004.

[42] H. Kuninaka and M. Matsushita, "Statistical properties of complex systems-lognormal and related distributions," in *AIP Conference Proceedings*, vol. 1468. AIP, 2012, pp. 241–251.

[43] B. E. Weeks, A. Ardèvol-Abreu, and H. Gil de Zúñiga, "Online influence? social media use, opinion leadership, and political persuasion," *International Journal of Public Opinion Research*, vol. 29, no. 2, pp. 214–239, 2017.

[44] A. Thieltges, O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich, "Effects of social bots in the iran-debate on twitter," *arXiv preprint*, 2018.

[45] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," *Nature communications*, vol. 9, no. 1, p. 4787, 2018.

[46] T. Correa, A. W. Hinsley, and H. G. De Zuniga, "Who interacts on the web?: The intersection of users personality and social media use," *Computers in Human Behavior*, vol. 26, no. 2, pp. 247–253, 2010.

[47] M. Duggan and J. Brenner, "The demographics of social media users - 2012," Pew Research Center's Internet & American Life Project Washington, DC, Tech. Rep., 2013.

[48] Y.-R. Lin, J. P. Bagrow, and D. Lazer, "More voices than ever? quantifying media bias in networks," in *Proceedings of the fifth International AAAI Conference on Web and Social Media*. AAAI, 2011.

[49] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge & Data Engineering*, vol. 9, pp. 1263–1284, 2008.

[50] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63–77, 2006.

[51] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang, "One-class collaborative filtering," in *IEEE eighth International Conference on Data Mining*. IEEE, 2008, pp. 502–511.

[52] C.-C. Lee, P.-C. Chung, J.-R. Tsai, and C.-I. Chang, "Robust radial basis function neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 6, pp. 674–685, 1999.

[53] H. W. Lin and M. Tegmark, "Criticality in formal languages and statistical physics," *arXiv preprint*, 2016.

[54] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint*, 2016.

[55] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," *arXiv preprint*, 2017.

[56] S. Krishnan, J. Patel, M. J. Franklin, and K. Goldberg, "A methodology for learning, analyzing, and mitigating social influence bias in recommender systems," in *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 2014, pp. 137–144.

[57] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl, "Is seeing believing?: how recommender system interfaces affect users' opinions," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2003, pp. 585–592.

[58] S. Engelmann, J. Grossklags, and O. Papakyriakopoulos, "A democracy called facebook? participation as a privacy strategy on social media," in *Annual Privacy Forum*. Springer, 2018, pp. 91–108.

[59] M. Falch, A. Henten, R. Tadayoni, and I. Windekilde, "Business models in social networking," in *CMI International Conference on Social Networking and Communities*, 2009.

[60] D. Ruths and J. Pfeffer, "Social media for large studies of behavior," *Science*, vol. 346, no. 6213, pp. 1063–1064, 2014.

[61] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, 2016.

[62] H. Nissenbaum, "From preemption to circumvention: if technology regulates, why do we need regulation (and vice versa)," *Berkeley Technology Law Journal*, vol. 26, p. 1367, 2011.

[63] J. Burrell, "How the machine thinks: Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, 2016.

[64] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, "Auditing algorithms: Research methods for detecting discrimination on internet platforms," in *Data and discrimination: converting critical concerns into productive inquiry*, 2014, pp. 1–23.

[65] N. Diakopoulos, "Algorithmic-accountability: the investigation of black boxes," Tow Center for Digital Journalism, Tech. Rep., 2014.

[66] F. Pasquale, *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

4.2 Political communication on social media: A tale of hyperactive users and bias in recommender systems

Authors

Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, Simon Hegelich

In

Online Social Networks and Media Volume 15, January 2020, 100058, Elsevier

Abstract

A segment of the political discussions on Online Social Networks (OSNs) is shaped by hyperactive users. These are users that are over-proportionally active in relation to the mean. By applying a geometric topic modeling algorithm (GTM) on German users' political comments and parties' posts and by analyzing commenting and liking activities, we quantitatively demonstrate that hyperactive users have a significant role in the political discourse: They become opinion leaders, as well as having an agenda-setting effect, thus creating an alternate picture of public opinion. We also show that hyperactive users strongly influence specific types of recommender systems. By training collaborative filtering and deep learning recommendation algorithms on simulated political networks, we illustrate that models provide different suggestions to users, when accounting for or ignoring hyperactive behavior both in the input dataset and in the methodology applied. We attack the trained models with adversarial examples by strategically placing hyperactive users in the network and manipulating the recommender systems' results. Given that recommender systems are used by all major social networks, that they come with a social influence bias, and given that OSNs are not per se designed to foster political discussions, we discuss the implications for the political discourse and the danger of algorithmic manipulation of political communication.

Contribution of thesis author

Theoretical design, model design and analysis, manuscript writing, revision and editing

Online Social Networks and Media 15 (2020) 100058

Political communication on social media: A tale of hyperactive users and bias in recommender systems

Orestis Papakyriakopoulos*, Juan Carlos Medina Serrano, Simon Hegelich

Bavarian School of Public Policy, Technical University of Munich, Germany

Keywords:
Social media
Political communication
Hyperactive users
Recommendation systems
Algorithmic bias
Recommender systems
Topic modeling
Computational social science
Political data science
Agenda setting

A B S T R A C T

A segment of the political discussions on Online Social Networks (OSNs) is shaped by hyperactive users. These are users that are over-proportionally active in relation to the mean. By applying a geometric topic modeling algorithm (GTM) on German users' political comments and parties' posts and by analyzing commenting and liking activities, we quantitatively demonstrate that hyperactive users have a significant role in the political discourse: They become opinion leaders, as well as having an agenda-setting effect, thus creating an alternate picture of public opinion. We also show that hyperactive users strongly influence specific types of recommender systems. By training collaborative filtering and deep learning recommendation algorithms on simulated political networks, we illustrate that models provide different suggestions to users, when accounting for or ignoring hyperactive behavior both in the input dataset and in the methodology applied. We attack the trained models with adversarial examples by strategically placing hyperactive users in the network and manipulating the recommender systems' results. Given that recommender systems are used by all major social networks, that they come with a social influence bias, and given that OSNs are not per se designed to foster political discussions, we discuss the implications for the political discourse and the danger of algorithmic manipulation of political communication.

1. Introduction

To day, the internet prevails as a prominent communication and information medium for citizens. Instead of watching TV or reading newspapers, increasing numbers of people become politically informed through online websites, blogs, and social media services. The latest statistics demonstrate that internet as a news source has become as important as television, with its share increasing year by year [1]. Given this shift in the means of news broadcasting, politicians have altered their tactics of communication to the society. OSNs, such as Twitter, Facebook and Instagram, have become a cornerstone of their public profiles as they use OSNs to transmit their activities and opinions on important social issues [2–4].

The growth of online communities on social media platforms have created a public amenable to political campaigning. Political parties and actors have adapted to the new digital environment [5], and besides new campaigning methods such as microtargeting [6], have created microblogs through which they can inform citizens of their views and activities. Furthermore, OSNs have enabled users to respond to or comment on politicians' messages, giving birth to a new type of political interaction and transforming the very nature of political communication itself.

On OSNs, the flow of information from politicians to citizens and back follows a different broadcasting model than the classical one [7]. Instead of journalists monitoring the political activity, political actors themselves produce messages and make them publicly available on the platforms. Each platform provides its users with access to the generated content, as well as distributes it to them through recommendation algorithms [8,9]. The received information is then evaluated both directly or indirectly [10,11]: The political message is interpreted immediately, or subsequently through further social interactions among citizens on the related topics. On OSNs, not only can users respond to politicians in the traditional manner i.e., through their political activity in the society-, but also respond to or comment on the politicians' views online. This creates a new type of interactivity, as users, who actively engage in online discussions sharing their views, are able to influence the way the initial political information will be assimilated by passive users as well as directly influencing political actors.

This new form of political interactivity transforms political communication. Given the possibility of users to directly respond

* Corresponding author.
E-mail addresses: orestis.papakyriakopoulos@tum.de (O. Papakyriakopoulos), juan.medina@tum.de (J.C.M. Serrano), simon.hegelich@hfp.tum.de (S. Hegelich).

https://doi.org/10.1016/j.osnem.2019.100058

to the political content set by political actors, and discuss political issues with other citizens online, OSNs emerge as a fruitful space for agonistic pluralism. They provide the necessary channels for different interests and opinions to be expressed, heard and counterposed; elements that constitute the very essence of political communication. If the discussions held are legitimized within a democratic framework, they form the basis for reaching a conflictual consensus [12], based on which societal decisions can be made. Hence, political communication on OSNs opens new possibilities for citizens to participate in the political shaping of the society, providing them with additional space to address their interests.

1.1. Motivation

Although the above type of political communication has the potential to improve the function of democracy, OSNs possess a structural property that obstructs the unbiased constructive interaction between political actors and citizens: The activities of users on OSNs follow an extreme value distribution [13–16]. Practically, this means that users are not equally active when using a specific OSN. Among others, the majority of users remain passive, or participate with a very low frequency; they either simply read the content or like, comment, tweet, etc. very rarely. On the contrary, a very small number of the users are hyperactive, as they over-proportionally interact with the platform they use. Thus, in political communication on OSNs, hyperactive users are citizens who over-proportionally externalize their political attitudes compared to the mean. This could be done by liking, commenting, tweeting or using any other interaction possibility provided by a platform to declare an attitude to a political issue.

The given activity asymmetry becomes a major issue, considering that a considerable part of the society is politically informed via OSNs. As hyperactive users externalize their political attitudes more than the others, they have the potential to distort political communication; political issues that are important to them become overrepresented on OSNs, while the views of normally active users become less visible. Hence, hyperactive users may influence the political discussions towards their ends, creating a deformed picture of the actual public opinion on OSNs. This fact violates the assumption of an equitable public political discourse as part of political communication [17], because the interests and views of normally active users appear as less important.

The above distortion of political communication is intensified by the business models of the OSN platforms. OSNs were not created to be arenas of political exchange. Their aim is to maximize the number of platform users, by keeping them satisfied [18], and to transform this social engagement to profits, i.a. through advertisement. Hence, on OSNs, users are both consumers and citizens [19]. In order to maximize their profits, OSN platforms adjust their recommendation algorithms to the content popularity, with a view to promoting information that most users will like. For example, the recommendation algorithm of Facebook aims to maximize how well users spend their time on the platform [20]. This is translated to the promotion of content that influences users to engage more with the website, i.e to generate more likes, comments or shares. Facebook also gives further incentives for users to engage more on the platform. It rewards users, giving them badges as 'top commenter' or 'top fan' that appear next to their name. This happens when they are really active on the platform or on a specific page. These awards give users a higher social status on the platform and actually aim the mobilization of users to engage more with the service. This can be then translated to more user data for the platform and consequently, yields better placed advertisement.

As hyperactive users asymmetrically influence the popularity of political content, recommender algorithms might also replicate this asymmetry. A platform might recommend content which is consistent with the political interests of hyperactive users. This phenomenon per se denotes a form of algorithmic manipulation of political communication: The platform unintentionally magnifies hyperactive users' interests, thus posing the risk of political invisibility for passive users [21]. The investigation of such an issue becomes politically important, considering that the most engaged news stories on social media platforms usually come from right-wing media sources and contain provocative content [5,20]. Nevertheless, understanding the interplay between hyperactive users and recommendation systems is not a trivial task. OSNs conceal exactly how their models work, as well as their effects on user behavior. Given these restrictions, the study tries to shed light on potential algorithmic distortions of content, partly by performing simulations that try to replicate real world OSN features.

Last but not least, aforementioned misrepresentations of public opinion have a direct impact on political campaigning. Contemporary political actors develop their influence strategies based on the perceived voter model [22], which presupposes the gathering of demographic and political data for the development of statistical models about the electorate's attitudes. As major part of these data is derived from social media, models that fail to take the effect of hyperactive users into account would face an important bias.

1.2. Research questions

Considering the above, we want to answer following questions:

RQ1: What is the impact of hyperactive users on political agenda-setting on OSNs?

RQ2: Does hyperactive behavior influence recommendation algorithms?

To coherently provide answers to the above questions, we analyze a dataset of posts and user reactions generated on political pages on Facebook. We classify the generated content into topics by applying a topic modeling algorithm and formulate five hypotheses to be tested.

For the investigation of the impact of hyperactive users on political agenda-setting, we want to understand: A. How the user activities are distributed on the political pages. B. If hyperactive users are interested in the same content as regular users. C. If hyperactive users' opinion is equally, more, or less important in the discussions taking place. Therefore we test the following hypotheses:

H1: The log-normal distribution is the optimal distribution for mathematically describing the users' liking and commenting activities on political pages on Facebook.

H2: The distribution of topics in which hyperactive users engage does not deviate from the distribution of topics in which the regular users engage.

H3: Hyperactive users' comments receive on average the same amount of likes as the comments of the rest of the users.

Besides understanding the role of hyperactive users on political discussions, we want to quantify the contribution of hyperactive users on the suggestions made by recommendation systems. We focus on two parts of the recommendation systems. A. Their training process, and how efficient different models are in learning from heavy-tailed distributions of political activities. B. Given that algorithmic recommendations on OSNs depend on the users' friendship network, we investigate how the suggested content can be manipulated by poisoning the friendship graphs, i.e. by adding users that show hyperactive behaviour. We achieve this by testing following hypotheses on a simulated political network:

H4: Given the heavy tailed distribution of activities and data sparsity, standard recommendation algorithms fail to suggest content that correspond to the users' political interests.

H5: Given the interconnectedness of recommender systems and friendship network structure, graph poisoning by the introduction of hyperactive users significantly distorts recommendations of political content.

1.3. Original contribution

- We mathematically define hyperactive users on OSN Facebook, and identify them on the public pages of the major German political parties.
- By applying a state-of-the-art topic modeling algorithm, we investigate whether they spread or like different messages on political issues other than normal users and politicians do. We prove that hyperactive users are not only responsible for a major part of online political discussions, but also externalize different attitudes than the average user, changing the discourse taking place.
- We quantify their effect on content formation by measuring their popularity and showing that they adopt an opinion leader status.
- We simulate a social network of political activities and train hybrid collaborative filtering and deep learning recommendation algorithms on it. We prove that recommendations are strongly influenced by hyperactive behavior. We also show that the election of an appropriate cost function can deal with the issue.
- We insert adversarial examples of hyperactive users and intentionally manipulate the recommender systems' suggestions. Given the above, we initiate an important discussion on OSNs as spaces of political communication.

2. Background and related work

Research proves that user activities on OSNs are not normally distributed. The interactivity time of users with the services, the frequency with which they generate content on the platforms, how often they like, tweet or comment, all follow an extreme value distribution [13,14,23]. This denotes a significant asymmetry existing between users' behaviour, which stops being a mere descriptive behavioral feature when it comes to analyzing phenomena as political communication. The fact that users contribute differently on political discussions and on news diffusion automatically becomes a political phenomenon, which needs to be analyzed by researchers. This is crucial for understanding how political processes take place in the digital era.

The given asymmetry becomes even more important when considering that most OSN users remain passive. They just browse and read political content [16], without contributing in information diffusion [24]. This automatically constitute active and hyperactive users as potential opinion influencers, as they shape the platforms' content. Research proves that users who obtain influential roles in networks have the ability to politically persuade the rest [25] as well as to guide information diffusion [26] in terms of news stories and political discussions [27]. Therefore, it is interesting to investigate and define the differences of normally active and hyperactive users, and their role in shaping political content. Hyperactive behaviour constitutes a widely used tactic in information operations, with automated or human accounts intensively trying to influence discussions taking place [28,29].

The extremely skewed distribution of user activities also becomes an issue when considering the function and impact of recommendation systems. All major OSNs deploy data-intensive algorithms for suggesting to the users contents to interact with, especially on the platforms' news feed. These algorithms consist usually of deep neural systems that take as input user specific features and return contents that maximize the probability of a user interacting with it [30]. Although most OSNs do not disclose the exact structure of their recommender systems, it is known that they use the users' social network structure to improve recommendations' accuracy [31,32]. In this way they take advantage of the homophily principle that states that a user's friends will have similar interests as the user themselves.

Nevertheless, research proves that the training of statistical models and neural architectures on extremely skewed and imbalanced datasets as that of user activities might yield poor accuracy results [33,34]. Therefore, the danger exists that algorithms might fail to learn and diffuse in a coherent way the users' political interests, resulting to the algorithmic manipulation of political communication [21]. To that phenomenon might contribute the fact that recommender systems come with a social influence bias: They actually have the power to alter public opinion [35,36].

Besides having the potential to influence the recommender systems' training process, hyperactive behavior is able to distort the recommendations of already trained models. Researchers prove that adversarial attacks of neural architectures used for recommendation systems can make the models fail. This can be done by both inserting adversarial observations at the level of the model, or by graph poisoning: altering the network structure in a way that specific information will become invisible or be overrepresented [36–40]. Given the number of hyperactive users that explicitly aim the diffusion of political content, it is important to investigate their impact on recommender systems' results.

3. Data and methods

The section provides an overview of the data and methods used to study the impact of hyperactive users on political agenda-setting (RQ1) and to test if hyperactive behavior influences recommendation systems (RQ2). Subsection 3.1 provides an overview of the data used to test the five formulated hypotheses and explains the related limitations. Subsections 3.2-3.4 present the tools used to quantify the presence and activities of hyperactive users on Facebook political pages (RQ1). For answering RQ2, subsections 3.4 and 3.5 provide details on the development of recommendation systems based on user activities and their differences and similarities from the actual algorithms used by OSNs. We describe how the systems make recommendations and how we attack the systems with adversarial examples to potentially distort their recommendations.

3.1. Data description & limitations

To investigate the effect of hyperactive users, we analyse the public Facebook pages of the main German political parties. Our sample includes CDU, CSU, SPD, FDP, Bündnis 90/Die Grünen, Die Linke, and AfD. CDU is the main conservative party of Germany, while CSU is the conservative party active in Bavaria. SPD represents the main German social-democratic party, and Die Linke the radical left. AfD has a nationalist, anti-immigrant, and neo-liberal agenda, while FDP is a conservative, neo-liberal party. Finally, Bündnis 90/Die Grünen is the German green party. We focus on Facebook, because the platform's API restrictions and its monitoring system largely prevent automated activities, such as those performed by social bots on other platforms [41,42]. Therefore we can evaluate the natural behavior of hyperactive users and not an algorithmically generated one.

We take into consideration all party posts and their reactions in the year 2016. This choice is made, because the acquirement of similar data is not feasible today, since Facebook has restricted access to its public API. In total, by accessing the Facebook Graph API in 2017, we collected 3,261 Posts, 3,084,464 likes and 382,768 comments, made by 1,435,826 users. The sample includes all posts

Table 1
Overview of activities on the 7 investigated Facebook pages for the year 2016. Ratio denotes the total number of likes and comments made on each page to the total number of active users.

	Subscribers	Active users	Posts	Comments	Likes	Reactions to users ratio
AfD	309,275	642,927	577	176,506	1,601,502	1,778,008 : 642,927
CDU	123,534	77,400	391	46,837	82,564	129,401 : 77,400
CSU	148,759	404,641	595	80,618	692,273	772,891 : 404,641
Die Grünen	132,630	135,217	209	24,750	153,500	288,717 : 135,217
Die Linke	167,570	135,085	356	14,257	181,161	195,418 : 135,085
FDP	56,581	76,919	579	9,177	126,584	135,761 : 76,919
SPD	121,128	179,970	566	30,630	246,887	277,517 : 179,970

Table 2
Vuong test results.

Log-normal vs	Likes LL-ratio (p-value)	Comments LL-ratio (p-value)
Power-law	15.1 (<0.01)	34.9 (<0.01)
Poisson	34.9 (<0.01)	12.7 (<0.01)
Exponential	12.7 (<0.01)	26.6 (<0.01)

and comments on the posts generated for the period under investigation.

Table 1 provides an overview of the data collected. We only concentrate on 7 German political pages, and we do not investigate the numerous other smaller party pages existing on the platform. Nevertheless, because our sample includes activities of more than 1,4 million people, we can draw consistent inferences about the statistical distribution of users related to political communication. From Table 1 it is visible that activities among the pages are not evenly distributed. This conforms with existing literature stating that AfD users are the most active German partisans online [43]. It also agrees with existing literature stating that OSNs come with a population bias: user preferences on OSNs do not describe accurately preferences of people offline [44]. This asymmetry does not influence our investigation. Although some pages might be more popular than others, and user behaviour might vary in between, our investigation focuses on revealing statistical properties of user activities on the party pages as an integrated ecosystem. This fact also deals with the limitation that we do not have the activities of users on other non-political pages. Our analysis deals only with the interaction of users with political content and wants to uncover features and issues of political communication taking place on OSNs.

3.2. Defining hyperactive users

For testing if the log-normal distribution is the optimal distribution for mathematically describing the users' liking and commenting activities on the investigated pages (H1), we make theoretic assumptions about hyperactive users and choose the formal framework that best complies with their properties.

We consider hyperactive users as people, whose behavior deviates from the standard on an OSN platform. To obtain an understanding of the overall behavior of the users, we fit discrete power-law and extreme value distributions to mathematically describe the users' like and comment activities. Additionally, we run bootstrapped and comparative goodness-of-fit tests based on the Kolmogorov–Smirnov [45] and the Vuong [46] statistic to evaluate the potential fits, as proposed by Clauset et al. [47]. The KS test examines the null hypothesis that the empirical sample is drawn from the reference distribution, while the Vuong test measures the log-likelihood ratio of two distributions and, based on it, investigates whether both empirical distributions are equally far from a third unidentified theoretical one.

In order to mathematically describe the activities of hyperactive users, we select to treat them as outliers of the standard OSN population. We adopt the definitions made by Barnett and Lewis [48], Johnson and Wichern [49] and Ben-Gal [50], and see outliers not as errors, or coming from a different generative process, but as data containing important information, which is inconsistent with and deviating from the remainder of the data-set. Therefore, given the extreme skewed distribution of the activities, we follow the method proposed by Hubert and Vam der Veeken [51] and Hubert and Vandervieren [52] for outlier detection. We calculate the quartiles of our data Q_1 and Q_3, the interquartile range $IQR = Q_3 - Q_1$ and the whiskers w_1 and w_2, which extend from the Q_1 and Q_3 respectively to the following limits:

$$[Q_1 - 1.5e^{-4MC}IQR, Q_3 + 1.5e^{3MC}IQR] \quad (1)$$

where MC is the medcouple [53], a robust statistic of the distribution skewness. Data beyond the whiskers are marked as outliers.

3.3. Topic modeling

After evaluating the likes and comments distributions, as well as identifying the existing hyperactive users, we prepare our data for the application of a topic modeling algorithm. By applying a topic modeling algorithm and assigning user activities to the different topics, we can test if the distribution of topics in which hyperactive users engage does not deviate from the distribution of topics in which the rest of the users engage (H2).

As it has been shown that a noun-only topic modeling approach yields more coherent topic-bags [54], we clean our posts and comments from the remaining part-of-speech types. To do so, we apply the spaCy pretrained convolutional neural network classifier [55] based on the Tiger [56] and WikiNER [57] corpuses, classify each word in our document collection, and keep only the nouns.

We want to investigate the various topics that users and parties discussed but do not want to differentiate on the way they talked about them. Parties usually use a more formal language when posting on a topic than users. Therefore, there is the risk that the topic modeling algorithm would create different topics on the same issue, one for the parties and one for the users. To avoid this, we fit our model on the user comments, and then classify the parties' posts through the trained model.

For our analysis, we apply a non-parametric Conic Scan-and-Cover (CoSAC) algorithm for geometric topic modeling [58]. Our decision is based on the fact that most topic modeling algorithms (e.g. LDA [59], NMF [60]) need a priori as input the number of topics to split the corpus. CoSAC has the advantage of electing itself the number of topics to find the most efficient topic estimates.

