



Agenda Formation and Prediction of Voting Tendencies for European Parliament Election using Textual, Social and Network Features

Gautam Kishore Shahi¹ · Ali Sercan Basyurt² · Stefan Stieglitz² · Christoph Neuberger³

Accepted: 2 December 2024 / Published online: 23 December 2024
© The Author(s) 2024

Abstract

As per agenda-setting theory, political agenda is concerned with the government's agenda, including politicians and political parties. Political actors utilize various channels to set their political agenda, including social media platforms such as Twitter (now X). Political agenda-setting can be influenced by anonymous user-generated content following the Bright Internet. This is why speech acts, experts, users with affiliations and parties through annotated Tweets were analyzed in this study. In doing so, the agenda formation during the 2019 European Parliament Election in Germany based on the agenda-setting theory as our theoretical framework, was analyzed. A prediction model was trained to predict users' voting tendencies based on three feature categories: social, network, and text. By combining features from all categories logistical regression leads to the best predictions matching the election results. The contribution to theory is an approach to identify agenda formation based on our novel variables. For practice, a novel approach is presented to forecast the winner of events.

Keywords 2019 European parliament election · Social media analytics · User's voting prediction · Agenda-setting theory · Political agenda

1 Introduction

Social media has developed into a platform that is frequently utilized by the public to communicate, express, share beliefs and opinions, engage with others, and disclose personal information voluntarily. Users can share content on social media without strict scrutiny (Myers West, 2018), which provides its users, in accordance with one of the principles of the Bright Internet, the freedom to express one's opinion publicly (Lee et al., 2018). Shared information can be aggregated and analyzed to make predictions using posts on social media and can positively benefit and aid decision-making, such as business decisions (Kushwaha et al., 2021), benefit the training of chatbot responses (Kushwaha & Kar, 2021), help evaluate buyer experience in buyer-supplier relationships (Kumar et

al., 2022) or even be incorporated into Smart Cities for economic growth and increased liveability (Allam & Dhunny, 2019), winner of election (Maldonado & Sierra, 2015).

During crises or elections, social media platforms are used for sharing opinions and interacting with other posts based on one's political beliefs. Agenda-setting theory describes the ability of expert actors such as news outlets or political journalists, users with political affiliations, and politicians to influence the public discussion and polarise belief systems (Baran, 2015). Examining the relation of agenda-setting by traditional media, politicians and political parties on social media indicates that they are closely intertwined with each other and that they influence each other (Gilardi et al., 2022). This is due to research indicating that the traditional media agenda, the social media agenda of parties, and the social media agenda of politicians all influence each other. However, no single agenda predominantly leads the others; instead, they exert mutual influence (Gilardi et al., 2022). Moreover, this indicates that campaigns may be important for constraining and enabling parties and politicians to push their own agendas (Gilardi et al., 2022). Furthermore, the consumption of media related to politics from the country of origin of external voters living in another country positively influences their election participation (Himmelroos

✉ Gautam Kishore Shahi
gautam.shahi@uni-due.de

¹ Faculty of Informatics, University of Duisburg-Essen, Duisburg, Germany

² Faculty of Economics and Social Sciences, University of Potsdam, Potsdam, Germany

³ Department of Political and Social Sciences, Free University of Berlin, Berlin, Germany

& von Schoultz, 2023). A positive effect