The algorithm presupposes that the optimal number of topics K are embedded in a V-1 dimensional probability simplex Δ^{V-1}, where V the number of words in the corpus. Each topic β_K corresponds to a set of probabilities in the word simplex. The totality of topics build a convex polytope $B = conv(\beta_1, ..., \beta_K)$. Each document corresponds to a point $p_m = (p_{m1}, ..., p_{mV})$ inside Polytope B, with $p_m = \sum_k \beta_k \theta_{mk}$. θ_{mk} denotes the proportion that topic k covers in document m. Finally each document is drawn from a multi-

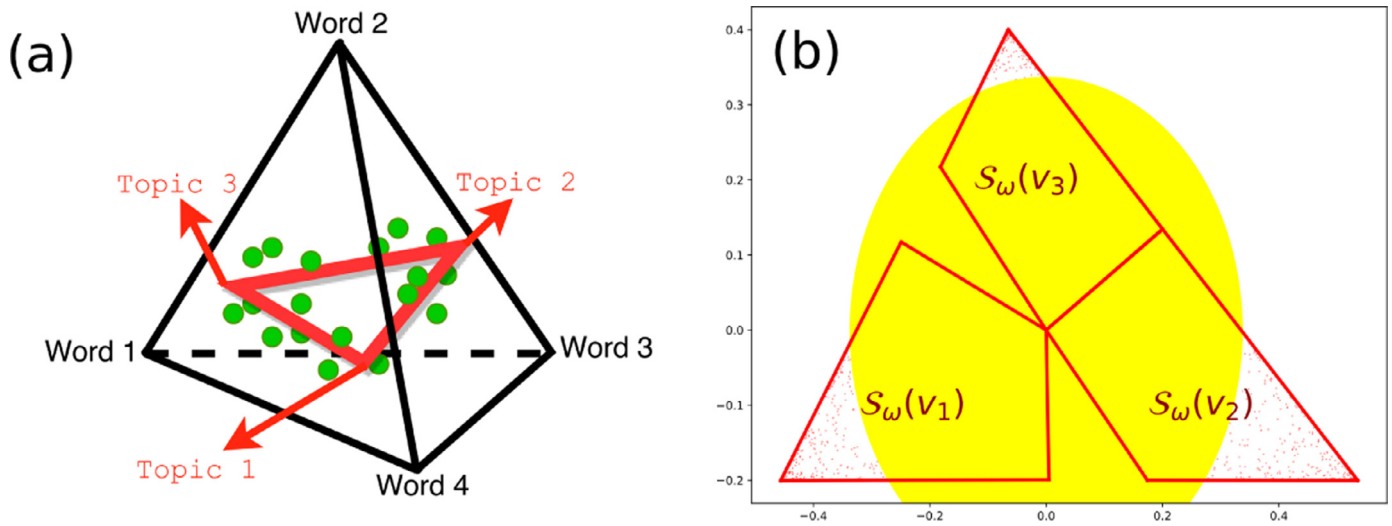

Fig. 1. (a) The topic polytope embedded in the word simplex. (b) Cones and sphere coverage of the polytope (figures from [58]).

nomial distribution of words: $w_m \sim Multinomial(p_m, N_m)$, where N_m the number of words in document m.

The CoSAC algorithm iteratively scans the polytope B and finds the furthest point from its center Cp. It then constructs a conical region with angle ω, starting from Cp and embedding the specific point (Fig. 1). All points within the cone are considered to belong in the same topic and are removed from the polytope. The procedure is iterated K−1 times, until almost no points remain in the polytope. A cone is considered sufficient if it covers at least a proportion of documents λ. After fitting the cones, CoSAC places a sphere with radius R to the polytope, to cover the remaining points. The K geometric objects and their respective points correspond to the K topics created by the algorithm. In our model, the hyperparameters were set to $\omega = 0.6, \lambda = 0.001$ and $R = 0.05$, as proposed by the authors.

3.4. Comparison of activities

Given our topics, we want to evaluate the differences in the activity of normal and hyperactive users (H2). Therefore, we calculate the empirical distributions $f(comment|topic)$ over all topics for the comments of normal and hyperactive users respectively. We pairwise compare the distributions for each topic, by applying a 2-Sample Anderson–Darling Test [61]. The test calculates the probability that the populations from which two groups of data were drawn are identical.

Besides testing the empirical comment-topic distributions, we assign to each comment the topic with the highest probability and compare the most commented topics for normal and hyperactive users. Similarly, we assign the classified party posts to their most probable topic and aggregate the likes of normal and hyperactive users. In this way, we are in the position to locate the concrete political interests of users.

To complete the investigation of RQ1, we calculate the average likes that the comments of hyperactive and normally active user' comments received. In this way, we test H3, namely if hyperactive users' comments receive on average the same amount of likes as the comments of the rest of the users.

3.5. Training recommendation algorithms

In this and the next subsection we describe the methodology followed to answer our second research question: Does hyperactive behavior influence recommendation algorithms? Here, we describe two types of recommendation systems developed to test H4, namely if the heavy tailed distribution of user activities influences standard recommendation algorithms in learning the actual political interests of users.

We explain our methodology and describe differences and similarities between our recommendation systems and the actual ones' used by OSNs. Our algorithms do not fully correspond to the one's actually deployed by Facebook. We are not in the position to do that, because Facebook and most platforms do not disclose the exact systems they use. Nevertheless, by training recommendation algorithms that share common properties with the actual ones, we want to illustrate potential issues related to political communication on OSNs, as well as point out the need for additional transparency on the algorithms used by the platforms and their impact on user behavior.

To test the effect of hyperactive behavior on the suggestions of recommendation algorithms, we simulate a social network of political activities. We use the Facebook social circles dataset developed by [62], which contains a real network of 4,039 Facebook users and their corresponding 88,234 friendship relations. The dataset also provides anonymized user metadata about their age, sex, education, work, hometown and location. The network consists of 10 communities. We merge them into 7, in order to correspond to the number of party pages we investigate. For each community, we only map users active on a specific party page to the nodes, maintaining homophilic features initially existing in the network. Homophily states that similar users tend to form ties with each other [63]. We preserve this similarity by keeping the meta-data and connections from the original dataset and extend it by adding an additional common feature i.e. the political page users are active on.

In total, we create a social network that includes users' political attitudes and nonpolitical properties, while also having a friendship structure corresponding to an actual Facebook sub-network. We do not assign to a user all the posts they reacted on, but use the developed topics as proxies of the user's interests i.e., we assign how many times a user liked a post from a specific topic or commented on a specific topic. We built recommender systems by using the tensorrec library [64] based on the users' meta-data, the user's political interests and the political interests of their friends. In this way we tried to replicate as closely as possible the function of the private recommendation systems of OSN platforms, which take into consideration user features, location settings and the activities of friends in order to tailor news feed suggestions [31,65,66].

Of course, the actual OSNs are trained online and have the actual posts users have interacted with or might interact with in the future, as parameters. They also take as input thousands of features we do not have access to. For example, they take into consideration the pages a user has made a meaningful interaction, regardless of it is related to politics or not. They also include features about the general popularity of content generated on the pages, how much time a user spends on specific content, what information they share on other services of the platforms, or what device a user uses. All these pieces of information are not available to us, and instead of speculating about their impact we leave them out of our analysis. Nevertheless, we train recommendation algorithms on the available data and illustrate the impact of hyperactive behaviour. Training models on topics as proxies of the actual posts does not reduce the credibility of the study, as we want to make inferences about the impact of the models on agenda-setting and their accuracy on suggesting political content (H4,H5).

We train two models, a hybrid collaborative filtering model (HCF) and a deep learning recommendation system (DNN-BWMRB). The first one maintains properties of the recommender systems used on OSNs, i.e. it takes into consideration user features and friends activities, but is not specially designed for extremely skewed input data. By contrast, DNN-BWMRB uses a special cost function to deal with sparse and asymmetric input and output data.

The HCF optimizes the cost function

$$L = \min_{W,V} |P - UW * IV|_F$$

where P the n by t matrix containing the number of interactions of n users on the t different topics; and U the n by m matrix of user meta-data containing m features per user and I the t by n friends' preferences matrix containing for each user how many times their friends interacted about a specific topic. W and V are n by t and t by t parameter matrices to be optimized, $*$ the dot product, and $|.|_F$ the frobenius norm. Similarly, DNN-BWMRB consists of two parallel architectures with 3 fully connected layers of 276 neurons each, which take as inputs matrices U and I^T and minimize the balanced weighted margin-Rank Batch loss [67] (BWMRB) of predicting matrix P. BWMRB optimizes the rank-sensitive cost function

$$L_{BWMRB} = \min \sum_{j=1}^{n} \sum_{k=1}^{t} ln\left(\frac{n_{j,k}}{pop(k)} rank(j,k) + 1\right)$$

where j is a user, k a topic, $n_{j,k}$ the interaction of user j with content related to topic k, where $pop(k)$ is the total number of interactions on topic k. Also, $rank(j, k)$ is the ranked importance of a topic for a user given by the negative sampling equation

$$rank(j,k) = \frac{|K|}{|Z|} \sum_{k' \subset Z} \max(1 - p(j,k) + p(j,k'), 0)$$

where $p(j, k)$, $p(j, k')$ the scores predicted for user j, topic k, user j and a negative topic k' respectively. $|Z|$ is a set of randomly selected negative samples and $|K|$ the set of all topics. We select the specific cost function because it is robust against asymmetric interactions of unique users on the topics, as well as total popularity asymmetries between topics. An overview of the architecture can be found in Fig. 2.

3.6. Designing attacks with adversarial examples

The last part of our study focuses on understanding the vulnerability of recommender systems to attempts of organized manipulation. This concludes the investigation of how hyperactive behavior might influence recommendation systems (RQ2) by testing if graph poisoning by the introduction of hyperactive users significantly distorts recommendations of political content (H5).

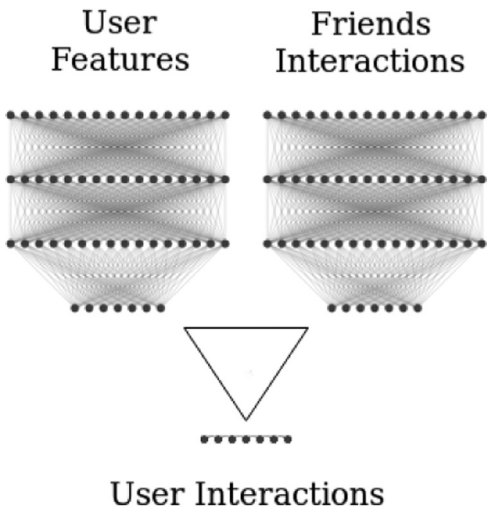

Fig. 2. Structure of DNN-BWMRB, consisting of two parallel neural networks leading to the balanced weighted margin-Rank Batch loss function.

We measure how models' recommendations change in the presence of adversarial examples that attempt to set the political agenda. After training the recommender systems on the initial network, we change the network structure by inserting new users, who have purposefully intensive interest on topic 10 i.e., on the deportation of immigrants. We perform a detailed sensitivity analysis, by varying the number of reactions they make on the topic, the number of users involved in the attack, as well as their position in the network. For that, we calculate the network eigenvalue centrality for each user in the friendship network, and adding the users in positions of high, moderate and low eigenvalue centrality. We sample from the already trained recommendation system's random suggestions, giving as input the initial and the attacked network. As recommendation algorithms take users' friends' interests as input, their suggestions adjust in respect to changes in the friendship network structure. By performing the aforementioned attack, we quantify via counterfactual scenarios how the suggestions made by the systems shift in the presence of adversarial examples.

4. Results

Our results are split in two parts, each one of them dedicated to answer a research question and its related hypotheses. The first part investigates hypotheses related to RQ1, i.e. on the impact of hyperactive users on agenda-setting on OSNs. In the second part we investigate if hyperactive behaviour impacts recommender systems' suggestions (RQ2).

4.1. What is the impact of hyperactive users on political agenda-setting on OSNs?

In order to address RQ1, we present first our findings on the general user distribution of the investigated pages and answer if the log-normal distribution is the optimal one to describe the data (H1). Based on that, we analyze the number and distribution of hyperactive users among the different pages. Then, we compare the behavior between hyperactive and normal users by taking into consideration the topic modeling results and further statistical tests. In this way, we both test whether hyperactive users engage on the same content with the rest (H2) and if they are equally popular to the other users (H3).

As a first result, we identify the log-normal distribution as the best measure for describing user activities (Fig. 3), confirming thus

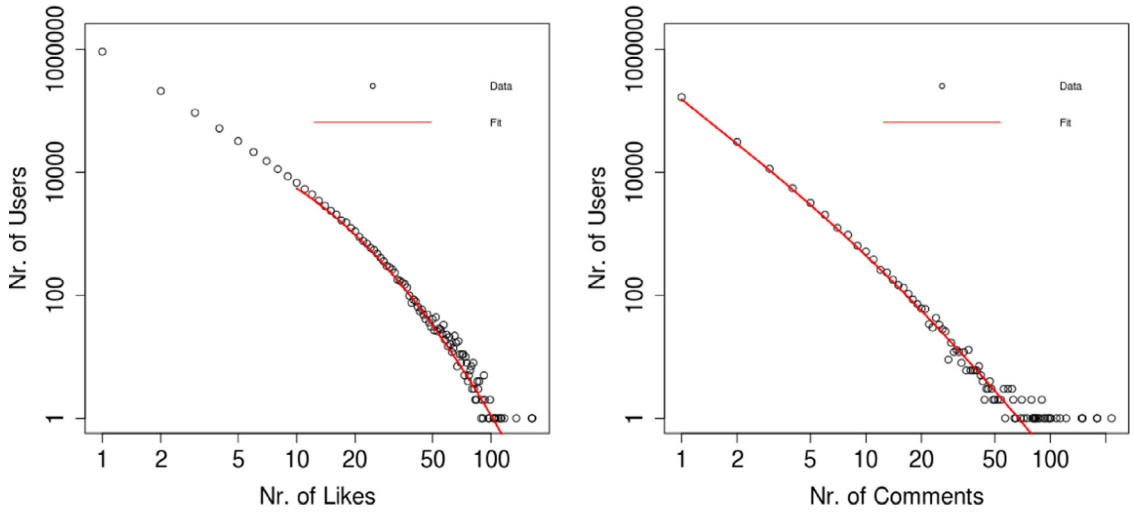

Fig. 3. Distribution of users' commenting and liking activites. The red line representes the log-normal fit.

H1. The bootstrapped KS-Tests (100 samples, two tailed) for both comments and likes fail to reject the null that our data come from a log-normal distribution (gof $<$ 0.01, p $>$ 0.05 and gof $<$ 0.01, p $>$ 0.2 respectively), while comparative Vuong tests show a better fit of the log-normal in comparison to the power-law, poisson and exponential distributions (Table I). Our results comply with the existing literature, which states that usually complex social network properties are log-normally distributed [15,68,69]. Fig. 3 shows the empirical frequencies of user activities and their respective log-normal fits.

After finding the distribution that best describes our data, we categorize users in hyperactive and normally active. Through our outlier detection methodology, we detect 12,295 hyperactive users on the comment section of pages, who correspond to 5.3% of the total users commenting on the pages. Due to the extreme skewness of the comments' distribution, a user is characterized as hyperactive if they made three or more comments. This is justified by the fact that 74% of the users under investigation made only one comment. Although hyperactive users represent 5.3% of the total commenting population, they account for 25.8% of the total comments generated on the parties' pages. Furthermore, 56% of these hyperactive users commented on two or more party pages, denoting that they generally interact with users across the political spectrum. By evaluating the popularity of the users' comments, we find that hyperactive users tend to get more support than the rest. Comments made by hyperactive users on average gained 3.52 likes, while normal users' comments only 3.07, a difference that is statistically significant (Mann-Whitney Test with continuity correction, one tailed: W = 1.4^{10}, $p < 0.01$). Thus, we reject the hypothesis that hyperactive users' comments are liked equally often as normal users' comments (H3). on the contrary, we show that hyperactive users are more appreciated in their contributions. This complies with previous research stating that highly active users tend to have the characteristics of opinion leaders [25].

Similarly, the evaluation of the pages' likes results in the characterization of 61,372 users as hyperactive, or 4.3% of the total users that liked the parties' posts. As before, the methodology labels users as hyperactive if they made three or more likes, because the majority of the active Facebook population rarely interacted with the related pages. The likes of these hyperactive users account for 26.4% of total likes, thus having a major effect on the overall content liked. In addition, 29% of hyperactive users like posts of more than one party, denoting again that their activities are spread over the entire parties' network. The overview of the hyperactive users' commenting and liking activities for each party can be found in Table 3. We also compare the like and comment distributions, by calculating their gini index. The measure provides a proxy for inequality, with 0 denoting perfect equality and 1 extreme inequality. In our case, we calculate a value of 0.35 and 0.45 for the comment and like distribution respectively. This denotes that like activities are more unequally distributed than the comment activities i.e., hyperactive users play a bigger role in the formation of likes. Additionally, the values denote a degree of inequality between normal and hyperactive users, but not an extreme one. Nevertheless this is misleading, because the measure does not take into consideration inactive users. Had that information been included, the gini index would have been much higher in both cases.

Based on the categorization of users as hyperactive or normal, we evaluate the results of the topic modeling algorithm. The model clustered the users' comments into 69 main topics. A major part of the topics concern the refugee crisis of 2016 and the related discussions about Islam. A set of topics aggregate comments on German Chancellor Merkel, on the leaders of other parties, on female and male politicians and the German parties in general. There is one topic summing up comments in English language, as well as a topic containing hyperlinks. Furthermore, the algorithm created policy related topics regarding foreign affairs, as well as the economy and labour market and the state in general. Other topics are related to the German national identity, society and solidarity, and the nature of democracy. Users also talk about family and gender policy, homeland security, transportation and environmental policy. There are topics that include wishes, fear, ironic and aggressive speech, as well as topics aggregating user thoughts. Finally, a set of topics is about political events and communications and a number of topics include comments against mainstream media and the political system.

The GTM algorithm is able to provide a broad picture of the discussion topics on the parties' pages, revealing numerous insights about the way Facebook users commented on the parties' posts. By splitting the comments into two categories, one for the ones generated by hyperactive users and one for the comments of normal users, and by assigning them to the topics to which they were mostly related, we create a a heatmap that provides a qualitative overview on the activities across pages (Fig. 4). The heatmap illustrates which topics are prominent on the users' comments on each political page. Users on the AfD page talk a lot about immigration, on Chancellor Merkel, and the party itself, as well as share a lot of youtube Videos. CDU users are mostly concerned about the

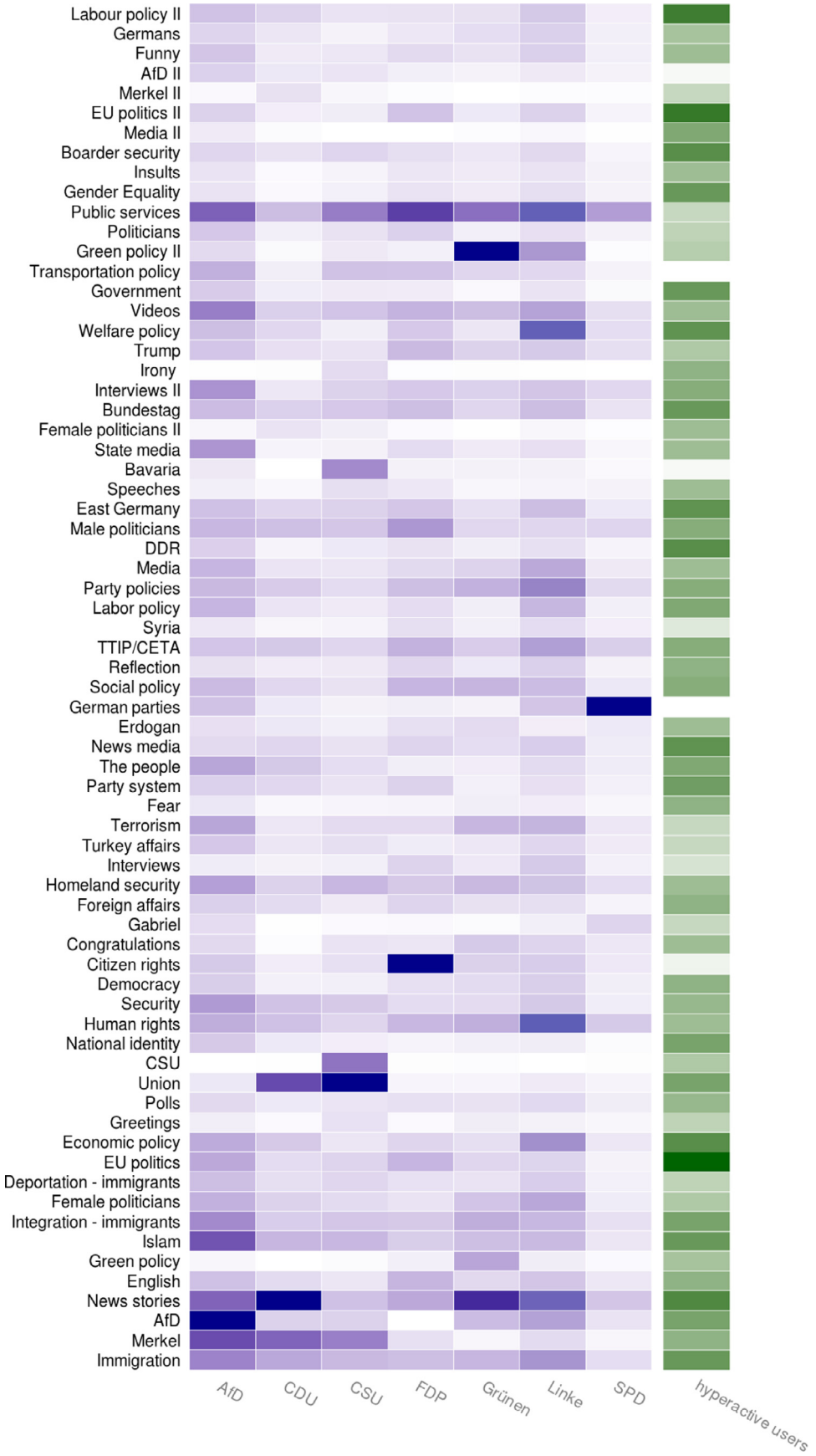

Fig. 4. Heatmap of user comments on the party pages by topic. The heatmap for hyperactive users aggregates both likes and comments.

Table 3
Hyperactive users per party – Likes & comments.

Party	Likes by Hyperactive Users	Likes Ratio	Comments by Hyperactive Users	Comments Ratio
AfD	555,564	0.35	43,017	0.24
CDU	16,997	0.2	20,929	0.45
CSU	139,493	0.2	18,312	0.22
FDP	20,188	0.16	1,400	0.15
Die Grünen	28,777	0.19	8,946	0.36
Die Linke	24,546	0.14	2,343	0.16
SPD	29,057	0.12	3,926	0.13

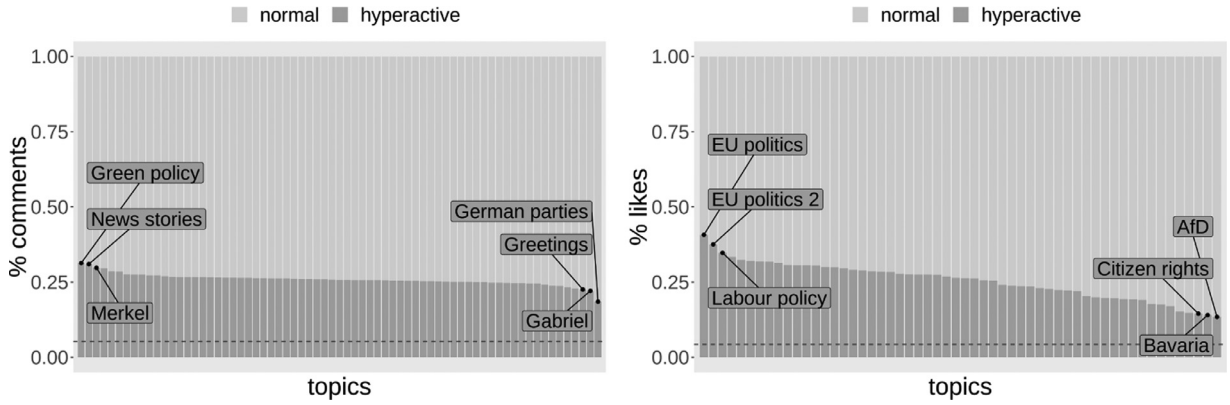

Fig. 5. Proportion of likes and comments generated by normal and hyperactive users. The dotted red line gives the expected proportion of activities for hyperactive users in cases where they were not hyperactive. The plot also illustrates the three most and least interesting topics for hyperactive users.

government and the cooperation with the CSU, topics that are also important for the users on the CSU page. CSU users also discuss a lot about Bavaria, which is the state on which the party is politically based. Users on the FDP page are mostly concerned about citizen rights, while users on the Green Party page about environmental issues. Users on the Left Party page talk about immigration, human rights and welfare policy. Finally, the users on the SPD are focused on party related discussions. Besides illustrating the discussions among pages, the heatmap illustrates how hyperactive users behave. Hyperactive users like and comment more contents related to EU politics and labour policy, and much less on Bavaria, Merkel, citizen rights and on discussions critical against the AfD. It is shown therefore that hyperactive users have their own preferences that do not necessarily concede with the preferences of the rest.

To test if hyperactive users have the same engaging preferences as the rest of the users (H2) we create a stacked chart illustrating the share of hyperactive users' comments and likes for every topic (Fig. 5). It is evident that hyperactive users cover a major part of the comments, and although more active, they comment more or less similarly to the normal users among the various topics. Despite that, the Anderson-Darling tests reject the null hypothesis that hyperactive and normal users' comments come from the same distribution for 54 out of the 69 topics. Practically, this means that the topic density distributions varies between the comments of normal and hyperactive users. This is caused when the comments contain different words in different proportions. Therefore, hyperactive and normal users use different vocabularies when referring to a topic and, consequently, externalize overall different views and sentiment, or focus on different issues in each case.

Besides the fact that hyperactive users have a different behavior on the posts' comments, our analysis shows that they also have different liking preferences. After classifying each party post to the most relevant topic, we count the likes of the posts that belong to each topic. We take into consideration only topics that are based on either political vocabulary or politicians, and ignore topics that contain aggressive speech or sentiments, because the related vocabulary is rarely used by the parties. In contrast to the comments' chart, it is obvious that hyperactive users like specific topics with different intensity than normal users. Even though hyperactive users perform on average 26.4% of the likes, they like much more content related to EU politics and labour policy, while they have less interest on the conservative party AfD, citizens' rights and the region of Bavaria. Therefore, it is clear that hyperactive users influence the like distribution of the public on party posts. Given the statistical analysis of the liking and commenting activities of users, we reject the hypothesis that the distribution of topics in which hyperactive users engage in do not deviate from the distribution of topics in which the rest of the users engage (H2). It must be noted that our analysis gives an overview of the content of posts and comments. It cannot identify sentiment, or specific predispositions of users. For a firm understanding of the issues that are over- or under-represented by hyperactive users an additional extensive analysis is needed, which is beyond the scope of this paper. Our analysis demonstrates that, both on commenting and liking, hyperactive users have a different behavior than the other users.

To further understand hyperactive behavior, we calculate the proportion of comments and likes made by hyperactive users for each post (Fig. 6). Although on average one fourth of the reactions are made by hyperactive users, this ratio is not stable among posts. The proportion of reactions by hyperactive users follows a skewed distribution, denoting that on some posts hyperactive users are almost not active at all, while on some other they are the major content generators. This is especially visible on the distribution of comments, where there are some posts with comments exclusively by hyperactive users, and on a set where hyperactive users are totally absent (around 4% of posts). Most of the posts with no hyperactive users' comments were posts generated on the FDP page. The fact that the above distributions are not normally distributed denotes again that hyperactive users influence asymmetrically the political discourse on the pages. This is an additional finding that reinforces the statement that hyperactive users have an influence on political agenda-setting on OSNs.

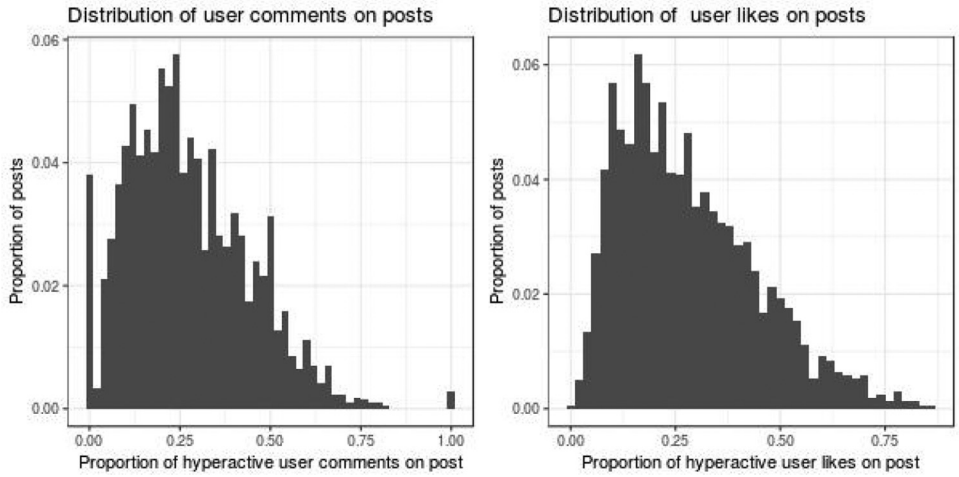

Fig. 6. Distribution of the proportions of comments and likes by hyperactive users on the investigated posts.

Table 4
Accuracy of the various recommender systems for the training and test phase. We report recall@10, NDGC@10 and each model's coverage.

Model	Recall@10 - train	Recall@10 - test	NDGC@10 - train	NDGC@10 - test	coverage - train	coverage - test
CF	0.23	0.2	0.13	0.11	**1**	**1**
HCF	0.24	0.2	0.13	0.13	**1**	0.9
DNN-BWMRB	**0.9**	**0.4**	**0.84**	**0.28**	0.95	0.76

Given the above results, we prove that hyperactive users account for a significant part of the total users' activities (around 25%). They participate in discussions differently from the rest, commenting and liking on different political content than the other users. Moreover, they become opinion leaders, as their comments become more popular than these of the normal users. These findings suggest that hyperactive users have a significant impact on political agenda setting on OSNs. Their preferences and behaviour are different in the discussions taking place, and their opinions are a more valued part of the political discourse.

4.2. Does hyperactive behavior influence recommendation algorithms?

After quantifying the impact of hyperactive users on the political discourse, we illustrate potential effects of hyperactive behaviour on recommendation systems. To do that, we test if the presence of hyperactive users influence recommender systems accuracy on suggesting content that actually corresponds to the political interests of the users (H4). We also test if inserting targeted adversarial attacks on the friendship network (graph poisoning) can result in the distortion of algorithmic recommendations (H5).

Given the extreme skewness of the distribution of user reactions and the high sparsity of the input data (2.9%), we train multiple recommender systems and evaluate their accuracy. We train a baseline collaborative filtering (CF) model that takes only the user friendship network activities as input, a HCF model that has user meta-data and the friendship network activities as input, and a DNN-BWMRB model that has user meta-data, the friendship network activities as inputs, and a cost function accounting for asymmetric user behavior. We split our dataset in train (90%) and test set (10%). To evaluate the models' ability to learn the signal in the dataset and make accurate predictions, we calculate for the training and test set the models' recall@10 and the normalized discounted cumulative gain (NDGC@10). Recall@10 gives the percent of relevant suggestions to the users in the top 10 recommendations. NDGC@10 measures how high in the top 10 recommendations appear the related recommendations (between 0 and 1, with 1 being the best). We also report the coverage of the models, i.e. the percent of classes suggested to the users. Coverage is used to assure that a model does not overfit, predicting only the most popular data classes and not the rest.

Table 4 provides an overview of the results. The baseline CF model yields the worst results, with the HCF providing a slightly better accuracy. Nevertheless, none of the collaborative filtering models is able to learn the true signal in the data, underfitting even at the training phase (CF: Recall@10 - 0.23, HCF: Recall@10 - 0.24). On the contrary, the DNN-BWMRB recommendation system outperforms the other models, being able to make consistent and diverse recommendations (test set: Recall@10 - 0.4, NDGC@10 - 0.28, coverage - 0.76). These values are more than satisfactory, when taking into consideration: (a) the sparsity of the data and the limited amount of observations, (b) the number of labels (69 topics) and the fact that we created our dataset by comprising features from two different user networks.

To further evaluate if the models are able to learn the actual political user interests, we treat the trained systems as generative models and sample from them 69000 recommendations, giving as input the training and the test set. Then, we calculate the proportion of suggestions for each topic and compare it with the actual reactions made by the users. Fig. 7 illustrates how accurately the models learn and replicate the users political interests. As it is visible, the DNN-BWMRB model is able to capture the actual user interests both in the training and test sets, making recommendations that are close to the actual distribution in the data. On the contrary, the HCF only moderately learns the actual distribution of the training set, while failing to generate accurate recommendations for the test set.

The inability of HCF to cope with data sparsity and skewness becomes apparent when sampling recommendations from models with and without hyperactive users (Table 5). While the mean topic proportion is around 1.4%, the M.A.E for the HCF models recommendations is higher (1.7% and 1.9% with and without hyperactive users respectively). This denotes that the models are not able to provide consistent recommendations, proposing significantly different topics to the users for each case. For example,

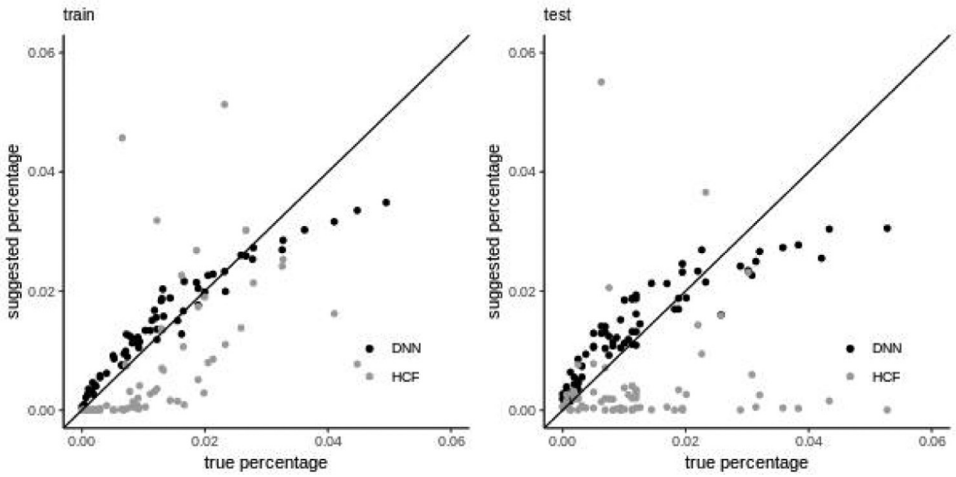

Fig. 7. Accuracy of the two different algorithms trained on the total friendship network containing hyperactive users and taking into consideration n=1403 person related features.

Table 5
Models' accuracy on the test set when taking into consideration or ignoring hyperactive users. The actual mean topic proportion for the users' interests is provided as a reference point.

	w. hyperactive users	w/o hyperactive users
mean topic proportion	0.014	0.014
HCF M.A.E.	0.017	0.019
DNN-BWRB M.A.E.	0.005	0.005

content containing news stories, and issues related to economic, labor, and green poicy would be over-represented by the model trained on all data compared to the model trained only on normal users. The opposite applies for contents related to immigration, citizen rights and youtube videos, which would have been significantly under-represented. To capture which content would have been recommended at each case, we sample posts given the distribution of recommended topics and create wordclouds showing the most prominent tokens for each case (Fig. 8). For the model including hyperactive users, the suggestions would over proportionally contain the token geld (money), together with youtube videos and content about Merkel. It would also suggest content related to the government (Regierung), and immigration (Flüchtling) more often that the model not accounting for hyperactive users. The latter would overproportionally suggest content about justice (Recht), youtube videos, content about Merkel and content containing the token "politik". It would also suggest more often content about the public (Volk) and the police in relation to model including hyperactive users.

Given that, it is clear that such a recommendation algorithm is unable to replicate interactions taking place on OSNs, neither accurately, nor consistently, as including or ignoring hyperactive users yields significantly different recommendations. This fact together with the low accuracy of CF models leads us to confirm H4: Standard recommendation systems fail to suggest content that correspond to the users' political interests, because of the heavy tailed distribution of activities and data sparsity.

On the contrary, the DNN-BWMRB architecture is able not only to provide higher accuracy, but also in a consistent and reliable way. Recommendations made when ignoring or accounting for hyperactive users are similar to each other, with the deviations of predictions from the true distribution of interactions being more or less the same (M.A.E. - 0.005). As the wordclouds in Fig. 8 suggest, both models recommend content related to Merkel, the country, and content from media agencies. They equally promote content related to the police and the euro. Nevertheless, they also illustrate some differences: Token "Politik" is favored by the model with hyperactive users, while token "Partei" is favored by the model without hyperactive users.

Overall, the DNN-BWMRB model is able to capture the signal in the data successfully. Nonetheless, although in a much less degree, the existence of hyperactive users leads the model to provide slightly different suggestions.

As the DNN-BWMRB model is able to capture and replicate the actual user interactions acceptably, we test its predictions when faced with politically malicious attacks (H5). First, we calculate the centrality of each user in the network. Then we insert an adversarial example at the position of the user, which is only engaging to content from topic 10. Next, we calculate the proportion of model suggestions related to topic 10. We repeat the process for different amounts of reactions (10–40) and for different amount of users (1–3). We calculate the bootstrapped mean (50 times, 100 users without replacement) and the maximum value suggested by the model for each setting.

Fig. 9 provides an overview of the results. It is clear that the insertion of adversarial examples influences the suggestions of the recommender systems. The more targeted reactions a malicious user performs, the more relevant suggestions are made by the system. Similarly, the more malicious users we place in the network, the more successful is the distortion of the recommendations to our ends. A simple attack of two users reacting 50 times can lead the recommender system in the network to suggest topic 10 80% of the times. The results of the placed attacks confirm H5, i.e. that graph poisoning can successfully distort recommendation systems suggestions.

By confirming H4 and H5, the study shows that recommendation algorithms are sensitive to hyperactive behaviour. We locate issues regarding the ability of the models to learn from extremely skewed sparse data and make coherent suggestions to the users. We also illustrate how simple it is to manipulate models' inputs in a way that distorts models' recommendations.

5. Discussion

Given that activity asymmetries and recommendation algorithms are a feature of OSNs, it is important to evaluate the consequences for both science and the wider society. Although our analysis was concentrated on Facebook, previous research has proven that hyperactive accounts, either human or automated, have the potential to equally influence political communication on other

(a) HCF with hyperactive users

(b) HCF without hyperactive users

(c) DNN-BWMRB with hyperactive users

(d) DNN-BWMRB without hyperactive users

Fig. 8. Most suggested tokens for different recommender models.

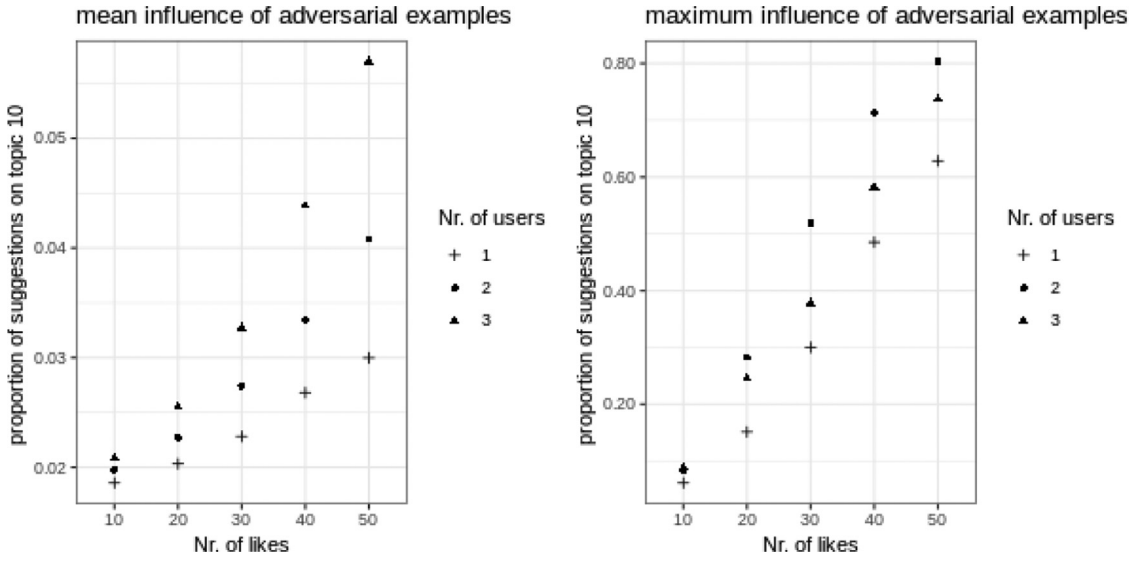

Fig. 9. Mean and maximum proportion of content belonging to Topic 10 after placing adversarial examples.

platforms [70,71]. The specific formation and distribution of political interactions on OSNs raises various questions regarding the role and impact of OSNs on a political level, on an algorithmic level, as well as on the intersection of both.

In the political dimension, the OSN activity asymmetries are transformed into an asymmetry of disseminated political content, as the attitudes and interests of hyperactive users appear disproportionately in the discussions taking place. Until now, research [72,73] has stated that OSNs suffer from a population bias: The people using OSNs are not representative of the actual society. On top of that, a content bias is now added: The content disseminated on OSNs is not even representative of the mean users' attitudes on the platform. We showed that hyperactive users have different attitudes than the rest, and different engaging behavior, altering how the public opinion appears macroscopically. This poses a scientific problem, as it might lead to erroneous research results. Equally important, it poses a political problem, because political discussions and opinion exchange is distorted by the effect of hyperactive accounts. This does not happen because the diffused information in the network is transformed or changed, but because hyperactive users strongly contribute to the type of information diffused. Their attitudes fill the communication space, leading to a bias on the political feedback to politicians, and to a shift on the issues that shape the political agenda. Last but not least, they acquire the status of opinion leaders. Given the above, although OSNs provide a more open environment to express opinions than traditional media, it ends up becoming partly a gathering of political echoes [74] that struggle to be imposed on each other.

In the algorithmic dimension, the extreme skewness of the activity distributions raises specific issues regarding the recommendation algorithms used by OSN platforms. The first problem is related to algorithmic accuracy: skewed data are imbalanced data, and their raw use, either as input features or as output labels, can yield weak classification results. The imbalanced learning problem applies to both standard statistical algorithms, collaborative filtering and neural networks [33,75,76], with algorithms over-estimating the importance of outliers and under-estimating the importance of the rest. This also happens in the case of a poor selection of a cost function [77]. Furthermore, it is proven that statistical models as Markov-chains might fail to capture the signal immanent in highly skewed data, while deep learning methods might face the same issue given power-law distributions of data [34]. We showed that collaborative filtering models come with the same weaknesses. In addition, we showed that the election of a proper cost-function can resolve a large part of the issue.

The second potential problem is that an algorithm might fail, not in the sense that it might be unable to learn from the data, but rather, it might learn the wrong signal. Hyperactive users can be seen as physical adversaries [78] of the mean user attitudes. Algorithms trained in the full data will include the bias, tracking and predicting behavioral associations that correspond to hyperactive users rather than to the population majority. Our study showed that even well trained recommendation systems will have slight deviations in their predictions when accounting for or ignoring hyperactive users. We also illustrated that recommendations of already trained models can easily be distorted by the addition of adversaries of hyperactive behavior. It is not coincidental that the detection of adversaries in machine learning can be done by sample distribution comparison [79], in the same way as we tracked the different preferences of hyperactive users.

Solutions to the aforementioned issues exist, and they were partially shown in the above study and are usually taken into consideration by data scientists, who develop recommendation algorithms. Nevertheless, in the case of political communication, an algorithmic issue automatically becomes a political one. Recommendation systems come with a social influence bias [35,36] i.e., have the power to change users' opinion. Hence, OSNs promoting biased political content might result in the algorithmic manipulation of political communication. This manipulation can be traced back either on the low accuracy of a model, on the attitudes of hyperactive users, or at the structure of the recommender systems themselves.

Social media platforms are not designed to foster political discourses [80], but rather aim at increasing active users, in order to sell advertising and attract funding from venture capitalists [81]. As part of these ends, the structure and impact of recommendation algorithms used distorts human behavior [44], having transformative effects that were not foreseen a priori [82]. For example, it has been stated that changes of the Facebook recommendation algorithm not only resulted in the diffusion of more right-wing content on the platform [20], but also altering in general the way users interact with on the service.

It is evident from the above, that each algorithm mediates and redefines the importance of political interests [83], raising further questions about the opacity of the recommendation systems [84]. In a political context, it becomes important to know as citizens, how, why and with what impact algorithms change political communication. This presupposes awareness of the data processed and, the mathematical method applied, as well as knowledge of what exactly a machine learning cost function optimizes and to what extent recommendation systems alter human behavior. Proposals for algorithmic transparency have already been made [85–87], and wait to be applied in practice.

Until now, most platforms do not report what measures they take to assure the consistency of their recommendations, as well as the exact effect their systems have on human behavior. They do not disclose the inputs and outputs of their models, and provide no information on how recommendations are influenced by malicious organized attacks. Our study tried to illustrate potential effects by quantifying the impact of hyperactive accounts and attacks on simulated case studies. Nevertheless, there is a need for transparency that can illustrate whether the actual recommendation systems used suffer from similar problems.

The above issues need to be extensively analyzed, in order to evaluate and shape the structure of political communication in the digital era. In this paper we laid the foundations for this discussion, by defining, demonstrating and quantifying the effect of hyperactive users on OSNs, through the example of Facebook. We also illustrated and defined the risks of algorithmic manipulation by the OSN recommendation systems. Future research needs to focus on the aforementioned consequences, evaluate the structure of OSNs ethically, politically and normatively as political intermediators, as well as propose and apply solutions to the newly posed problems.

6. Conclusion

Our study showed that political communication on German political Facebook pages is constituted by the behavior of hyperactive users. By describing the users' liking and commenting activities on the pages, we characterized users as hyperactive or normal through outlier detection. We showed that hyperactive users account for a significant part of the total users' activities; they participate in discussions differently from the rest, and they like different content. Moreover, they become opinion leaders, as their comments become more popular than these of the normal users. Taking a set of Facebook pages as an example, we showed that user activities on OSNs might neither be equally nor evenly distributed.

We also studied the influence of hyperactive users on recommendation systems. Standard factorization models are not able to coherently describe the skewed distribution of activities on OSNs. On the other hand, models that balance these asymmetries in their cost function provide results that are closer to the population char-

acteristics. Nevertheless, hyperactive behavior still plays a role on the recommendations generated by the systems. Finally, by adding adversarial examples in the trained models, we illustrated that it is relatively easy to bias recommendations, and potentially distort the political agenda.

Our recommender systems analysis tried to simulate properties of the actual recommender systems used on OSNs. This does not per se mean that the actual systems face the same problems, nor that they function in the exact same way. The aim of the study was to illustrate potential issues related to political communication that might emerge by the application of recommender systems. Given these problems, we want to point out the need for additional transparency on the actual data-intensive recommendation systems used by the platforms, in order to resolve potential algorithmic interferences in the political discourse.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Orestis Papakyriakopoulos: Writing - review & editing, Writing - original draft, Data curation. **Juan Carlos Medina Serrano:** Data curation, Writing - review & editing. **Simon Hegelich:** Writing - review & editing.

References

[1] J. Gottfried, E. Shearer, Americans' online news use is closing in on tv news use, 2017, http://www.pewresearch.org/fact-tank/2017/09/07/americans-online-news-use-vs-tv-news-use/.

[2] S. Hegelich, M. Shahrezaye, The communication behavior of german MPS on twitter: preaching to the converted and attacking opponents, Eur. Policy Anal. 1 (2) (2015) 155–174.

[3] G.S. Enli, E. Skogerbø, Personalized campaigns in party-centred politics: Twitter and Facebook as arenas for political communication, Inf. Commun. Soc. 16 (5) (2013) 757–774.

[4] V. Arnaboldi, A. Passarella, M. Conti, R. Dunbar, Structure of ego-alter relationships of politicians in twitter, J. Comput. Mediated Commun. 22 (5) (2017) 231–247, doi:10.1111/jcc4.12193.

[5] J.C.M. Serrano, S. Hegelich, M. Shahrezaye, O. Papakyriakopoulos, Social Media Report: The 2017 German Federal Elections, 1 ed., TUM.University Press, 2018.

[6] O. Papakyriakopoulos, S. Hegelich, M. Shahrezaye, J.C.M. Serrano, Social media and microtargeting: Political data processing and the consequences for germany, Big Data Soc. 5 (2) (2018), doi:10.1177/2053951718811844. 2053951718811844

[7] M.E. McCombs, D.L. Shaw, The agenda-setting function of mass media, Public Opin. Q. 36 (2) (1972) 176–187.

[8] E. Bakshy, S. Messing, L.A. Adamic, Exposure to ideologically diverse news and opinion on Facebook, Science 348 (6239) (2015) 1130–1132.

[9] Twitter, 2018, (https://help.twitter.com/en/using-twitter/twitter-trending-faqs). Online; accessed 24 August 2018.

[10] M. Hilbert, J. Vásquez, D. Halpern, S. Valenzuela, E. Arriagada, One step, two step, network step? Complementary perspectives on communication flows in twittered citizen protests, Soc. Sci. Comput. Rev. 35 (4) (2017) 444–461.

[11] S. Choi, The two-step flow of communication in twitter-based public forums, Soc. Sci. Comput. Rev. 33 (6) (2015) 696–711.

[12] C. Mouffe, The Democratic Paradox, Verso, 2000.

[13] N. Blenn, P. Van Mieghem, Are human interactivity times lognormal?, arXiv preprint (2016).

[14] P. Van Mieghem, N. Blenn, C. Doerr, Lognormal distribution in the digg online social network, Eur. Phys. J. B 83 (2) (2011) 251.

[15] K. Lerman, R. Ghosh, Information contagion: An empirical study of the spread of news on digg and twitter social networks, in: Proceedings of the Fourth International AAAI Conference on Web and Social Media, AAAI, 2010, pp. 90–97.

[16] F. Benevenuto, T. Rodrigues, M. Cha, V. Almeida, Characterizing user behavior in online social networks, in: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, ACM, 2009, pp. 49–62.

[17] A. Schaap, Agonism in divided societies, Philos. Soc. Crit. 32 (2) (2006) 255–277.

[18] N. Shi, M.K. Lee, C.M. Cheung, H. Chen, The continuance of online social networks: how to keep people using Facebook? in: Proceedings of the Fourty-Third Hawaii International Conference on System Sciences, IEEE, 2010, pp. 1–10.

[19] C.R. Sunstein, # Republic: Divided Democracy in the Age of Social Media, Princeton University Press, 2018.

[20] N. Analytics, The 2019 guide to Facebook publishing, http://go.newswhip.com/2019_03-FacebookPublishing_LP.html.

[21] T. Bucher, Want to be on the top? Algorithmic power and the threat of invisibility on Facebook, New Media Soc. 14 (7) (2012) 1164–1180.

[22] E.D. Hersh, Hacking the Electorate: How Campaigns Perceive Voters, Cambridge University Press, 2015.

[23] C. Boldrini, M. Toprak, M. Conti, A. Passarella, Twitter and the press: an ego–centred analysis, in: Proceedings of the The Web Conference, International World Wide Web Conferences Steering Committee, 2018, pp. 1471–1478.

[24] D.M. Romero, W. Galuba, S. Asur, B.A. Huberman, Influence and passivity in social media, in: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2011, pp. 18–33.

[25] B.E. Weeks, A. Ardèvol-Abreu, H. Gil de Zúñiga, Online influence? Social media use, opinion leadership, and political persuasion, Int. J. Public Opin. Res. 29 (2) (2017) 214–239.

[26] V. Arnaboldi, M. Conti, A. Passarella, R.I. Dunbar, Online social networks and information diffusion: the role of ego networks, Online Soc. Netw. Media 1 (2017) 44–55.

[27] S. Priya, R. Sequeira, J. Chandra, S.K. Dandapat, Where should one get news updates: Twitter or Reddit, Online Soc. Netw. Media 9 (2019) 17–29.

[28] S. Hegelich, D. Janetzko, Are social bots on twitter political actors? Empirical evidence from a ukrainian social Botnet, in: Proceedings of the Tenth International AAAI Conference on Web and Social Media, 2016, pp. 1–4.

[29] J. Weedon, W. Nuland, A. Stamos, Information operations and Facebook, Retrieved from Facebook: https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf (2017).

[30] P. Covington, J. Adams, E. Sargin, Deep neural networks for Youtube recommendations, in: Proceedings of the 10th ACM Conference on Recommender Systems, ACM, 2016, pp. 191–198.

[31] The Facebook algorithm explained and how to work it, https://www.brandwatch.com/blog/the-facebook-algorithm-explained/.

[32] Q. Yuan, S. Zhao, L. Chen, Y. Liu, S. Ding, X. Zhang, W. Zheng, Augmenting collaborative recommender by fusing explicit social relationships, in: Proceedings of the Workshop on Recommender Systems and the Social Web, RECSYS, 2009, pp. 1–8.

[33] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 9 (2008) 1263–1284.

[34] H.W. Lin, M. Tegmark, Criticality in Formal Languages and Statistical Physics, Entropy 19 (2017) 299, doi:10.3390/e19070299.

[35] S. Krishnan, J. Patel, M.J. Franklin, K. Goldberg, A methodology for learning, analyzing, and mitigating social influence bias in recommender systems, in: Proceedings of the 8th ACM Conference on Recommender systems, ACM, 2014, pp. 137–144.

[36] D. Cosley, S.K. Lam, I. Albert, J.A. Konstan, J. Riedl, Is seeing believing?: How recommender system interfaces affect users' opinions, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2003, pp. 585–592.

[37] X. Chen, C. Liu, B. Li, K. Lu, D. Song, Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning, http://arxiv.org/abs/1712.05526 [Cs] (2017). (accessed December 19, 2019).

[38] B. Li, Y. Wang, A. Singh, Y. Vorobeychik, Data poisoning attacks on factorization-based collaborative filtering, in: Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 1885–1893.

[39] A. Bojchevski, S. Günnemann, Adversarial attacks on node embeddings via graph poisoning, in: Proceedings of the International Conference on Machine Learning, 2019, pp. 695–704.

[40] D. Zügner, S. Günnemann, Adversarial Attacks on Graph Neural Networks via Meta Learning, http://arxiv.org/abs/1902.08412 [Cs, Stat] (2019). (accessed December 19, 2019).

[41] Facebook, Using the graph API – documentation, https://developers.facebook.com/docs/graph-api/using-graph-api/.

[42] O. Varol, E. Ferrara, C.A. Davis, F. Menczer, A. Flammini, Online human-bot interactions: detection, estimation, and characterization, in: Proceedings of the Eleventh International AAAI Conference on Web and Social Media, 2017, pp. 1–7.

[43] J.C.M. Serrano, M. Shahrezaye, O. Papakyriakopoulos, S. Hegelich, The rise of Germany's AFD: a social media analysis, in: Proceedings of the 10th International Conference on Social Media and Society, in: SMSociety '19, ACM, New York, NY, USA, 2019, pp. 214–223, doi:10.1145/3328529.3328562.

[44] D. Ruths, J. Pfeffer, Social media for large studies of behavior, Science 346 (6213) (2014) 1063–1064.

[45] T.B. Arnold, J.W. Emerson, Nonparametric goodness-of-fit tests for discrete null distributions., R J. 3 (2) (2011).

[46] Q.H. Vuong, Likelihood ratio tests for model selection and non-nested hypotheses, Econom. J. Econom. Soc. 57 (2) (1989) 307–333.

[47] A. Clauset, C.R. Shalizi, M.E. Newman, Power-law distributions in empirical data, SIAM review 51 (4) (2009) 661–703.

[48] V. Barnett, T. Lewis, Outliers in Statistical Data, Wiley, 1974.

[49] R. Johnson, D. Wichern, Applied Multivariate Statistical Analysis., Pretice Hall, 1998.

[50] I. Ben-Gal, Outlier detection, in: Data Mining and Knowledge Discovery Handbook, Springer, 2005, pp. 131–146.

[51] M. Hubert, S. Van der Veeken, Outlier detection for skewed data, J. Chemom. J. Chemom. Soc. 22 (3–4) (2008) 235–246.

[52] M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions, Comput. Stat. Data Anal. 52 (12) (2008) 5186–5201.

[53] G. Brys, M. Hubert, A. Struyf, A robust measure of skewness, J. Comput. Gr. Stat. 13 (4) (2004) 996–1017.
[54] F. Martin, M. Johnson, More efficient topic modelling through a noun only approach, in: Proceedings of the Australasian Language Technology Association Workshop 2015, 2015, pp. 111–115.
[55] M. Honnibal, M. Johnson, An improved non-monotonic transition system for dependency parsing, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1373–1378.
[56] S. Brants, S. Dipper, P. Eisenberg, S. Hansen-Schirra, E. König, W. Lezius, C. Rohrer, G. Smith, H. Uszkoreit, Tiger: linguistic interpretation of a German corpus, Res. Lang. Comput. 2 (4) (2004) 597–620.
[57] J. Nothman, N. Ringland, W. Radford, T. Murphy, J.R. Curran, Learning multilingual named entity recognition from wikipedia, Artif. Intell. 194 (2013) 151–175.
[58] M. Yurochkin, A. Guha, X. Nguyen, Conic scan-and-cover algorithms for nonparametric topic modeling, in: Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 3878–3887.
[59] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[60] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788.
[61] F.W. Scholz, M.A. Stephens, K-sample Anderson–darling tests, J. Am. Stat. Assoc. 82 (399) (1987) 918–924.
[62] J. Leskovec, J.J. Mcauley, Learning to discover social circles in ego networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 539–547.
[63] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annu. Rev. Sociol. 27 (1) (2001) 415–444.
[64] J. Kirk, Tensorrec: a tensorflow recommendation algorithm and framework in python, 2019, https://github.com/jfkirk/tensorrec [Online; accessed 04.05.2019].
[65] J. Constine, How facebook news feed works techcrunch, 2016, https://techcrunch.com/2016/09/06/ultimate-guide-to-the-news-feed/
[66] Twitter trends Faqs, https://help.twitter.com/en/using-twitter/twitter-trending-faqs.
[67] K. Liu, P. Natarajan, WMRB: Learning to Rank in a Scalable Batch Training Approach, http://arxiv.org/abs/1711.04015 [Cs, Stat] (2017). (accessed December 19, 2019).
[68] K. Sun, Explanation of log-normal distributions and power-law distributions in biology and social science, Technical Report, Department of Physics, University of Illinois at Urbana-Champaign, 2004.
[69] H. Kuninaka, M. Matsushita, Statistical properties of complex systems-lognormal and related distributions, in: Proceedings of the AIP Conference Proceedings, 1468, AIP, 2012, pp. 241–251.
[70] A. Thieltges, O. Papakyriakopoulos, J.C.M. Serrano, S. Hegelich, Effects of Social Bots in the Iran-Debate on Twitter, http://arxiv.org/abs/1805.10105 [Cs] (2018). (accessed December 19, 2019).
[71] C. Shao, G.L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, F. Menczer, The spread of low-credibility content by social bots, Nat. Commun. 9 (1) (2018) 4787.
[72] T. Correa, A.W. Hinsley, H.G. De Zuniga, Who interacts on the web?: the intersection of users personality and social media use, Comput. Hum. Behav. 26 (2) (2010) 247–253.
[73] M. Duggan, J. Brenner, The demographics of social media users - 2012, Technical Report, Pew Research Center's Internet & American Life Project Washington, DC, 2013.
[74] Y.-R. Lin, J.P. Bagrow, D. Lazer, More voices than ever? Quantifying media bias in networks, in: Proceedings of the Fifth International AAAI Conference on Web and Social Media, AAAI, 2011, pp. 1–7.
[75] Z.-H. Zhou, X.-Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Trans. Knowl. Data Eng. 18 (1) (2006) 63–77.
[76] R. Pan, Y. Zhou, B. Cao, N.N. Liu, R. Lukose, M. Scholz, Q. Yang, One-class collaborative filtering, in: Proceedings of the IEEE Eighth International Conference on Data Mining, IEEE, 2008, pp. 502–511.
[77] C.-C. Lee, P.-C. Chung, J.-R. Tsai, C.-I. Chang, Robust radial basis function neural networks, IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) 29 (6) (1999) 674–685.
[78] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world, http://arxiv.org/abs/1607.02533 [Cs, Stat] (2017). (accessed December 19, 2019).
[79] K. Grosse, P. Manoharan, N. Papernot, M. Backes, P. McDaniel, On the (Statistical) Detection of Adversarial Examples, http://arxiv.org/abs/1702.06280 [Cs, Stat] (2017). (accessed December 19, 2019).
[80] S. Engelmann, J. Grossklags, O. Papakyriakopoulos, A democracy called Facebook? Participation as a privacy strategy on social media, in: Proceedings of the Annual Privacy Forum, Springer, 2018, pp. 91–108.
[81] M. Falch, A. Henten, R. Tadayoni, I. Windekilde, Business models in social networking, in: Proceedings of the CMI International Conference on Social Networking and Communities, 2009, pp. 1–7.
[82] B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: Mapping the debate, Big Data Soc. 3 (2) (2016) 2053951716679679.
[83] H. Nissenbaum, From preemption to circumvention: if technology regulates, why do we need regulation (and vice versa), Berk. Technol. Law J. 26 (2011) 1367.
[84] J. Burrell, How the machine thinks: understanding opacity in machine learning algorithms, Big Data Soc. 3 (1) (2016) 2053951715622512.
[85] C. Sandvig, K. Hamilton, K. Karahalios, C. Langbort, Auditing algorithms: research methods for detecting discrimination on internet platforms, in: Data and Discrimination: Converting Critical Concerns into Productive Inquiry, 2014, pp. 1–23.
[86] N. Diakopoulos, Algorithmic-accountability: the investigation of Black Boxes, Technical Report, Tow Center for Digital Journalism, 2014.
[87] F. Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information, Harvard University Press, 2015.

5 Word embeddings and unfair algorithms

5.1 Bias in word embeddings

Authors

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, Fabienne Marco

In

In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 446–457. DOI:https://doi.org/10.1145/3351095.3372843

Abstract

Word embeddings are a widely used set of natural language processing techniques that map words to vectors of real numbers. These vectors are used to improve the quality of generative and predictive models. Recent studies demonstrate that word embeddings contain and amplify biases present in data, such as stereotypes and prejudice. In this study, we provide a complete overview of bias in word embeddings. We develop a new technique for bias detection for gendered languages and use it to compare bias in embeddings trained on Wikipedia and on political social media data. We investigate bias diffusion and prove that existing biases are transferred to further machine learning models. We test two techniques for bias mitigation and show that the generally proposed methodology for debiasing models at the embeddings level is insufficient. Finally, we employ biased word embeddings and illustrate that they can be used for the detection of similar biases in new data. Given that word embeddings are widely used by commercial companies, we discuss the challenges and required actions towards fair algorithmic implementations and applications.

Contribution of thesis author

Theoretical design, model design and analysis, manuscript writing, revision and editing

Bias in Word Embeddings

Orestis Papakyriakopoulos
Technical University of Munich
Munich, Germany
orestis.p@tum.de

Simon Hegelich
Technical University of Munich
Munich, Germany
simon.hegelich@hfp.tum.de

Juan Carlos Medina Serrano
Technical University of Munich
Munich, Germany
juan.medina@tum.de

Fabienne Marco
Technical University of Munich
Munich, Germany
fabienne.marco@tum.de

ABSTRACT

Word embeddings are a widely used set of natural language processing techniques that map words to vectors of real numbers. These vectors are used to improve the quality of generative and predictive models. Recent studies demonstrate that word embeddings contain and amplify biases present in data, such as stereotypes and prejudice. In this study, we provide a complete overview of bias in word embeddings. We develop a new technique for bias detection for gendered languages and use it to compare bias in embeddings trained on Wikipedia and on political social media data. We investigate bias diffusion and prove that existing biases are transferred to further machine learning models. We test two techniques for bias mitigation and show that the generally proposed methodology for debiasing models at the embeddings level is insufficient. Finally, we employ biased word embeddings and illustrate that they can be used for the detection of similar biases in new data. Given that word embeddings are widely used by commercial companies, we discuss the challenges and required actions towards fair algorithmic implementations and applications.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → **Machine learning**; • **Information systems** → **Data mining**.

KEYWORDS

word embeddings, bias, detection, diffusion, mitigation, fairness, sexism, racism, homophobia

ACM Reference Format:
Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in Word Embeddings. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, Barcelona, Spain, 12 pages. https://doi.org/10.1145/3351095.3372843

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02.
https://doi.org/10.1145/3351095.3372843

1 INTRODUCTION

The growing ubiquity of algorithms in society poses questions about their social, political, and ethical consequences [77]. One of the issues research focuses on is algorithmic bias, which denotes the deviation of the algorithmic results from specific social expectations, based on epistemic or normative reasons [75].

Prior research has shown that algorithmic bias might result in unfair or discriminative decisions and statements, initiating a multi-level debate on the ethical use of algorithms [62, 102]. Under that framework, researchers, decision makers and institutions try to answer the following questions:

- What definitions of fairness and discrimination are appropriate and under what conditions? [15, 62]
- At which part of an algorithm does bias emerge and in what form? [42, 85, 93]
- What are the actual consequences of biased algorithms and who is accountable for them? [6, 68, 76]
- How can researchers and decision makers mitigate the detected bias? [8, 13]

Problem Statement

This study investigates bias in word embeddings, a set of natural language processing techniques for the mapping of words into numerical vectors. These vectors can then be used for the improvement of the predictions and inferences of other machine learning models [91]. Previous work has proven that word embeddings contain bias [13], and researchers have already developed methodologies for tracing, quantifying, and mitigating it [12, 16]. Recently, researchers have also started to develop methods for comparing biases existing in different datasets [40, 64].

Despite recent scientific findings, computer scientists in the industry widely use word embeddings for the development of highly accurate models that perform text generation, translation, classification and regression, without taking into consideration the impact of their inherent biases. Similarly, researchers have not yet investigated the diffusion and impact of biased word embeddings on further machine learning algorithms. Therefore, we want to provide a complete overview of bias in word embeddings: its detection in the embeddings, its diffusion in algorithms using the embeddings, and its mitigation at the embeddings level and at the level of the algorithm that uses them. We also investigate whether the employment of biased word embeddings contributes to the location of the bias in new data. The study raises additional awareness about a technique, whose implementation can lead to unfair algorithmic

decisions and inferences. We achieve this, by seeking the answer to the following research questions:

RQ1: How can we evaluate and mitigate the word embeddings' bias diffusion in further machine learning algorithms?

RQ2: Can we employ bias in word embeddings for tracing bias in new data?

Original Contribution

- We train state-of-the-art word embeddings based on the German version of Wikipedia and on unique social media data in the German language. For that, we gather over 22 million tweets and Facebook comments related to German politics. We develop a new method for locating biases in gendered languages, trace niches of sexist, xenophobic and homophobic prejudice and stereotypes on the two sets of vectors, and quantify the overall bias for each dataset.
- We transform and compare the vector spaces without distorting the immanent bias by borrowing techniques from embeddings translation [87]. We then compare the spaces and prove that the social media data containd a higher level of intergroup prejudice, while Wikipedia data contain a stronger bias in terms of stereotypes.
- We create a sentiment classifier based on the two embedding datasets and show how the model replicates bias immanent in the embeddings.
- We compare methodologies to mitigate bias without distorting the accuracy of the classifier. We compare debiasing at the embeddings level and at the level of the classifier. We illustrate that the standard technique for mitigating bias at the embeddings level [13] is insufficient for removing biases completely.
- We develop a new sexism dataset by labeling 100.000 Facebook comments as sexist or neutral and illustrate that embeddings with bias similar to the one in the target data perform better on the classification task.
- Finally, we discuss the issues, possibilities and challenges that accompany the use of biased word embeddings.

Paper Organization

The paper is organized as follows. Section 2 presents the theoretic background and related work. Section 3 describes the data and methodology we followed. Section 4 presents the results. Section 5 discusses the results, demonstrates the implications of the study and concludes the analysis.

2 BACKGROUND AND RELATED WORK

2.1 Algorithms, Bias & Fairness

A prerequisite for understanding bias in word embeddings is to evaluate how the method's results deviate from given social expectations. To do so, we adopt the Friedman et al. [38] proposed framework for analyzing algorithmic bias. They state that algorithms might face three types of bias: 1. preexisting, 2. technical, and 3. emergent.

Preexisting bias is related to the input data. Social or personal attitudes integrated in the input dataset might lead to the deviation of algorithmic inferences from a hypothesized social objective. For example, white hosts on online lodging marketplaces charge more than their non-white counterpart hosts [33]. Algorithms might replicate the asymmetry, valuing an apartment as more expensive only because the owner is white.

Technical bias emerges when there are software, hardware or mathematical constraints. An overfitted algorithm is biased, because its inferences are perfect on the training data but non-generalizable for new cases [39]. Different mathematical models trained on the same data have different prediction accuracies and consequently different amounts of bias, because of the different cost function that they optimize [69]. A computer with RAM limitations will not allow the development of a model on the full dataset, leading to the creation of model predictions that might miss important information.

Emergent bias appears at the evaluation of results and the context of their application. Forming a decision based on algorithmic results might pose an ethical problem, when the decision or inference proposed contradicts existing normative values in the society. Research and policy makers are investigating and trying to define when the emergent bias is transformed to unfair or discriminative decisions. Among others, Goodman [44] refers to algorithms as unfair, when a specific group or individual receive unfavorable treatment as a result of algorithmic decision-making. Cowgill and Tucker [23] argue that algorithmic results should always be compared to a counterfactual ideal case, in relation to which it will be decided when and how an algorithm discriminates. Overall, no unique definition of fairness is available, making each algorithmic application a distinct case to be studied.

Word embeddings might face all three types of bias. It is proven that social attitudes such as sexism and ethnic stereotypes in the initial dataset are transferred to the embeddings [13, 40], denoting the presence of a preexisting bias. Technical bias also appears. Word embeddings trained by different models yield different results on benchmark tests [19, 70, 74, 79]. Word embeddings might also result in emergent biases. Generalizations on social relations based on the distance of words immanent in an embedding space, or by inserting the embeddings in another model for prediction or inference might result in the formation of decisions that deviate from given social imperatives.

Word embeddings are used widely in commercial systems, inter alia for ad generation [46, 47], music and hospitality recommender systems [10, 45], and by tech companies who use them to develop models and offer tools [35, 58]. they constitute decisions that influence multiple social groups and individuals. It is important to understand the appearance of bias in them, the related dangers and possible reactions to them. This will not only contribute towards fairness, but will provide the foundations for creating applications that respect the rights of social groups and individuals [43, 51]. Given that the prominent bias form in word embeddings is related to the input dataset, we investigate preexisting biases and their connection to emergent biases in related applications.

2.2 Text & Social Discrimination

The reason why preexisting biases are imprinted in word embeddings is related to the nature of text. Because text is a medium for

communicating and projecting human interactions, it carries features that constitute the social world. In human history, text has not only been used to organize and comprehend sociopolitical events, but also to shape the way these events are perceived and interpreted [60]. Therefore, power relationships, social discrimination, and social asymmetries are always imprinted in text.

In this study, we narrow the investigation to bias related to social discrimination. We investigate how existing forms of social discrimination in text are diffused and influence word embeddings and further models. Social discrimination refers to discrimination emerging from members of one social group towards members of another [81], thus forming a self-other duality. By the time the distinction of people into groups takes place, group members automatically start to assign different properties to in-group members and other properties to members of the 'competing' group [89]. Social theory states that attitudes of dominant social groups are imprinted in the use of language [14]. Consequently, the bases of social discrimination are diffused through statements of prejudice and stereotypes in text, directly and indirectly [88]. Social discrimination can be not only hostile, but also benevolent. Depending social conditions and group relations, stereotypes and prejudice might be both positive or negative in nature. Regardless of their polarity, they are always a result of group antagonism [41, 54, 97].

The understanding of how social discrimination is projected into text is not a trivial task. For a thorough understanding of the process, text must be analyzed and so should the conditions of its production, its context, and use [3, 36]. To achieve that, researchers have developed extensive qualitative frameworks that take into consideration the sociopolitical conditions that lead to the emergence and formulation of lingual symbols [20, 37]. By taking into consideration political and social structures, ethical values, biases, predispositions, and social group perceptions [21, 92], relations and intentions of speakers and receivers in a social situation, researchers have studied language to understand sexism, racism, and other forms of social discrimination [52, 83]. Because of the complexity and types of social discrimination, the detection and quantification of social discrimination is not always possible by the use of formal mathematical techniques. Therefore, for the analysis of bias in word embeddings and further models, we restrict our study to the detection of forms of social discrimination that are detectable by concept comparison tests (e.g. the adjective check list [98], Implicit association test [49], Bem Sex-Role Inventory [56], polarity tests [31, 84]). These methods locate regularities such as stereotypes or prejudices, rather than explaining why they emerged. An explanation would require additional qualitative analysis, which is beyond the scope of this paper.

2.3 Social Discrimination & Word Embeddings

Researchers have proven that word embeddings contain forms of prejudice and stereotypes related to sexism and racism [13, 40]. Based on that, we study how and when biases result in further socially discriminative algorithmic behavior. To do that, we develop methodologies for tracing biases in gendered languages. Existing methods [13, 17] for detecting biases in word embeddings are grounded in qualitative techniques of concept associations, on which we also rely [49, 84, 98]. They analyze qualities of groups and their relation to other concepts, assuming that in an ideal society these concepts would have been either equally assigned to these groups or not at all. For example, ideally an occupation should not be connected more to one sex than the other, nor should one sex be treated more positively or negatively than the other. These assumptions might hide the actual reasons and conditions for the emergence of the specific associations and their straightforward connection to social discrimination, but are the same assumptions used in standard models for measuring social discrimination based on qualitative techniques [17, 49, 84, 98]. Because existing methods are developed primarily for the English language, we develop a new model that can account for gendered versions of words.

Until now, researchers have investigated various dimensions of bias in word embeddings. Garg et al. [40] show how prejudice evolving over time is imprinted in word embeddings. Kozlowski et al. [64] use the positional change of word embeddings to describe semantic transformation. Dev et al. [27] show that names in word embeddings function as proxies of bias against social groups. Arora et al. [4] prove that different meanings of words are 'encoded' in word embeddings and can be retrieved. Zhao et al. [101] propose a methodology to train word embeddings without sexist bias in them. Brunet et al. [16] develop a technique to trace the origin of bias in embeddings back to the original text. Caliscan et al. [17] introduce a general methodology to trace bias in word embeddings, while Drozd [30] et al. automatize the process. Drawing from previous research, we want to provide a complete picture of bias in the embeddings, its diffusion and mitigation.

2.4 Bias Prediction

Another objective of the study is to test whether bias in word embeddings can be used constructively. To that end, we also investigate whether biased word embeddings can contribute toward detecting bias in new text. Research shows that bias detection, especially in cases of social discrimination such as sexism or racism is very complicated. Park et al. [78] have attempted to create classifiers that detect abusive language. Dahou et al. [24] developed models for sentiment analysis. Kathrik et al. [28] tried to automatically trace cyberbulling in Youtube, while Levy [67] tried to detect sexism in newspaper covers. Overall, the performed attempts yield moderate results, especially when using only text as classification inputs because of the complex nature of human language [61]. It is a challenge to test if bias in word embeddings would lead to the improvement of classifiers predicting social discrimination.

3 DATA AND METHODS

3.1 Word Embeddings

To be able to investigate bias in word embeddings trained on different datasets, we collected data from Facebook, Twitter and Wikipedia. For Facebook and Twitter, we used the application programming interfaces (APIs) of each platform. From the social media channels, we collected data for the six main political parties in Germany: CDU, Germany's main conservative party; CSU, the sister party of the CDU in Bavaria; Bundnis90/Die Grünen, the green party in Germany; FDP, a neo-liberal party; SPD, Germany's social-democratic party; Die Linke, the radical left party and Alternative für Deutschland, the extreme right populist party. For Facebook, we

retrieved the posts from the political parties and their comments during the period between January 2015 and May 2018. For Twitter, we collected the tweets from political parties' Twitter accounts between January and October 2018. We also included tweets from users that mentioned or retweeted the political parties, as well as tweets that included the names of the political parties for the same period. Overall, we gathered 24 million posts, comments, and tweets, which comprised our social media political dataset (485 mil. tokens). For Wikipedia, we collected the complete German wikipedia as bulk file from the official repository. [1] The Wikipedia dataset consisted of 2.2 million articles (850 mil. tokens). All texts were originally written in German. Therefore, related biases existed in the original text and were not added to it by further textual processing (e.g. translation from other languages to German). We present our results in English for readability purposes.

For training the embeddings on the two datasets, we used GloVe, developed by Pennington et al. [79]. The model creates vectors of words by taking into consideration the word co-occurrence frequencies in the dataset and optimizing

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j + logX_{ij})^2,$$

where V is the vocabulary, i and j two words, w_i the word vector of word i, $\tilde{w}_j$ the context vector of word j, b_i and $\tilde{b}_j$ their biases, and X_{ij} the co-occurrence number of the words for a given window. $f(x) = (x/x_{max})^a$ if x lower than a chosen x_{max}, otherwise $f(x) = 1$, with a a hyper-parameter. Following the authors' recommendations, we tokenized the texts using the nltk tokenizer [71], our window size was 10, $a = 3/4$ and $x_{max} = 100$. Overall, the datasets for Wikipedia and social media contained 390.000 and 200.000 word vectors respectively.

3.2 Vector Space Transformation

Optimizing the GloVe cost function results in the nonlinear map

$$N : C, V \mapsto W$$

where C is the corpus, V the vocabulary and W the word embeddings vector space. Given that the corpora for Wikipedia and social media C_w, C_{sm} vary, as well as the two vocabularies V_w and V_{sm}, the generated vector spaces W_w and W_{sm} are not comparable to each other. A comparison presupposes the projection of the one space on the other, given a Transformation matrix T that preserves the bias in the vector spaces. Both Smith et al. [87] in embeddings translation and Hamilton et al. [53] in measuring semantic change obtain the transformation matrix by solving the Orthogal Procrustes problem

$$L = \underset{\Omega}{\operatorname{argmin}} \|\Omega A - B\|_F \text{ subject to } \Omega^T \Omega = I,$$

where A and B, two word embeddings vector spaces and Ω the transformation matrix. The problem is solvable by applying a singular value decomposition algorithm as proposed by Schönman [86]. The specific transformation places all words from vector space A as close as possible to their corresponding words in vector space B. As transformation is linear, the normalized distance between words does not change, thereby preserving bias in the embeddings.

3.3 Bias Detection

For detecting bias in word embeddings we must develop a generally applicable formula. The method proposed by Bolukbasi et al. [13] defines an inter-group direction $\vec{g}$. Then it quantifies the bias of a random word by the cosine distance between the word vector and $\vec{g}$. For example, the vector of the word *nurse* should be independent of the inter-group direction between man and woman. Usually it is not, since society stereotypically sees nursing as a female profession. Nevertheless, this definition does not cover cases that words ought to have an inter-group component. For example, words in German are sex-dependent. There is a male and a female version, denoting that mathematically the vectors of the words and of the sex direction should be dependent. To overcome that, we develop an alternate methodology. First, we define pairs of theory specific words for each type of discrimination. [2] Then we introduce a list of concepts for which we want to measure the bias. If concepts change based on the social groups, e.g. they have male and female versions, they are represented by word-pairs. We calculate the general bias in the embeddings by the equation

$$B_g = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} |cos(w_{j1}, P_{n1}) - cos(w_{j2}, P_{n2})|,$$

where N is the number of concepts, K the number of theory specific pairs, w_{j1} and w_{j2} the embeddings for the jth pair of theory specific words and P_{n1} and P_{n2} the embeddings for nth concept pair in the list. When $P_{n1} = P_{n2}$, i.e. when we investigate concepts that are not variable with respect to the social groups under investigation, we use the equation

$$B_g = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} |cos(w_{j1}, P_n) - cos(w_{j2}, P_n)|$$

The general bias equation compares the magnitude of dependence between a concept and the two groups. If the concept vector has a higher cosine distance to one group vector than to the other, then the concept is biased in that direction. We apply the above equation to two tasks. We create a list containing 1600 professions for a profession-related bias task. We also use the sentiment list developed by Remus et al.[80] that contains words of positive and negative polarity for a sentiment bias task.

3.4 Bias Diffusion

In order to measure bias diffusion, we need to capture whether a model that takes biased word embeddings as input will also give biased output. The bias of the model needs to be theoretically comparable to the bias in the embeddings. Therefore, we used the sentiment dictionary of Remus et al. [80], which contains a set of positive-laden and negative-laden words. We trained a linear support vector machine classifier. The classifier took as input the

[1] https://dumps.wikimedia.org/dewiki/

[2] Man-Woman, German-Foreigners, Straight-Gay, for sexist, xenophobic and homophobic prejudice respectively. Both gender and sexuality are spectra. We did not analyze here biases related to the rest social groups for simplicity reasons, as the methodology deals with group dualities.

embedding vector of a word and predicted if it had a positive or negative sentiment. We modified the output of the classifier by transforming the class probabilities to a sentiment score by applying the equation

$$S_w = -[log_{10}((P(w = positive) - log_{10}(P(w = negative))],$$

where S_w is the sentiment score for word w, P(w=positive) and P(w=negative) the model's assigned probability that the word is positive and negative respectively. Then, we designed an experimental setting by which we could measure the level of sexist and xenophobic prejudice that the classifier learned. We used the first names list developed by Winkelmann [100], and acquired male and female stereotypical first names for nine population groups: German, Turkish, Polish, Italian, Greek, French, US American, Russian and Arabic. We then fed the embeddings of the words into the algorithm and measured the sentiment score across different sexes and populations. We claim that ideally a name should have a sentiment score equal to zero, because it should be polarity-independent. We then defined the sentiment bias $B_{s,c}$ of the algorithm being equal to the classifiers' sentiment score and the classifiers' social discrimination bias for a specific social discrimination concept as

$$B_c = \left| \frac{1}{N} \sum_{i=1}^{N} B_{s,ci1} - \frac{1}{K} \sum_{j=1}^{K} B_{s,cj2} \right|,$$

where N and K are the number of names for each of the two investigated social groups and $B_{s,ci1}$ and $B_{s,cj2}$ the sentiment bias of the classifier for a word in each group respectively. This metric quantifies the difference in the assigned sentiment of the classifier for the names of each group. For investigating the statistical significance of our results we apply Mann-Whitney U and Kruskal-Wallis H tests to compare biases among two or more groups.

3.5 Bias Mitigation

A sentiment analysis algorithm has no social discrimination bias when it predicts equal sentiment for names of different sexes or populations. In order to achieve that we try two approaches. In the first case, we adopt and extend the methodology proposed by Bolukbasi et al. [13]. As the classifier assigns a sentiment polarity value for each input word, we define a sentiment direction $\vec{s} \in R^d$, where d is the dimension of a word vector $\vec{w}$. The direction is calculated by forming pairs of theory specific dualities (e.g. good - bad, positive - negative, etc., see Table 1) that are theory specific and taking the difference of their word vectors. Afterwards, we apply PCA, with the resulting first component being the sentiment direction $\vec{s}$. We also define the set $N = \{\vec{w_1}, ..., \vec{w_n}\}$ corresponding to the vectors of theory neutral words. Then we hard neutralize these words by applying

$$\vec{w_i'} = \vec{w_i} - \frac{\vec{w_i} \cdot \vec{s}}{\vec{s} \cdot \vec{s}} \vec{s},$$

where $\vec{w_i'}$ is the debiased non-normalized vector for word w_i. By doing this, we make the vectors of theory neutral words orthogonal to the sentiment vector. We then feed the neutralized embeddings into the classifier and calculate the sentiment for the different groups. This methodology tries to mitigate bias at the word embeddings level. As non-neutral words are not debiased, the accuracy of the classifier does not change.

Table 1: Word pairs used for the calculation of the sentiment direction translated from German.

Positive	Negative
good	bad
positive	negative
happy	sad
peace	war
cheap	expensive
love	hate

In the second case, we try to mitigate the bias at the level of the classifier. The linear SVM classifier learns to split the classes given a linear hyperplane, which is defined by a normal vector $\vec{p}$. This vector actually corresponds to the sentiment direction as learned by the classifier. Therefore, we hard-neutralize the theory neutral vectors given vector $\vec{p}$ by applying the same formula as above.

3.6 Bias Prediction

The last part of our study focuses on understanding whether biased word embeddings can help detecting bias in new text. For this scope, we manually labeled 100,000 user comments from German political parties Facebook pages and created a sexism dataset. We categorized each comment as sexist or neutral based on the following criteria: 1. the existence of a sexist buzzword, 2. the formulation of sex-related compliments, 3. the expression of statements against the equality of sexes, and 4. the assignment of stereotypical roles to persons based on their sex. Each of the four categories denoted a different label in the dataset and its formation was based on previous theoretic work. We traced sexist buzzwords under the notions of traditional sexism [32, 73], while we defined and searched sex related compliments given theories of benevolent sexism [9, 59]. We located statements against sex equality in comments that the users explicitly argued about the topic and we defined stereotypical roles of the sexes based on the works of Eckes [32], Tilegea [90], and Benokraitis et al. [9].

To efficiently code the dataset, we created sound recordings of each comment, because it has been shown that hearing a sentence rather than just reading it improves content understanding [34]. Two coders reviewed the sound corpus, assigning to each comment one or more of the four labels and giving a concrete reason for their decision. In cases of coders' disagreement, the comments were reviewed by one additional coder. For these comments we accepted labels assigned by more than one reviewer. Comments that were not assigned a label at all were then classified as non-sexist, while comments having at least one of the four labels as sexist. Overall, we detected 1,988 sexist comments. We then sampled an equal number of neutral comments, creating a balanced dataset, which we split into a train and a test set. We evaluated the biased word embeddings on the classification test. We created models that included long-short-memory network (LSTM) and attention based architectures, and investigated their accuracy on the test set, with 1. random word embeddings, 2. the embeddings from the Wikipedia data, 3. the embeddings from the social media data and 4. embeddings

trained on the sexism dataset. Furthermore, we investigate which properties of the word embeddings are responsible for accuracy improvement. For that, we transformed and compared the word embeddings from the sexism dataset to the other embeddings by calculating their mean weighted cosine similarities, as given by the equation

$$Sim_{s,i} = \frac{\sum_{n=1}^{N} f_n cos(\vec{w_{n,s}}, \vec{w_{n,i}})}{\sum_{n=1}^{N} f_n}, \text{ with } n = 1, ..., N \in s \cap i,$$

where s is the word embeddings trained on the sexism dataset, i is another word embeddings dataset, N is the number of common words in the two datasets and f_n is the frequency of appearance of a common word n in the sexism dataset. We also perform the sentiment task with the sexism dataset embeddings and calculate the level and type of sexist prejudice within them.

4 RESULTS

The results are split into three parts. First, we present our findings on bias within the Wikipedia and social media word embeddings. Second, we analyze how the bias was diffused and how we mitigated it. We also illustrate the efficiency of biased word embeddings when used as sexism detection models. In the last part of the section, we evaluate bias in word embeddings.

4.1 Bias in Word Embeddings

The word embeddings generated on the Wikipedia and social media corpora contained 390,000 and 200,000 vectors respectively. In both cases, the profession and sentiment task revealed intensive stereotypical features assigned to each examined social group. In both Wikipedia and social media spaces, women were mostly associated with professions like nurses and secretaries. On the other hand, men were associated with stereotypical male roles, like policemen and commanders. The aforementioned assigned professions highly correlate with the actual profession distribution in society [1], denoting that the actual social asymmetry is imprinted in the vectors. For Wikipedia, women were strongly associated with concepts related to marriage, while men were linked to concepts related to war and power. This could be because Wikipedia extensively includes biographies of historical figures, in which women are typically associated with marriage and familial relations, while men are associated with concepts such as war and governance [95, 96]. In social media, the female sex was closer to positive feelings such as love and maturity, but also to negative ones like stubbornness and agitation. Men were closer to concepts related to aggression and fighting, with most of them being negative. The stereotypes found in the social media dataset comply with previous research findings [99], which found the existence of power related stereotypes for men and sentiment related stereotypes for women.

In both Wikipedia and social media, Germans were intensively associated with jobs related to governance and journalism, while foreigners either to blue collar jobs or to professionals dealing with foreign populations such as aid officials, politicians or tour guides. Foreigners were generally linked to sentiment concepts related to immigration, law and crime, while Germans to positive feelings such as charm and passion (social media), as well as to cooperation and union (Wikipedia). The association of foreigners to immigration related concepts and professions can be traced back to the refugee crisis taking place in Europe over the last few years, which has a prominent position in the public agenda [65]. Similarly, researchers have proven the existence of biased slants related to immigration issues on wikipedia [48]. Given that both German Wikipedia and the German social media discussions are primarily produced by Germans, we can attribute the inherent positivity and negativity on Germans and foreigners on the intergroup prejudice existing in the society [2, 72]

Table 2: Extreme words for each task and group using the embeddings from Wikipedia data

Wikpedia

Sexist prejudice			
Profession		Sentiment	
Woman	Man	Woman	Man
Nurse	Officer	Wedding	Reinforcement
Secretary	Hunter	Divorce	Attack
Teacher	Commander	Anulment	Combat
Saleswoman	Guard	Engagement	Power
Actress	Cameraman	Marry	Decrease
Population Prejudice			
Profession		Sentiment	
Foreigners	German	Foreigners	German
Aid official	Author	Refugee	Champion
Craftsman	Journalist	Unauthorized	Cooperation
Bank Assistant	Historian	Lawful	Union
Tour guide	Director	Tax	New
Foreman	Painter	Accumulate	Assignment
Sexual Orientation Prejudice			
Profession		Sentiment	
Homosexuality	Heterosexuality	Homosexuality	Heterosexuality
Artist	Singing teacher	Corruption	Unserious
Art dealer	Copywriter	Violence	Nice
Actress	Forest manager	Adultery	Fantastic
Cook	Track driver	Known	Smart
Shoemaker	Carpenter	Prohibited	Fair

The stereotypes were equally intensive for sexual orientation. Homosexuals were related to stereotypical roles such as artists (Wikipedia) and hairdressers (social media), while persons of heterosexual orientation were related to blue collar professions or positions in science. Strikingly, homosexuality was related in both datasets with very negative concepts: from violence, prohibition and adultery (Wikipedia), to death sentencing, abuse and harassment (social media). On the complete opposite side, heterosexuality was closely positioned to inherently positive sentiments such as fantastic and smart (Wikipedia) and to concepts like friendship and deliberation (social media). These findings comply with historic negative social attitudes against homosexuality, where conservative groups state that it is abnormal and that should be prohibited by law [26]. Regarding positive concept relations to homosexuality, researchers have found similar associations in concept association tests [94], illustrating that biases in social media and wikipedia

Table 3: Extreme words for each task and group using the embeddings from social media data

Social Media

Sexist prejudice			
Profession		Sentiment	
Woman	Man	Woman	Man
Nurse	Policeman	Agitation	Robber
Secretary	Musician	Mature	Attacker
Pharmacist	Priest	Love	Injured
Religion teacher	Coach	Increase	Fascist
Correspondent	Paramedic	Stubborness	Overwhelmed
Population Prejudice			
Profession		Sentiment	
Foreigners	German	Foreigners	German
Newspaper	Government Official	Criminal	Mature
Skilled worker	Correspondent	Exclude	Beauty
Politician	Notary	Refugee	Charm
Consultant	Butler	Increase	Passion
Teacher	Reporter	Frustration	Love
Sexual Orientation Prejudice			
Profession		Sentiment	
Homosexuality	Heterosexuality	Homosexuality	Heterosexuality
Artist	Streetworker	Death sentence	Friendly
Scrap dealer	Political scientist	Discrimination	Moving
Hairdresser	Political economist	Abuse	Deliberation
Interviewer	Mediator	Harassment	Increasing
Consultant	Biologist	Violence	Unecessary

correspond to the ones found offline. An overview of the most extreme concept associations for all groups can be found in tables 2 and 3. The results demonstrate strong stereotypical associations for all groups. Overall, the calculated general bias was higher for almost all categories and tasks for the Wikipedia dataset (table 4), denoting that Wikipedia introduces more severe stereotypes for each social group than the examined social media content. The calculated scores are of similar magnitude to those calculated by Bolukbasi et al. [13], who calculated a general bias of 0.08 on the profession task for the two sexes on an English Google news corpus.

Table 4: General bias for each intergroup comparison, bias task and embeddings dataset.

	Wikipedia		Social Media	
	Profession	Sentiment	Profession	Sentiment
Sex	0.080	0.087	0.077	0.037
Population	0.066	0.063	0.054	0.056
Sex orientation	0.064	0.087	0.0619	0.084

The presented associations only reveal partial bias in the embeddings. Indeed, stereotypes are a base of social discrimination, and someone can qualitatively evaluate how specific social groups are presented in the datasets by checking the mostly associated concepts. Nevertheless, this does not per se signify that a specific group is generally favored over another, which would provide evidence of prejudice. To achieve that, we calculated the mean polarity score for the sentiment concepts being closer to each social group, and then extracted the difference for each intergroup comparison. The results are given in Figure 1. For both Wikipedia and social media, Germans were depicted much more positively than foreigners. The

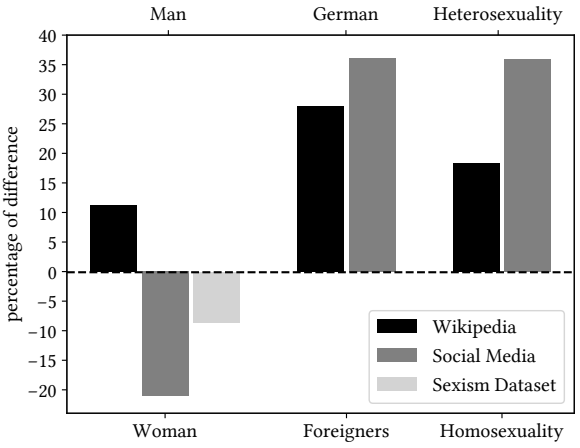

Figure 1: Intergroup positive sentiment difference in the embeddings.

same applies for heterosexuals in comparison to homosexuals. Both results are in accordance to the sentiment task results, as Germans and heterosexuals were associated with much more positive feelings and concepts, confirming the existence of biases that favor privileged social groups [26, 50].

In German Wikipedia, men were generally depicted more positively. On the other hand, in the social media dataset, women were associated with more positive words. One explanation is that in Wikipedia men were described by stereotypical concepts like power, attack and reinforcement, which are labeled as positive in the polarity dictionary. In contrast, the social media data also related men to concepts like fascism and robbery, i.e. words with highly negative sentiment. That could also be rooted in the nature of German language, which uses the male plural when making colloquial general claims. Because negative statements about groups on social media were generated in a male form, this bias could have been replicated by the model. Furthermore, the sentiment difference does not fully replicate bias in text. For example, in social media data, women are often associated with the term 'mother', for which the sentiment lexicon assigns a positive score. Nevertheless, the actual combination of words in a political context corresponds to sexist speech, as numerous users refer to female politicians as mothers in order to undermine their political abilities.

The above results illustrate that word embeddings contain a high level of bias in them in terms of group stereotypes and prejudice. The intergroup comparison between sexes, populations, and sexual orientations revealed the existence of strong stereotypes and unbalanced evaluations of groups. Although Wikipedia contained stronger bias in terms of stereotypes, social media contained a higher bias in terms of group prejudice.

4.2 Bias Diffusion, Mitigation & Prediction

Our analysis shows that the above bias was diffused further into the trained sentiment classifiers. We trained one classifier for each embedding dataset, with both having a test set accuracy of around

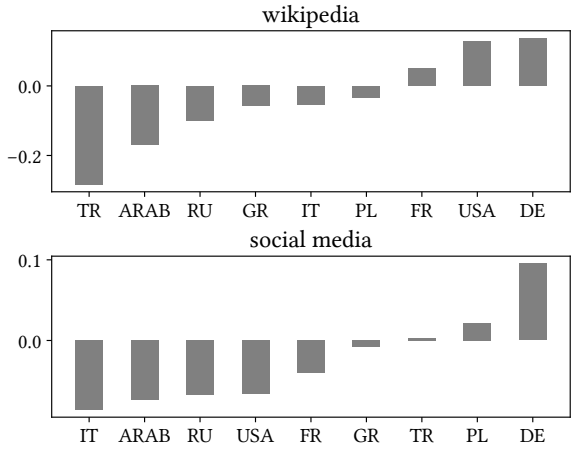

Figure 2: Predicted score of the sentiment classifier for stereotypical names of different populations

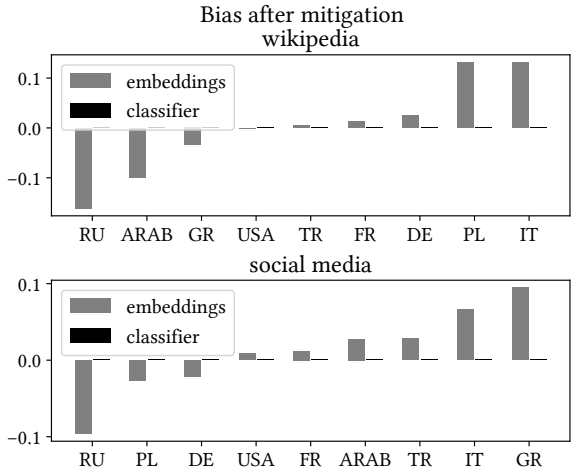

Figure 3: Bias in the sentiment classifier for stereotypical names of various populations after mitigation at (a) the embeddings' level, (b) the level of the classifier.

85%. The classification task for stereotypical names of different communities illustrated a preference for German names (Figure 2). In both embedding datasets, German names were assigned the highest average sentiment score. In contrast, most of the foreign names were assigned negative sentiment values. Arabic and Russian names were negatively associated in both datasets, which can be grounded both on existing social stereotypes against Russians and Arabs in the society [5, 7], as well as mainstream media representations of the ethnicities [55, 66]. Greek, Polish and Turkish names were seen much more positively by the social media classifier. This comes in contrast to what someone would actually expect, since a large part of contemporary German public opinion holds strong negative stereotypes against Greek, Polish and Turkish populations due to economic and migration issues [5, 11, 63]. French and US-American stereotypical names were classified much more positively by the Wikipedia classifier. The result related to French names was not intuitive, given the historical conflicts between Germany and France that are extensively covered in Wikipedia [57]. In contrary, researchers illustrate that non-English Wikipedia pages on U.S.-American persons generally contain positive cues [18], explaining also the favoritism of the classifier for U.S.-American names. Overall, the classifiers' social discrimination biases for the models trained on the Wikipedia and the social media data were $B_{c,wiki} = 0.23$ and $B_{c,sm} = 0.14$ respectively. The bias of the classifier was similar to the bias in the embeddings, as in both cases German concepts were evaluated much more positively. For both classifiers the Kruskal–Wallis tests were significant (sm classifier: H=101.95, p-value: <0.01; wiki classifer: H=37.36, p-value < 0.01), denoting that the mean bias for each ethnicity varies significantly from the others.

We concluded with similar findings when predicting the sentiment of male and female names. The classifiers exactly replicated the prejudice as measured in the word embeddings (Figure 4). The Wikipedia classifier predicted a higher average sentiment score for male names. In contrast, the social media classifier assigned a much more positive overall score to female names. This complies with the results from the intergroup positive sentiment difference in the embeddings, where women were associated with more positive concepts than men in the social media dataset, while the opposite happened in the Wikipedia embeddings. Hence, we proved that classifiers trained in biased word embeddings replicate the bias existing in the vectors. Overall, the classifiers' social discrimination biases were $B_{c,wiki} = 0.011$ and $B_{c,sm} = 0.068$ respectively. The Mann-Whitney U test was significant for the social media classifier (U = 1027471, p-value < 0.01), but not for the Wikipedia classifier (U = 1069947, p-value = 0.23). This does not mean that there is no bias between sexes in the second case. By breaking down names by ethnicity and comparing them, we get significant results for German (U = 91356, p-value = 0.001), Polish (U = 19, p-value = 0.01), Greek (U = 90, p-value = 0.003) and U.S.-American (U = 63128, p-value = 0.02) names.

The study proves that the diffused bias can be mitigated. Both methodologies for bias mitigation reduced bias significantly. Mitigation at the embeddings level resulted in social discrimination biases of the classifiers of $B_{c,wiki} = 0.027$ and $B_{c,sm} = 0.035$ for the population comparison. Similarly, when predicting the sentiment of male and female names, the bias of the classifiers after mitigation was $B_{c,wiki} = 0.009$ and $B_{c,sm} = 0.018$ respectively. Mitigation at the level of the classifier was by far more efficient: In all possible tasks, the overall social discrimination bias vanished. Figure 4 presents an overview of bias before and after mitigation for each case. In order to understand why the second methodology provides better results, we calculated the cosine distance between the sentiment vectors of the embeddings and the classifier, which were used for de-biasing. The value was close to 0.9, denoting that the classifier actually learns a significantly different sentiment direction than the one defined by the methodology proposed by Bolukbasi et al.[13]. Actually, the classifier learns further associations between the vectors, which are not taken into consideration when debiasing at the embeddings level. Debiasing at the embeddings level

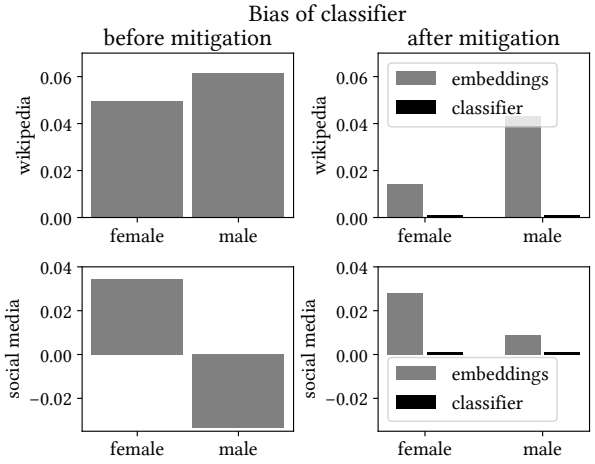

Figure 4: Predicted score of the sentiment classifier for male and female names, before and after mitigation by applying two different methods.

results in the diffusion of a different bias in the classifier. As Figure 3 shows, although bias related to the favored group was highly reduced, remaining patterns in the data resulted in a totally different bias diffusion. This bias was not universally distributed in all cases, but resulted in asymmetries in certain cases. For example, for the classifier trained on the Wikipedia embeddings the mean bias difference between German and Russian (U = 23363, p-value = 0.065), Arabic (U = 90, p-value = 0.045) and Italian (U = 86625, p-value = 0.065) names remained statistically significant. This was not the case for the sex names comparison in either classifier (sm: U=1060104, p-value=0.12; wiki: U=1065656, p-value=0.18) or ethnicity names comparison for the classifier trained on the social media embeddings (Kruscal-Wallis H=4.2, p-value=0.83). Hence, we show that debiasing at the classifier level is a much better and safer methodology to follow. Because of the mathematical definition of the linear support vector classifier, it was straightforward to mitigate the bias in it. For other cases, where non-linearity prevails, more sophisticated methodologies are needed.

Our last finding states that biased word embeddings can be useful for bias prediction tasks. We trained and deployed various models on the sexism prediction task, with and without the trained biased word embeddings. On the first test, we created a simple LSTM model, which had as inputs either a random dataset, Wikipedia, social media, or the sexism dataset word embeddings. We restricted the embeddings from being trainable, in order to evaluate their actual influence on the results. In addition, we only inserted the values of the word embeddings for the words that were common in the datasets. In this way, we could assure that if an embedding dataset had more impact on the results, that it would be because of the type of information encoded into the vectors and not the amount of words existing in the dataset. The models with the trained embeddings provided higher test accuracy and F1 scores. The model with the sexism dataset vectors yielded the best results. The social media embeddings provided better results than the Wikipedia vectors. The calculated weighted mean cosine similarity between the sexism dataset vectors and the social media and the Wikipedia datasets was 0.49 and 0.39 respectively. This denotes that social media vectors are more similar to the vectors of the sexism classifier, which in turn signifies that more similar meanings and, consequently, biases were encoded in them. This is also proven by the sentiment task, for which the sexism dataset vectors had similar prejudice with the social media vectors (Figure 1). Thus, the more similar the bias in the embeddings with the target data, the higher the ability of the classifier to detect the bias.

On the second task, we used additional architectures for the prediction task. We allowed the embeddings to be freely trainable, and used all the available vectors to predict sexism. The best model contained an attention layer and provided an accuracy of 80%. Then, we removed all test observations that contained words that did not appear in the training process, and recalculated the accuracy. We obtained an overall score of 92% on the test data. Given the general difficulty in the detection of sexism and hate-speech by machine learning models [25, 29], the results are more than satisfactory. The model's input was text without any punctuation, nor any other metadata that generally help in detecting social discrimination [82]. Therefore, we showed that biased word embeddings can substantially help in sexism detection, while attention based networks can provide really high accuracy in detecting sexism. An overview of all models can be found in table 5.

Table 5: Classification results for the sexism task

Model	Embeddings	Trainable	Accuracy	F1 - sexist	F1 - neutral
LSTM	Random	False	0.57	0.55	0.62
LSTM	Wiki - common	False	0.68	0.65	0.70
LSTM	SM - common	False	0.70	0.69	0.70
LSTM	Sexism - common	False	**0.75**	**0.75**	**0.75**
Attention	Sexism - all	True	0.80	0.80	0.81
Attention	Sexism - all - filtered	True	**0.92**	**0.92**	**0.91**

4.3 Evaluating biased word embeddings

The analysis provided a thorough description of bias in word embeddings. We proved that the technique replicates biases related to sexism, homophobia, and xenophobia immanent in the original text. We showed that Wikipedia data mediates to the word embeddings stronger stereotypes, while political social media data imprints stronger forms of group favoritism into the vectors.

The study illustrated that the use of biased word embeddings results in the creation of biased machine learning classifiers. Models trained on the embeddings replicate the preexisting bias. Bias diffusion was proven both for sexism and xenophobia, with sentiment classifiers assigning positive sentiments to Germans and negative sentiments to foreigners. In addition, the amount of polarity for men and women in the embeddings was diffused unaltered into the models. We used two methods for bias mitigation, one at the level of the embeddings and one at the level of the classifier. In both cases, we lowered the bias, while mitigation at the level of the classifier was the optimal one.

The analysis also showed that biased word embeddings can be beneficial for bias prediction. Embeddings containing bias similar to the one in the investigated dataset can help in the classification task.

We showed that text-only models for bias prediction can provide more than satisfactory results by using embeddings. Among the various models developed, we found that simple attention-based neural networks yielded the best results. Of course, the developed models are in the position to detect forms of sexism similar to that defined by the inter-subjective coding process and its theoretical assumptions. The models are not generalizable to other forms of sexism that were not taken into consideration at the development of the dataset. Nevertheless, the study provides promising findings for the detection of biases in text by the use of word embeddings and deep neural architectures.

Overall, the study provided a full evaluation of biased word embeddings. It showed how bias can be detected, its diffusion, and how it can be mitigated. It also proved that different forms of bias influence further models differently. In addition, we showed positive aspects of word embeddings. Not only can they be used for bias detection, but most importantly, they can help understand and evaluate sociopolitical relations immanent in text.

5 DISCUSSION

The findings of the study provide a complete picture of the issues, limits, and possibilities of biased word embeddings at the algorithmic level. In the discussion, we go one step further and analyze the societal importance of the aforementioned findings. We illustrate the emerging opportunities for the use of biased word embeddings, while we explain their negative properties. Last but not least, we describe the related challenges that researchers and decision makers need to deal with, in order to assure a just application of algorithmic systems based on word vectors.

On the positive side, the ability of word embeddings to absorb semantic relations of the social world prevails as their main advantage. Being able to quantify bias existing in the society, latent political relations and properties of language and text, has always been a scientific challenge, and until now a privilege of qualitative social science [20]. Word embeddings constitute a way to mathematically grasp and describe sociopolitical relations through the analysis of text, allowing the quantification of phenomena as racism, sexism and social discrimination in general. Based on the vectors, it is possible to evaluate social phenomena, compare and measure their magnitude for different conditions and context. A systematic analysis of word embeddings can result in the creation of new scientific knowledge about the social world, redefining and developing further existing theories. Furthermore, developing models for bias detection by using biased word embeddings can be beneficial. Word embeddings generally improve the accuracy of machine learning models, and we proved that this was also the case in bias prediction, a task which is highly difficult.

On the negative side, the dependence of word embeddings on the nature of the input data is an open methodological issue. There is no such thing as naturally developed neutral text, because the semantic content of words is always bound with the sociopolitical relations of a society [14]. The study illustrates that even text generated in a formal and controlled environment like Wikipedia, results in biased word embeddings. Furthermore, the preexisted bias becomes even more graspable when evaluating the vectors and using them in further algorithms. The algorithms associate stereotypes and concepts to specific social groups, while containing latent prejudice. These associations are usually not directly perceivable in the initial text, nor are they uniformly distributed within it. Nevertheless, the projection of words in a mathematical space by the embeddings consolidates stereotyping and prejudice, assigning static properties to social groups and individuals. Relations are no longer context-dependent and dynamic, and embeddings become deterministic projections of the bias of the social world. This bias is diffused into further algorithms unchanged, resulting in socially discriminative decisions.

Word embeddings are a valuable tool for improving machine learning models and for understanding the social world. Managing their bias prevails as an open challenge for ethical and fair algorithmic applications. Until now, researchers and commercial companies train and integrate word embeddings uncontrollably in their models, without taking into consideration the potential impact and societal implications. The study showed that bias in word embeddings can result in algorithmic social discrimination, yielding negative inferences on specific social groups and individuals. Therefore, it is necessary not only to reflect on the related issues, but also to develop frameworks of action for the just use of word embeddings. To achieve that, it is necessary to develop frameworks that detect bias in concrete algorithmic applications of the embeddings and quantify their impact on individuals and the society [22]. This presupposes commercial companies becoming more transparent regarding the exact algorithms and data they use in their products and decisions. Only through detailed auditing can it be possible to fully understand the issues and start implementing measures that assure algorithmic justice. These measures include the hard mitigation of the bias at the level of the end product, in such a way that no individual is negatively influenced or discriminated against.

It also includes the development of artificial datasets that comply with certain social expectations, on which the embeddings can be trained on. Until now, word embeddings are either trained on text related to a specific algorithmic application or context, or on huge freely accessible corpora. In both cases, bias in the text is always imprinted in the embeddings, and therefore also diffused in further models. It is necessary to search for alternatives, in order to remove preexisting bias in an optimal way.

Our study provides a complete overview on the issue of bias in word embeddings. Not only does it describe the problems and possible solutions, but also initiates an important discussion on the implementation of the vectors in commercial applications. The presented results denote the need for more transparency in the use of word embeddings, in order to ensure their ethical algorithmic implementation. The mathematical tools for model evaluations are already provided; actions from the related stakeholders need to follow.

REFERENCES

[1] [n.d.]. Which jobs do men and women do? Occupational breakdown by gender. https://careersmart.org.uk/occupations/equality/which-jobs-do-men-and-women-do-occupational-breakdown-gender

[2] Richard Alba, Peter Schmidt, and Martina Wasmer. 2004. *Germans or foreigners? Attitudes toward ethnic minorities in post-reunification Germany.* Springer.

[3] Michael W Apple. 1992. The text and cultural politics. *Educational Researcher* 21, 7 (1992), 4–19.

[4] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association of Computational Linguistics* 6 (2018), 483–495.

[5] Frank Asbrock. 2010. Stereotypes of social groups in Germany in terms of warmth and competence. *Social Psychology* (2010).

[6] Solon Barocas, Sophie Hood, and Malte Ziewitz. 2013. Governing algorithms: A provocation piece. *Available at SSRN 2245322* (2013).

[7] Rupprecht S Baur and Stefan Ossenberg. 2017. Zur Verbindung von Stereotypen und Komik am Beispiel deutsch-russischer Witze. In *(Un) Komische Wirklichkeiten*. Springer, 329–342.

[8] Yahav Bechavod and Katrina Ligett. 2017. Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044* (2017), 1733–1782.

[9] Nijole Vaicaitis Benokraitis and Joe R Feagin. 1995. *Modern sexism: Blatant, subtle, and covert discrimination*. Pearson College Div.

[10] Erik Bernhardsson. 2013. Model benchmarks. https://erikbern.com/2013/11/02/model-benchmarks.html

[11] Hans Bickes, Tina Otten, and Laura Chelsea Weymann. 2014. The financial crisis in the German and English press: Metaphorical structures in the media coverage on Greece, Spain and Italy. *Discourse & Society* 25, 4 (2014), 424–445.

[12] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121* (2016).

[13] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[14] Pierre Bourdieu. 1991. *Language and symbolic power*. Harvard University Press.

[15] Danah Boyd, Karen Levy, and Alice Marwick. 2014. The networked nature of algorithmic discrimination. *Data and Discrimination: Collected Essays. Open Technology Institute* (2014).

[16] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2018. Understanding the Origins of Bias in Word Embeddings. *arXiv preprint arXiv:1810.03611* (2018).

[17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[18] Ewa S Callahan and Susan C Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology* 62, 10 (2011), 1899–1915.

[19] Yanqing Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2013. The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226* (2013).

[20] Paul Chilton. 2004. *Analysing political discourse: Theory and practice*. Routledge.

[21] Paul Chilton and Christina Schäffner. 2002. *Politics as text and talk: Analytic approaches to political discourse*. Vol. 4. John Benjamins Publishing.

[22] Sasha Costanza-Chock. 2018. Design justice: Towards an intersectional feminist framework for design theory and practice. *Available at SSRN 3189696* (2018).

[23] Bo Cowgill and Catherine Tucker. 2017. *Algorithmic Bias: A Counterfactual Perspective*. Technical Report. Working Paper: NSF Trustworthy Algorithms.

[24] Abdelghani Dahou, Shengwu Xiong, Junwei Zhou, Mohamed Houcine Haddoud, and Pengfei Duan. 2016. Word embeddings and convolutional neural network for arabic sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2418–2427.

[25] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

[26] Connie De Boer. 1978. The polls: Attitudes toward homosexuality. *The Public Opinion Quarterly* 42, 2 (1978), 265–276.

[27] Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word Vectors. *arXiv preprint arXiv:1901.07656* (2019).

[28] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web* 11, 02 (2011), 11–17.

[29] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. ACM, 29–30.

[30] Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 3519–3530.

[31] Alice H Eagly and Antonio Mladinic. 1989. Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin* 15, 4 (1989), 543–558.

[32] Thomas Eckes. 2008. Geschlechterstereotype: Von Rollen, Identitäten und Vorurteilen. In *Handbuch Frauen-und Geschlechterforschung*. Springer, 171–182.

[33] Benjamin Edelman, Micahel Luca, et al. 2014. *Digital Discrimination: The Case of Airbnb. com*. Technical Report. Harvard Business School.

[34] K Anders Ericsson and Herbert A Simon. 1984. *Protocol analysis: Verbal reports as data*. the MIT Press.

[35] Facebook. 2018. Research in Brief: Dynamic Meta-Embeddings improve AI language understanding. https://code.fb.com/ai-research/dynamic-meta-embeddings/

[36] Norman Fairclough. 1992. *Discourse and social change*. Vol. 10. Polity press Cambridge.

[37] Michel Foucault. 2013. *Archaeology of knowledge*. Routledge.

[38] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[39] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, NY, USA:.

[40] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[41] Peter Glick and Susan T Fiske. 2018. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. In *Social Cognition*. Routledge, 116–160.

[42] Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813* (2016).

[43] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57.

[44] Bryce W Goodman. 2016. Economic Models of (Algorithmic) Discrimination. In *29th Conference on Neural Information Processing Systems*, Vol. 6.

[45] Mihajlo Grbovic. 2018. Listing Embeddings in Search Ranking. https://medium.com/airbnb-engineering/listing-embeddings-for-similar-listing-recommendations-and-real-time-personalization-in-search-601172f7603e

[46] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, Ricardo Baeza-Yates, Andrew Feng, Erik Ordentlich, Lee Yang, and Gavin Owens. 2016. Scalable semantic matching of queries to ads in sponsored search advertising. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 375–384.

[47] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. 2015. E-commerce in your inbox: Product recommendations at scale. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1809–1818.

[48] Shane Greenstein and Feng Zhu. 2012. Is Wikipedia Biased? *American Economic Review* 102, 3 (2012), 343–48.

[49] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74, 6 (1998), 1464.

[50] Louk Hagendoorn. 1995. Intergroup biases in multiple group systems: The perception of ethnic hierarchies. *European review of social psychology* 6, 1 (1995), 199–228.

[51] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

[52] Kira Hall and Mary Bucholtz. 2012. *Gender articulated: Language and the socially constructed self*. Routledge.

[53] William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096* (2016).

[54] Deborah Hellman. 2008. *When is discrimination wrong?* Harvard University Press.

[55] Seth M Holmes and Heide Castañeda. 2016. Representing the "European refugee crisis" in Germany and beyond: Deservingness and difference, life and death. *American Ethnologist* 43, 1 (2016), 12–24.

[56] Cheryl L Holt and Jon B Ellis. 1998. Assessing the current validity of the Bem Sex-Role Inventory. *Sex roles* 39, 11-12 (1998), 929–941.

[57] Michael Howard. 2013. *The Franco-Prussian War: The German Invasion of France 1870–1871*. Routledge.

[58] IBM. 2019. Word Embedding Generator. https://developer.ibm.com/exchanges/models/all/max-word-embedding-generator/

[59] Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*. 7–16.

[60] John E Joseph. 2006. *Language and politics*. Edinburgh University Press.

[61] Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are Word Embedding-based Features Useful for Sarcasm Detection? *arXiv preprint arXiv:1610.00883* (2016).

[62] Keith Kirkpatrick. 2016. Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Commun. ACM* 59, 10 (2016), 16–17.

[63] Andreas Klink and Ulrich Wagner. 1999. Discrimination Against Ethnic Minorities in Germany: Going Back to the Field 1. *Journal of Applied Social Psychology* 29, 2 (1999), 402–423.

[64] Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The Geometry of Culture: Analyzing Meaning through Word Embeddings. *arXiv preprint arXiv:1803.09288* (2018).

[65] Michał Krzyżanowski, Anna Triandafyllidou, and Ruth Wodak. 2018. The mediatization and the politicization of the "refugee crisis" in Europe.

[66] Walter Laqueur. 2018. *Russia and Germany: Century of Conflict.* Routledge.

[67] Susan Leavy. 2014. *Detecting Gender Bias in the Coverage of Politicians in Irish Newspapers Using Automated Text Classification.* Ph.D. Dissertation. Trinity College Dublin.

[68] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.

[69] Tjen-Sien Lim, Wei-Yin Loh, and Yu-Shan Shih. 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine learning* 40, 3 (2000), 203–228.

[70] Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical Word Embeddings.. In *AAAI*. 2418–2424.

[71] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).

[72] Bart Maddens, Jaak Billiet, and Roeland Beerten. 2000. National identity and the attitude towards foreigners in multi-national states: the case of Belgium. *Journal of ethnic and migration studies* 26, 1 (2000), 45–60.

[73] Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication.*

[74] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems.* 3111–3119.

[75] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 2053951716679679.

[76] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How search engines reinforce racism.* NYU Press.

[77] SC Olhede and PJ Wolfe. 2018. The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018), 20170364.

[78] Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206* (2017).

[79] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 1532–1543.

[80] Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS-A Publicly Available German-language Resource for Sentiment Analysis.. In *LREC*.

[81] Katherine J Reynolds, John C Turner, and S Alexander Haslam. 2000. When are we better than them and they worse than us? A closer look at social discrimination in positive and negative domains. *Journal of personality and social psychology* 78, 1 (2000), 64.

[82] Abigail Riemer, Stephenie Chaudoir, and Valerie Earnshaw. 2014. What looks like sexism and why? The effect of comment type and perpetrator type on women's perceptions of sexism. *The Journal of general psychology* 141, 3 (2014), 263–279.

[83] Celia Roberts, Evelyn Davies, and Tom Jupp. 2014. *Language and discrimination.* Routledge.

[84] Paul Rosenkrantz, Susan Vogel, Helen Bee, Inge Broverman, and Donald M Broverman. 1968. Sex-role stereotypes and self-concepts in college students. *Journal of consulting and clinical psychology* 32, 3 (1968), 287.

[85] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014), 1–23.

[86] Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31, 1 (1966), 1–10.

[87] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859* (2017).

[88] Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social communication* (2007), 163–187.

[89] Henri Tajfel. 1970. Experiments in intergroup discrimination. *Scientific American* 223, 5 (1970), 96–103.

[90] Cristian Tileaga. 2014. Prejudice as collective defi nition: ideology, discourse and moral exclusion. In *Rhetoric, Ideology and Social Psychology.* Routledge, 85–96.

[91] Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics.* Association for Computational Linguistics, 384–394.

[92] Teun A Van Dijk. 2002. Political discourse and political cognition. *Politics as text and talk: Analytic approaches to political discourse* 203 (2002), 203–237.

[93] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare).* IEEE, 1–7.

[94] Denise C Viss and Shawn M Burn. 1992. Divergent perceptions of lesbians: A comparison of lesbian self-perceptions and heterosexual perceptions. *The Journal of social psychology* 132, 2 (1992), 169–177.

[95] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media.*

[96] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science* 5, 1 (2016), 5.

[97] Bernard E Whitley Jr and Mary E Kite. 2016. *Psychology of prejudice and discrimination.* Routledge.

[98] John E Williams and Susan M Bennett. 1975. The definition of sex stereotypes via the adjective check list. *Sex roles* 1, 4 (1975), 327–337.

[99] John E Williams, Robert C Satterwhite, and Deborah L Best. 1999. Pancultural gender stereotypes revisited: The five factor model. *Sex roles* 40, 7-8 (1999), 513–525.

[100] Matthias Winkelmann. 2016. firstname database. https://doi.org/10.5281/zenodo.15991

[101] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496* (2018).

[102] James Zou and Londa Schiebinger. 2018. Design AI so that it's fair. *Nature* 559, 7714 (2018), 324–326.

6 Discussion

In this chapter, I summarize my findings on algorithmic influence on political machines. I describe the thesis' contributions in understanding political dimensions of algorithmic implementations 1. for data-driven microtargeting, 2. political content recommendations on social media, and 3. text-based algorithms for ADM systems. I describe implications and open challenges for social media and ADM systems as political machines. Then, I show possible avenues for future work that can additionally shed light on political machines. Finally, I discuss on how algorithmic implementations should be integrated in technological ecosystems, in order to ensure fair, inclusive, and diverse social machines.

6.1 Summary

The aim of the thesis was to uncover further political dimensions of algorithmic influence on social machines, which have not been studied extensively by prior researchers in computational social science. To achieve that, I investigated three specific case-studies in which algorithms interact with individuals. I analyzed data-driven microtargeting on social media, how recommender systems intervene in political communication on social media, and how word embeddings-based ADM systems can result in unfair inferences. The analysis of each case study provided additional knowledge on the phenomena. This knowledge in turn revealed how algorithms reform social machines and create specific systems' equilibria. It also shed light on specific challenges regarding the function of social machines. For each case study, I revise my research findings, and discuss the related implications and challenges.

Social media and data-driven microtargeting

The analysis of the performance of data-driven microtargeting on social media (chapter two) revealed legal and technical possibilities and restrictions. Depending on the country, political actors have different legal freedom in collecting personal data, processing it, and deploying models for targeting individuals. The difference in available data also translates to a difference in datafication of political campaigns. For example, parties in Germany deploy far less data modeling techniques for assessing the attitudes of the electorate, in contrary to the US where campaigning is fully datafied. Regardless of the country, social media platforms prevail as key stakeholders in political data collection and deployment. They collect huge amounts of data about individuals, and this thesis showed how trivial it is to create voter profiles out of this data. Consequently they lie in the epicenter of political data-driven microtargeting, with political parties and other actors using the platforms' services for placing personalized ads. Although data-driven microtargeting is one of the most important parts in electoral campaigning today, my investigation shows that evidence proving its efficiency is sparse. Because assessing and measuring ads' influence on political voting is almost impossible, there are no scientific

studies that prove that microtargeting can influence public opinion to the exent that it will shape an election outcome.

Implications & Challenges

The above results have straightforward implications for understanding social media as spaces of political algorithmic campaigning. Not only do platforms' owners as designers decide how microtargeting will take place, making political actors dependent on their decisions, but they also hold the responsibility of protecting and meeting the expectations set by legislation regarding data privacy. Although users are the actual product of the platforms, since their data is collected and processed, their capacity to decide how microtargeting should take place remains restricted. The same applies for states and political actors. Existing regulations set standards and limits on the collected data, but they do not pose restrictions on how personalized advertising should take place, even though it directly intervenes with the function of the political system and individuals rights. The thesis results also show how socially constructed these processes are. Although there is no evidence showing that microtargeting influences more individuals than classical campaigning in their voting decisions, a whole political industry has been developed around this campaign method. Given the above, multiple scientific and political challenges remain open regarding microtargeting. First, there is an open challenge to quantify algorithmic influence on voters. Such a finding can help not only political actors in designing their campaigns, but also help in evaluating misinformation attempts based on personalized advertising and to generate regulatory frameworks that assure political justice. Second, it is a challenge to decouple the campaigning technique from the decisions of tech companies and platforms that were not per se designed to foster political communication. The ability of a candidate to pay advertising space on social media should not be a criterion for their potential success in political processes. Third, there is a further need for opacity in evaluating data-driven microtargeting. That means that both platforms and political actors should become much more transparent regarding the methods, data, and strategies they deploy to influence individuals and social groups. The resolution of the above issues can result not only in a general understanding of political microtargeting on social media, but also create the ground for changes that can promote diverse, open, and legitimized political conduct.

Political communication and recommender systems

In chapter three I analyzed how recommender systems interfere in political discourse taking place on social media platforms. The analysis provided insights both on the behaviour of political users, but also on how recommender systems interact with them. The investigation of users revealed that most of them remain politically passive, while a small proportion is responsible for a large part of the discourse. This activity asymmetry does not only exist in how much people participate in political discussions, but also about what they discuss. Hyperactive users are interested in different content than regular users. They also become opinion leaders, given that their generated content becomes more popular in discussions than those of the regular users. Analysis also showed that algorithms for content curation can suffer from serious biases because of the above asymmetries. The skewed user behavior can be a serious problem for training trustworthy models. Furthermore, even when models are trained ideally, inserting adversarial

examples in the network can easily distort system's suggestions. The analysis once again revealed the importance of platforms in political communication. The design of recommender systems and their objective is dependent on the platform owner's imperatives, explicitly decoupled from any notion of fair political communication. Furthermore, services are consciously not disclosing the systems' function, making it possible to study them only by simulations.

Implications & Challenges

The above results provide important insights on user behavior, political communication on social media, and point out dangers related to algorithmic influence. The first implication is related to user participation in political discourse. Social media is not an equitable space for political expression. Those who shout louder are the ones that are going to be heard and form political attention and political discussions. Second, algorithms de facto influence political communication, because they select contents seen by individuals. Because of their opacity, it is difficult to assess algorithmic influence, how exactly they interfere in political communication, as well as whether they are resilient to extreme user behavior and malicious attacks. It is a challenge, therefore, to define structures that can transparently point to the exact ways that these algorithms participate in agenda setting and content priming. Furthermore, it is an open challenge to regulate social media as spaces for political communication. Not only does user participation deviate from notions of inclusivity and uniform visibility, but so does algorithmic implementations violate assumptions of unbiased political exchange. Again, social media owners become key stakeholders in a political process, in which they had no intention of getting involved in. It is important, therefore, to reassess the responsibilities and rights of users, platform owners and political actors that use social media for political purposes. It is also crucial to evaluate whether algorithms that filter information should play the role of political communication mediator, an question of big theoretical and social importance.

Word embeddings and unfair algorithms

The third case study that the thesis dealt with was unfair inferences drawn by text-based ADM systems. Specifically, chapter five provided an overview of word embeddings, a set of techniques that map words into numerical vectors and are able to capture semantic properties of language. The investigation analyzed preexisting bias in the input text and showed how it is imprinted in word embeddings in terms of sexism, homophobia, and xenophobia. It illustrated that a difference in biases in the input text result in different types of bias in the embeddings. Furthermore, it investigated how biases in word embeddings are diffused within further models. Because the vectors are used in further machine learning models to improve their accuracy, the study showed that a sentiment classifier trained on the embeddings will suffer from similar biases, resulting in discrimination of individuals and social groups. I then analyzed two mitigation techniques and results demonstrated that bias mitigation should always take place at the final model level, due to the complexity of bias structure. Although bias in word embeddings is a serious issue, the study showed that biased word embeddings can be used for the detection of similar biases on new data. Overall, the study proved that word embeddings can lead to unfair

algorithms, and measures should be taken to ensure that any inferences are decoupled from any biases in the specific modeling technique.

Implications & Challenges

The analysis provides important insights on the use of word embeddings in ADM systems. Given that most models that use text today, such as text-based systems for translation, sentiment analysis, and question answering, employ word embeddings, these findings can be generalized. The first implication is that since text is always a projection of parts of the social world, it is inevitable that it will reproduce biases and asymmetries existing in that world. Second, because most scientists use freely accessible corpora to train their models, they incorporate biases into their models' inferences. Even for scientists who use standard mitigation techniques, the study showed that instead of removing biases they distort it. Third, there is currently no regulatory framework or standard guidelines that try to control biases in text-based ADM systems. Because of these important implications, a set of challenges emerge for ensuring fairness in social machines. A challenge is to find texts that models can be trained on, without resulting to biased inferences. At the very least, researchers should investigate techniques for bias mitigation at the models' final inference, regardless of the models' structure. Another challenge is the deployment of frameworks that can guide researchers to develop unbiased text-based models, as well as to integrate and study text-based models further in terms of fair outcomes, discrimination and hate speech. Given that such ADM systems are deployed in many services for communication, translation and decision making, such biases can violate laws and lead to socially unacceptable outcomes.

The above three case studies provided important insights on algorithmic influence on political machines. They revealed how designers' choices strongly shape how algorithms interact with individuals, how political communication on social media takes place, as well as whether a model will result in discriminative outcomes. Furthermore, the analysis dealt with the opacity of the investigated systems, which made research attempts highly difficult. The studies also pointed out that users and individuals have restricted power on the function of the algorithms, while a huge gap in regulation has been discovered, with governments not being in a position to assess algorithmic influence nor to provide frameworks that guarantee socially just processes. On a more general level, and given the five-point framework developed in the introduction, this thesis shows that any evaluation of algorithmic influence is contingent on and should always take into consideration dimensions of symbolic influence, political conduct, design and regulation. Changes in these fields change de facto how algorithmic influence will take place, as well shape how researchers can study algorithms and their impact on sociotechnological ecosystems.

6.2 Future work

This work shed light on multiple dimensions of political machines, uncovering important issues in the deployment of algorithms on social media platforms and ADM systems. Given the generated new knowledge and the wide space of interactions that can be studied, I provide possible avenues for further research.

The study on data driven microtargeting revealed specific research gaps and obstacles, as well as new opportunities. First, social media platforms are starting to make

targeting data public in the form of ad libraries. The analysis of these ad libraries has the potential to contribute to more transparent political interactions and facilitate the understanding of political actors. Furthermore, the use of the platforms' targeting tools in experimental settings can function as an auditing measure on algorithms and their function. The same applies for controlled and long term experiments on individuals exposed to microtargeted advertisement. Researchers can study whether the tools result in the mobilization, polarization or opinion change of individuals. Equally important is the further definition of what is legal and ethical in a politically personalized advertisement, as well as the investigation of what is a political advertisement and what is not. These directions can prepare the ground for changes in social media and political campaigning policy.

Similar to microtargeting on social media, the study on recommender systems and user political behavior raised new questions regarding social media as a space for political communication. Researchers should investigate whether hyperactive behavior overlaps with malicious, inauthentic and coordinated user behavior with the aim of spreading specific messages on the platforms. The study can be extended to further social media platforms, because news consumption and political discourse are multi- and crossplatform processes. Further research attempts should also focus on the influence of recommender systems on political attitudes and dynamic information diffusion. Such a study should quantify dangers and issues mentioned in this thesis, making recommender systems in political communication more tangible. Such findings need to be backed up and integrated into social theories about political communication, equality and participation. In this way, researchers and policy makers can quantify how political communication should take place online and if recommender systems should or should not be active mediators of publics' interests and opinions.

The investigation of algorithms in text-based ADM systems lies in an even more premature stage. The study showed that the existence of biases in word embeddings can lead to discrimination of individuals and social groups. Therefore, researchers need to investigate whether it is possible to generate artificial corpora that, on the one hand, preserve semantic associations in language, but also do not include text that is responsible for the emerging biases. Another important pathway is the study of contextual word embeddings, which are used by the major transformer architectures in NLP. Such embeddings encode much more complicated information within them. Therefore, the auditing and the understanding of that information is a challenging task that needs to be performed. Another future direction is to integrate text-based algorithms on general auditing and explainability tools that try to synchronize computing bias with legal frameworks. This presupposes that theoretic investigations define the boundaries of discrimination in text and decision making.

Overall, each case study created fertile ground for further research. From a general perspective, the further study of political machines needs access to the actual data used in the interactions, which are usually private or inaccessible. Therefore there is a need for the creation of mechanisms that give scientists the relevant information and tools in order to reveal properties of political machines. New privacy mechanisms such as differential privacy that guarantee the privacy rights of individuals might be a solution to that issue. Further questions arising when studying political machines should also be addressed, such as how algorithms in political machines should be audited and whether and how privacy and property rights should be changed in the digital era. These issues

lead to further legal and political questions about how to regulate political machines and how should political conduct take place on them. Equally important is the definition of ethical frameworks about equality, justice, and discrimination of individuals, social groups, political, public and private actors. Since political machines transform constantly, there are endless research pathways that scientists can take in order to generate new knowledge and support the functioning of society.

6.3 Outro

Algorithms and individuals interact constantly in the society in unexpected ways. Because sociotechnological ecosystems reform social and political processes, this thesis contributed to the understanding of political dimensions of algorithmic influence. From a descriptive perspective, researchers can effectively study phenomena under a cybernetic framework. Understanding political machines can contribute to understanding how technology constitutes individual decisions and socialization. It can also help in predicting possible future social states, detecting issues and challenges in the interactions taking place, and evaluating the implementation of technology as whole. Datafication and ubiquitous computing have created a social reality where data-intensive algorithms, statistical, machine and deep learning models are a cardinal part of the world. Political machines as a framework is able to evaluate the deployment of such models and offer insights on the complex interactions between algorithms and individuals.

The case studies of the thesis provided insights into the functioning of prominent social machines and phenomena taking place on them. Results showed that the use of algorithms for campaigning, content curation and ADM systems can lead to reformation and distortion of political processes, as well as lead to biases towards individuals and social groups. Furthermore, they demonstrated the existence of a political equilibrium in the machines, where platforms' owners and algorithmic designers are the central tensors and controllers in the systems' function. This systemic feature initiates a discussion about how social machines should be and how they should be designed. Most algorithmic implementations today are part of the commercial sector, with states having marginal control on them and regulators facing serious challenges. Furthermore, individuals and social groups are the most passive participants in the systems, usually taking either the role of the consumer, or being projected into datafied artifacts. From a normative perspective, society should reflect on these roles, and imagine and prescribe how socioalgorithmic systems should ideally be.

Centering the civic interest, and the idea that technology should serve individuals and the society in a way that ensures equality, justice, political freedom and social inclusiveness, the study of political machines should be extended to the study of civic machines. Researchers should not only describe how political machines function, but also define principles, frameworks, and constraints that can lead to the creation of sociotechnological ecosystems that serve the public interest. The design of civic machines prevails as a necessity in an environment where technological and algorithmic implementations influence society in an unexpected ways, transforming the political essence of society. This thesis made a first towards that direction, by defining political machines and unraveling issues and challenges of algorithmic influence. There is an endless space for further scientific investigation, and the new knowledge can be used to create sociotechnological ecosystems, by the society and for the society.

Bibliography

[1] O. Papakyriakopoulos, M. Shahrezaye, A. Thieltges, J. C. M. Serrano, and S. Hegelich. Social media und microtargeting in deutschland. *Informatik-Spektrum*, 40(4):327–335, 2017.

[2] O. Papakyriakopoulos, S. Hegelich, M. Shahrezaye, and J. C. M. Serrano. Social media and microtargeting: Political data processing and the consequences for germany. *Big Data & Society*, 5(2):2053951718811844, 2018. URL: `https://doi.org/10.1177/2053951718811844`, `arXiv:https://doi.org/10.1177/2053951718811844`, `doi:10.1177/2053951718811844`.

[3] O. Papakyriakopoulos, M. Shahrezaye, J. C. M. Serrano, and S. Hegelich. Distorting political communication: The effect of hyperactive users in online social networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 157–164. IEEE, 2019.

[4] O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich. Political communication on social media: A tale of hyperactive users and bias in recommender systems. *Online Social Networks and Media*, 15:100058, 2020.

[5] O. Papakyriakopoulos, S. Hegelich, J. C. M. Serrano, and F. Marco. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 446–457, New York, NY, USA, 2020. Association for Computing Machinery. URL: `https://doi.org/10.1145/3351095.3372843`, `doi:10.1145/3351095.3372843`.

[6] J. C. M. Serrano, S. Hegelich, M. Shahrezaye, and O. Papakyriakopoulos. *Social Media Report: The 2017 German Federal Elections*. TUM University Press, 2018.

[7] S. Engelmann, J. Grossklags, and O. Papakyriakopoulos. A democracy called facebook? participation as a privacy strategy on social media. In *Privacy Technologies and Policy: 6th Annual Privacy Forum, APF 2018, Barcelona, Spain, June 13-14, 2018, Revised Selected Papers*, pages 91–108. Springer, 2018.

[8] M. Shahrezaye, O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich. Estimating the political orientation of twitter users in homophilic networks. In *2019 AAAI Spring Symposium Series*, 2019.

[9] J. C. M. Serrano, M. Shahrezaye, O. Papakyriakopoulos, and S. Hegelich. The rise of germany's afd: A social media analysis. In *Proceedings of the 10th International Conference on Social Media and Society*, pages 214–223. ACM, 2019.

[10] M. Shahrezaye, O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich. Measuring the ease of communication in bipartite social endorsement networks: A proxy to study the dynamics of political polarization. In *Proceedings of the 10th International Conference on Social Media and Society*, pages 158–165. ACM, 2019.

[11] R. W. Emerson. *The conduct of life*, volume 6. Harvard University Press, 2003.

[12] V. Mayer-Schönberger and K. Cukier. *Big Data - A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt, Orlando, 2013.

[13] N. Shadbolt, D. De Roure, and W. Hall. *The Theory and Practice of Social Machines*. Springer, 2019.

[14] M. Foucault. *The history of sexuality: An introduction*. Vintage, 1990.

[15] N. Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.

[16] N. Luhmann. *Soziale systeme*. Suhrkamp, 1985.

[17] H. Von Foerster. *Understanding understanding: Essays on cybernetics and cognition*. Springer Science & Business Media, 2007.

[18] M. Mead. Cybernetics of cybernetics. *Purposive Systems. New York: Spartan Books*, 1968.

[19] W. R. Ashby. *An introduction to cybernetics*. Chapman & Hall Ltd., 1957.

[20] A. Rosenblueth, N. Wiener, and J. Bigelow. Behavior, purpose and teleology. *Philosophy of science*, 10(1):18–24, 1943.

[21] S. Brand. For god's sake, margaret: conversation with gregory bateson and margaret mead. *CoEvolution Quarterly*, 10:32–44, 1976.

[22] G. Bateson. A theory of play and fantasy. *The game design reader: A rules of play anthology*, pages 314–328, 2006.

[23] D. A. Novikov. *Cybernetics: From past to future*, volume 47. Springer, 2015.

[24] J. Ruesch, G. Bateson, E. C. Pinsker, and G. Combs. *Communication: The social matrix of psychiatry*. Routledge, 2017.

[25] R. Glanville. Second order cybernetics. *Systems Science and Cybernetics*, 3:59–85, 2002.

[26] 2020. URL: `https://www.youtube.com/watch?v=QJWDJvbeaOM&t=1s`.

[27] J. Hendler and A. M. Mulvehill. *Social machines: the coming collision of artificial intelligence, social networking, and humanity*. Apress, 2016.

[28] P. R. Smart and N. R. Shadbolt. Social machines. In *Encyclopedia of Information Science and Technology, Third Edition*, pages 6855–6862. IGI Global, 2015.

[29] D. Murray-Rust, S. Tarte, M. Hartswood, and O. Green. On wayfaring in social machines. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1143–1148. ACM, 2015.

[30] W. Roush. Social machines: Computing means connecting. *Technology Review*, 108(8):44, 2005.

[31] S. R. Meira, V. A. Buregio, L. M. Nascimento, E. Figueiredo, M. Neto, B. Encarnacao, and V. C. Garcia. The emerging web of social machines. In *2011 IEEE 35th Annual Computer Software and Applications Conference*, pages 26–27. IEEE, 2011.

[32] M. Aakhus. Understanding information and communication technology and infrastructure in everyday life: Struggling with communication-at-a-distance. In *Machines that become us*, pages 27–42. Routledge, 2017.

[33] U. Martin and A. Pease. Mathematical practice, crowdsourcing, and social machines. In *International Conference on Intelligent Computer Mathematics*, pages 98–119. Springer, 2013.

[34] T. Berners-Lee and M. Fischetti. *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. Diane Publishing Company, 2001.

[35] E. D. Hersh. *Hacking the electorate: How campaigns perceive voters*. New York: Cambridge University Press, 2015.

[36] D. Kreiss. *Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy*. New York: Oxford University Press, 2016.

[37] D. Murray-Rust and D. Robertson. Bootstrapping the next generation of social machines. In *Crowdsourcing*, pages 53–71. Springer, 2015.

[38] W. Hall, D. De Roure, and N. Shadbolt. The evolution of the web and implications for eresearch. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890):991–1001, 2008.

[39] J. E. Katz. *Machines that become us: The social context of personal communication technology*. Routledge, 2017.

[40] D. De Roure, C. Hooper, K. Page, S. Tarte, and P. Willcox. Observing social machines part 2: How to observe? In *Proceedings of the ACM Web Science Conference*, page 13. ACM, 2015.

[41] R. Tinati and L. Carr. Understanding social machines. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 975–976. IEEE, 2012.

[42] H. D. Lasswell. *Politics: Who gets what, when, how*. Pickle Partners Publishing, 2018.

[43] G. H. Mead. *Mind, self and society*, volume 111. Chicago University of Chicago Press., 1934.

[44] F. De Saussure. *Nature of the linguistic sign*. 1916.

[45] L. Vygotski. *Thought and language*. MIT press, 2012.

[46] T. A. Van Dijk. 18 critical discourse analysis. *The handbook of discourse analysis*, pages 349–371, 2001.

[47] N. Fairclough. *Critical discourse analysis: The critical study of language*. Routledge, 2013.

[48] P. Bourdieu. Symbolic power. *Critique of anthropology*, 4(13-14):77–85, 1979.

[49] J. Piaget. *The Psychology of Intelligence*. Routledge, 1947.

[50] B. Shneiderman and C. Plaisant. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India, 2010.

[51] D. Benyon. *Designing interactive systems: A comprehensive guide to HCI, UX and interaction design*. Pearson Edinburgh, 2014.

[52] R. King, E. F. Churchill, and C. Tan. *Designing with Data: Improving the User Experience with A/B Testing*. " O'Reilly Media, Inc.", 2017.

[53] L. Rainie, A. Smith, K. L. Schlozman, H. Brady, and S. Verba. Social media and political engagement. *Pew Internet & American Life Project*, 19:2–13, 2012.

[54] N. Gustafsson. The subtle nature of facebook politics: Swedish social network site users and political participation. *New Media & Society*, 14(7):1111–1127, 2012.

[55] B. D. Loader and D. Mercea. Networking democracy? social media innovations in participatory politics: Brian d. loader and dan mercea. In *Social Media and Democracy*, pages 12–21. Routledge, 2012.

[56] N. Fenton and V. Barassi. Alternative media and social networking sites: The politics of individuation and political participation. *The Communication Review*, 14(3):179–196, 2011.

[57] S. Joseph. Social media, political change, and human rights. *BC Int'l & Comp. L. Rev.*, 35:145, 2012.

[58] C. A. Rentschler. Rape culture and the feminist politics of social media. *Girlhood studies*, 7(1):65–82, 2014.

[59] G. Wolfsfeld, E. Segev, and T. Sheafer. Social media and the arab spring: Politics comes first. *The International Journal of Press/Politics*, 18(2):115–137, 2013.

[60] Beyond the vast wasteland: briefing congresspeople for the aspen institute. URL: `http://www.ethanzuckerman.com/blog/2019/07/31/beyond-the-vast-wasteland-briefing-congresspeople-for-the-aspen-institute/`.

[61] S. Hegelich and M. Shahrezaye. The communication behavior of german mps on twitter: Preaching to the converted and attacking opponents. *European Policy Analysis*, 1(2):155–74, 2015.

[62] F. Zuiderveen Borgesius, J. Möller, S. Kruikemeier, R. Ó Fathaigh, K. Irion, T. Dobber, B. Bodo, and C. H. de Vreese. Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, 14(1):82–96, 2018.

[63] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. M. Menczer. Detecting and tracking political abuse in social media. In *Fifth international AAAI conference on weblogs and social media*, 2011.

[64] P. N. Howard and S. Woolley. Political communication, computational propaganda, and autonomous agents-introduction. *International Journal of Communication*, 10(2016), 2016.

[65] R. Effing, J. Van Hillegersberg, and T. Huibers. Social media and political participation: are facebook, twitter and youtube democratizing our political systems? In *International conference on electronic participation*, pages 25–35. Springer, 2011.

[66] Y. Jin, B. F. Liu, and L. L. Austin. Examining the role of social media in effective crisis management: The effects of crisis origin, information form, and source on publics' crisis responses. *Communication research*, 41(1):74–94, 2014.

[67] P. N. Howard and M. R. Parks. Social media and political change: Capacity, constraint, and consequence, 2012.

[68] E. Zuckerman. New media, new civics? *Policy & Internet*, 6(2):151–168, 2014.

[69] W. L. Bennett. The personalization of politics: Political identity, social media, and changing patterns of participation. *The ANNALS of the American Academy of Political and Social Science*, 644(1):20–39, 2012.

[70] H. Idrees, M. Shah, and R. Surette. Enhancing camera surveillance using computer vision: a research note. *Policing: An International Journal*, 41(2):292–307, 2018.

[71] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[72] K. D. Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[73] N. Just and M. Latzer. Governance by algorithms: reality construction by algorithmic selection on the internet. *Media, Culture & Society*, 39(2):238–258, 2017.

[74] E. Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[75] E. Ruppert, E. Isin, and D. Bigo. Data politics. *Big Data & Society*, 4(2):2053951717717749, 2017.

[76] N. Seaver. Knowing algorithms. *digitalSTS: A Field Guide for Science & Technology Studies*, page 412, 2019.

[77] D. Lazer. The rise of the social algorithm. *Science*, 348(6239):1090–1091, 2015.

[78] E. Bozdag. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3):209–227, 2013.

[79] L. Taylor. What is data justice? the case for connecting digital rights and freedoms globally. *Big Data & Society*, 4(2):2053951717736335, 2017.

[80] F. R. McKelvey. Algorithmic media need algorithmic methods: Why publics matter. *Canadian Journal of Communication*, 39(4), 2014.

[81] D. Beer. The social power of algorithms, 2017.

[82] K. L. Mosier and L. J. Skitka. Human decision makers and automated decision aids: Made for each other? In *Automation and human performance*, pages 201–220. Routledge, 2018.

[83] J. Larus, C. Hankin, S. G. Carson, M. Christen, S. Crafa, O. Grau, C. Kirchner, B. Knowles, A. McGettrick, D. A. Tamburri, et al. When computers decide: European recommendations on machine-learned automated decision making, 2018.

[84] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*, 2017.

[85] I. Ajunwa, S. Friedler, C. E. Scheidegger, and S. Venkatasubramanian. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*, 2016.

[86] L. Introna and H. Nissenbaum. Defining the web: The politics of search engines. *Computer*, 33(1):54–62, 2000.

[87] C. Lustig, K. Pine, B. Nardi, L. Irani, M. K. Lee, D. Nafus, and C. Sandvig. Algorithmic authority: the ethics, politics, and economics of algorithms that interpret, decide, and manage. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1057–1062. ACM, 2016.

[88] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126. ACM, 2016.

[89] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[90] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[91] A. Newell, H. A. Simon, et al. *Human problem solving*. Prentice-hall Englewood Cliffs, NJ, 1972.

[92] F. Pasquale. *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press, 2015.

[93] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Data and discrimination: converting critical concerns into productive inquiry*, pages 1–23, 2014.

[94] P. Hitlin and L. Rainie. Facebook algorithms and personal data. *Pew Research Center, Jan*, 16, 2019.

[95] S. Barocas, S. Hood, and M. Ziewitz. Governing algorithms: A provocation piece. *Available at SSRN 2245322*, 2013.

[96] G. Langlois and G. Elmer. The research politics of social media platforms. *Culture machine*, 14, 2013.

[97] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Governance in social media: A case study of the wikipedia promotion process. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[98] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, 2018.

[99] M. Ananny. Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, 41(1):93–117, 2016.

[100] C. Castoriadis. *The imaginary institution of society.* Cambridge, MA: MIT Press, 1997.

[101] J. C. Bertot, P. T. Jaeger, S. Munson, and T. Glaisyer. Social media technology and government transparency. *Computer*, 43(11):53–59, 2010.

[102] E.U.-Directive. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119:1–88, May 2016.

[103] J. Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2016.

[104] T. Zarsky. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1):118–132, 2016.

[105] M. Ziewitz. Governing algorithms: Myth, mess, and methods. *Science, Technology, & Human Values*, 41(1):3–16, 2016.

[106] N. Seaver. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 4(2):2053951717738104, 2017.

[107] J. C. Bertot, P. T. Jaeger, and D. Hansen. The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government information quarterly*, 29(1):30–40, 2012.

[108] L. Introna and D. Wood. Picturing algorithmic surveillance: The politics of facial recognition systems. *Surveillance & Society*, 2(2/3):177–198, 2004.

[109] N. Diakopoulos. Algorithmic accountability reporting: On the investigation of black boxes, 2014.

[110] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[111] J. Pfeffer, T. Zorbach, and K. M. Carley. Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, 20(1-2):117–128, 2014.

[112] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 211–223. ACM, 2014.

[113] P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier. Evolution of reddit: from the front page of the internet to a self-referential community? In *Proceedings of the 23rd international conference on world wide web*, pages 517–522. ACM, 2014.

[114] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

[115] M. M. Malik and J. Pfeffer. Identifying platform effects in social media data. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[116] H. Schoen, D. Gayo-Avello, P. Takis Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor. The power of prediction with social media. *Internet Research*, 23(5):528–543, 2013.

[117] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*, 2015.

[118] O. Varol, E. Ferrara, F. Menczer, and A. Flammini. Early detection of promoted campaigns on social media. *EPJ Data Science*, 6(1):13, 2017.

[119] A. Thieltges and S. Hegelich. Manipulation in sozialen netzwerken. *ZfP Zeitschrift für Politik*, 64(4):493–512, 2017.

[120] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[121] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[122] S. Hegelich and D. Janetzko. Are social bots on twitter political actors? empirical evidence from a ukrainian social botnet. In *Tenth International AAAI Conference on Web and Social Media*, pages 1–4, 2016.

[123] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[124] C. Wagner, S. Mitter, C. Körner, and M. Strohmaier. When social bots attack: Modeling susceptibility of users in online social networks. In *# MSM*, pages 41–48, 2012.

[125] A. Thieltges, F. Schmidt, and S. Hegelich. The devil's triangle: Ethical considerations on developing bot detection methods. In *2016 AAAI Spring Symposium Series*, 2016.

[126] P. Covington, J. Adams, and E. Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198. ACM, 2016.

[127] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, and E. H. Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 456–464, New York, NY, USA, 2019. ACM, Association for Computing Machinery. URL: `https://doi.org/10.1145/3289600.3290999`, `doi:10.1145/3289600.3290999`.

[128] M. A. DeVito. From editors to algorithms: A values-based approach to understanding story selection in the facebook news feed. *Digital Journalism*, 5(6):753–773, 2017.

[129] E. Rader and R. Gray. Understanding user beliefs about algorithmic curation in the facebook news feed. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 173–182. ACM, 2015.

[130] J. Yang. Effects of popularity-based news recommendations ("most-viewed") on users' exposure to online news. *Media Psychology*, 19(2):243–271, 2016.

[131] S. Krishnan, J. Patel, M. J. Franklin, and K. Goldberg. A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 137–144. ACM, 2014.

[132] R. Faris, H. Roberts, B. Etling, N. Bourassa, E. Zuckerman, and Y. Benkler. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center Research Publication*, 6, 2017.

[133] S. Stieglitz and L. Dang-Xuan. Social media and political communication: a social media analytics framework. *Social network analysis and mining*, 3(4):1277–1291, 2013.

[134] Z. Tufekci and C. Wilson. Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of communication*, 62(2):363–379, 2012.

[135] K. Holt, A. Shehata, J. Strömbäck, and E. Ljungberg. Age and the effects of news media attention and social media use on political interest and participation: Do social media function as leveller? *European Journal of Communication*, 28(1):19–34, 2013.

[136] I. Himelboim, S. McCreery, and M. Smith. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of computer-mediated communication*, 18(2):154–174, 2013.

[137] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[138] B. E. Weeks, A. Ardèvol-Abreu, and H. Gil de Zúñiga. Online influence? social media use, opinion leadership, and political persuasion. *International Journal of Public Opinion Research*, 29(2):214–239, 2017.

[139] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 211–220. ACM, 2009.

[140] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332, 2014.

[141] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[142] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[143] S. Flaxman, S. Goel, and J. M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

[144] K. Endres. The accuracy of microtargeted policy positions. *PS: Political Science & Politics*, 49(4):771–774, 2016.

[145] S. Kruikemeier, M. Sezgin, and S. C. Boerman. Political microtargeting: Relationship between personalized advertising on facebook and voters' responses. *Cyberpsychology, Behavior, and Social Networking*, 19(6):367–372, 2016.

[146] B. C. Schipper and H. Woo. Political awareness, microtargeting of voters, and negative electoral campaigning. *Microtargeting of Voters, and Negative Electoral Campaigning (May 2, 2017)*, 2017.

[147] S. Hegelich and J. C. M. Serrano. Microtargeting.

[148] S. Kruschinski and A. Haller. Restrictions on data-driven political micro-targeting in germany. *Internet Policy Review*, 6(4), 2017.

[149] S. Shorey and P. Howard. Automation, big data and politics: A research review. *International Journal of Communication*, 10, 2016.

[150] J. Burkell and P. M. Regan. Voter preferences, voter manipulation, voter analytics: policy options for less surveillance and more autonomy, December 2019. URL: `https://policyreview.info/articles/analysis/voter-preferences-voter-manipulation-voter-analytics-policy-options-less`.

[151] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[152] J. Pfeffer, K. Mayer, and F. Morstatter. Tampering with twitter's sample api. *EPJ Data Science*, 7(1):50, 2018.

[153] D. Ruths and J. Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.

[154] M. Klašnja, P. Barberá, N. Beauchamp, J. Nagler, and J. Tucker. Measuring public opinion with social media data. In *The Oxford handbook of polling and survey methods*. Oxford University Press, 2017.

[155] M. Strohmaier. A few thoughts on engineering social machines. In *WWW (Companion Volume)*, pages 919–920, 2013.

[156] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*, pages 183–194. ACM, 2008.

[157] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

[158] S. Athey. The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*, pages 507–547. University of Chicago Press, 2018.

[159] R. Parasuraman and M. Mouloua. *Automation and human performance: Theory and applications*. Routledge, 2018.

[160] S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

[161] K. Allix, T. F. Bissyandé, Q. Jérome, J. Klein, Y. Le Traon, et al. Empirical assessment of machine learning-based malware detectors for android. *Empirical Software Engineering*, 21(1):183–211, 2016.

[162] M. Lorent, H. Maalmi, P. Tessier, S. Supiot, E. Dantan, and Y. Foucher. Meta-analysis of predictive models to assess the clinical validity and utility for patient-centered medical decision making: application to the cancer of the prostate risk assessment (capra). *BMC medical informatics and decision making*, 19(1):2, 2019.

[163] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence, and C. P. Langlotz. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*, 287(1):313–322, 2018.

[164] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline. Machine learning for medical imaging. *Radiographics*, 37(2):505–515, 2017.

[165] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[166] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

[167] N. Mehrabi, T. Gowda, F. Morstatter, N. Peng, and A. Galstyan. Man is to person as woman is to location: Measuring gender bias in named entity recognition, 2019. arXiv:1910.10872.

[168] H. Heidari, C. Ferrari, K. Gummadi, and A. Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, pages 1265–1276, 2018.

[169] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[170] S. Verma and J. Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[171] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[172] C. Russell, M. J. Kusner, J. Loftus, and R. Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[173] D. C. Hsia. Credit scoring and the equal credit opportunity act. *Hastings LJ*, 30:371, 1978.

[174] R. K. Berg. Equal employment opportunity under the civil rights act of 1964. *Brook. L. Rev.*, 31:62, 1964.

[175] M. Orfield. Racial integration and community revitalization: Applying the fair housing act to the low income housing tax credit. *Vand. L. Rev.*, 58:1747, 2005.

[176] A. Xiang and I. D. Raji. On the legal compatibility of fairness definitions. *arXiv preprint arXiv:1912.00761*, 2019.

[177] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[178] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[179] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[180] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, page 1–1, 2019. URL: http://dx.doi.org/10.1109/TVCG.2019.2934619, doi:10.1109/tvcg.2019.2934619.

[181] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 2016.

[182] B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, July 1996. URL: `https://doi.org/10.1145/230538.230561`, `doi:10.1145/230538.230561`.

[183] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[184] R. Dobbe, S. Dean, T. Gilbert, and N. Kohli. A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *arXiv preprint arXiv:1807.00553*, 2018.

[185] K. Holstein, H. Daumé III, M. Dudík, and H. Wallach. Opportunities for machine learning research to support fairness in industry practice. In *Workshop on Critiquing and Correcting Trends in Machine Learning at the Conference on Neural Information Processing Systems*, 2018.

[186] A. Chouldechova and A. Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[187] S. Bird, K. Kenthapadi, E. Kiciman, and M. Mitchell. Fairness-aware machine learning: Practical challenges and lessons learned. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 834–835, 2019.

[188] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996.

[189] W. Pietsch. Aspects of theory-ladenness in data-intensive science. *Philosophy of Science*, 82(5):905–916, 2015.

[190] Z. Obermeyer and E. J. Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

[191] R. Kitchin. Big data, new epistemologies and paradigm shifts. *Big data & society*, 1(1):2053951714528481, 2014.

[192] D. Boyd and K. Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[193] J. K. Blitzstein and J. Hwang. *Introduction to probability*. Chapman and Hall/CRC, 2014.

[194] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57–57, 1992.

[195] J. Frederick, W. Gravetter, B. Larry, and L.-A. B. Forzano. *Essentials of Statistics for the Behavioral Sciences*. Cengage Learning, 2013.

[196] A. D. Broido and A. Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1017, 2019.

[197] M. L. Weitzman. Fat-tailed uncertainty in the economics of catastrophic climate change. *Review of Environmental Economics and Policy*, 5(2):275–292, 2011.

[198] G. Csányi and B. Szendrői. Structure of a large social network. *Physical Review E*, 69(3):36131, 2004.

[199] A. C. Harvey. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, volume 52. Cambridge University Press, 2013.

[200] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 324–332. ACM, 2018.

[201] L. Muchnik, S. Pei, L. C. Parra, S. D. Reis, J. S. Andrade Jr, S. Havlin, and H. A. Makse. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, 3:1783, 2013.

[202] G. S. Maddala and K. Lahiri. *Introduction to econometrics*, volume 2. Macmillan New York, 1992.

[203] G. A. Seber and A. J. Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[204] J. M. Hilbe. *Logistic regression models*. Chapman and hall/CRC, 2009.

[205] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[206] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of machine learning research*, 2(Dec):125–137, 2001.

[207] S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[208] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[209] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[210] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. In *Linear Algebra*, pages 134–151. Springer, 1971.

[211] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[212] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[213] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 2010.

[214] A. Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[215] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, 2002.

[216] T. Joachims. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pages 200–209, 1999.

[217] D. R. Cox. *Principles of statistical inference*. Cambridge university press, 2006.

[218] I. Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[219] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[220] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[221] D. Anderson and K. Burnham. Model selection and multi-model inference. *Second. NY: Springer-Verlag*, 63, 2004.

[222] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, 2009.

[223] I. Rivals and L. Personnaz. On cross validation for model selection. *Neural computation*, 11(4):863–870, 1999.

[224] D. K. Tasoulis, N. M. Adams, and D. J. Hand. Unsupervised clustering in streaming data. In *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*, pages 638–642. IEEE, 2006.

[225] R. Kneser and H. Ney. Improved clustering techniques for class-based statistical language modelling. In *Third European Conference on Speech Communication and Technology*, 1993.

[226] K. P. Burnham and D. R. Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[227] C. D. Manning, C. D. Manning, and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[228] D. Sarkar. *Text Analytics with python*. Springer, 2016.

[229] T. Kudo and J. Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

[230] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[231] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012. URL: https://doi.org/10.1145/2133806.2133826, doi:10.1145/2133806.2133826.

[232] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288, 2013.

[233] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[234] C. Roberts, E. Davies, and T. Jupp. *Language and discrimination*. Routledge, 2014.

[235] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[236] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[237] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007. URL: https://doi.org/10.1162/neco.2007.19.10.2756, arXiv:https://doi.org/10.1162/neco.2007.19.10.2756, doi:10.1162/neco.2007.19.10.2756.

[238] Berné, O., Joblin, C., Deville, Y., Smith, J. D., Rapacioli, M., Bernard, J. P., Thomas, J., Reach, W., and Abergel, A. Analysis of the emission of very small dust particles from spitzer spectro-imagery data using blind signal separation methods ***. *A&A*, 469(2):575–586, 2007. URL: https://doi.org/10.1051/0004-6361:20066282, doi:10.1051/0004-6361:20066282.

[239] Z. Akata, C. Thurau, and C. Bauckhage. Non-negative Matrix Factorization in Multimodality Data for Segmentation and Label Prediction. In A. Wendel, S. Sternig, and M. Godec, editors, *16th Computer Vision Winter Workshop*, Mitterberg, Austria, February 2011. URL: https://hal.inria.fr/hal-00652879.

[240] L. Taslaman and B. Nilsson. A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data., 2012. URL: https://lup.lub.lu.se/search/ws/files/4197352/3735521.pdf, doi:10.1371/journal.pone.0046331.

[241] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 569–577, New York, NY, USA, 2008. Association for Computing Machinery. URL: https://doi.org/10.1145/1401890.1401960, doi:10.1145/1401890.1401960.

[242] J. Foulds, L. Boyles, C. DuBois, P. Smyth, and M. Welling. Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 446–454, New York, NY, USA, 2013. Association for Computing Machinery. URL: https://doi.org/10.1145/2487575.2487697, doi:10.1145/2487575.2487697.

[243] R. Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 391–402, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[244] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 262–272, USA, 2011. Association for Computational Linguistics.

[245] F. Krasnov and A. Sen. The number of topics optimization: Clustering approach. *Machine Learning and Knowledge Extraction*, 1(1):416–426, 2019. URL: https://www.mdpi.com/2504-4990/1/1/25.

[246] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1105–1112, New York, NY, USA, 2009. Association for Computing Machinery. URL: https://doi.org/10.1145/1553374.1553515, doi:10.1145/1553374.1553515.

[247] R. Deveaud, E. SanJuan, and P. Bellot. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1):61–84, 2014.

[248] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[249] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. arXiv:1810.04805.

[250] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[251] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[252] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2018. URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[253] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019. arXiv: 1906.08237.

[254] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C. J. Kuo. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8, 2019. URL: http://dx.doi.org/10.1017/ATSIP.2019.12, doi:10.1017/atsip.2019.12.

[255] A. Bakarov. A survey of word embeddings evaluation methods, 2018. arXiv: 1801.09536.

[256] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/P14-1023, doi:10.3115/v1/P14-1023.

[257] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. arXiv:1301.3781.

[258] P. Blair, Y. Merhav, and J. Barry. Automated generation of multilingual clusters for the evaluation of distributed representations, 2016. arXiv:1611.01547.

[259] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997. doi:10.1162/neco.1997.9.8.1735.

[260] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. arXiv:1406.1078.

[261] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980. URL: https://doi.org/10.1007/BF00344251, doi:10.1007/BF00344251.

[262] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. arXiv:1706.03762.

[263] P. Covington, J. Adams, and E. Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 191–198, New York, NY, USA, 2016. Association for Computing Machinery. URL: https://doi.org/10.1145/2959100.2959190, doi:10.1145/2959100.2959190.

[264] B. Smith and G. Linden. Two decades of recommender systems at amazon.com. *IEEE Internet Computing*, 21(3):12–18, May 2017. doi:10.1109/MIC.2017.72.

[265] A. Greenstein-Messica and L. Rokach. Personal price aware multi-seller recommender system: Evidence from ebay. *Knowledge-Based Systems*, 150:14–26, 2018. URL: http://www.sciencedirect.com/science/article/pii/S0950705118300893, doi:https://doi.org/10.1016/j.knosys.2018.02.026.

[266] A. R. Gilpin. Table for conversion of kendall's tau to spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and psychological measurement*, 53(1):87–92, 1993.

[267] M. Kendall and J. Gibbons. *Rank Correlation Methods*. Charles Griffin Book. E. Arnold, 1990. URL: https://books.google.de/books?id=ly4nAQAAIAAJ.

[268] H. J. Ahn. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information Sciences*, 178(1):37–51, 2008. URL: http://www.sciencedirect.com/science/article/pii/S0020025507003751, doi:https://doi.org/10.1016/j.ins.2007.07.024.

[269] URL: https://sifter.org/~simon/journal/20061211.html.

[270] R. Kumar, B. Verma, and S. S. Rastogi. Social popularity based svd++ recommender system. *International Journal of Computer Applications*, 87(14), 2014.

[271] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' guide*, volume 9. Siam, 1999.

[272] Facebook. Using the graph api - documentation. URL: https://developers.facebook.com/docs/graph-api/using-graph-api/.

[273] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 173–182, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. URL: https://doi.org/10.1145/3038912.3052569, doi:10.1145/3038912.3052569.

[274] H. Wang, N. Wang, and D.-Y. Yeung. Collaborative deep learning for recommender systems, 2014. arXiv:1409.2944.

[275] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He. Deepfm: A factorization-machine based neural network for ctr prediction, 2017. arXiv:1703.04247.

[276] J. Tang and K. Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 565–573, New York, NY, USA, 2018. Association for Computing Machinery. URL: https://doi.org/10.1145/3159652.3159656, doi:10.1145/3159652.3159656.

[277] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec. Graph convolutional neural networks for web-scale recommender systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, 2018. URL: http://dx.doi.org/10.1145/3219819.3219890, doi:10.1145/3219819.3219890.

[278] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[279] P. Mohapatra, M. Rolínek, C. Jawahar, V. Kolmogorov, and M. Pawan Kumar. Efficient optimization for rank-based loss functions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[280] P. Adamopoulos and A. Tuzhilin. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):54, 2015.

[281] T. Silveira, M. Zhang, X. Lin, Y. Liu, and S. Ma. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10(5):813–831, 2019.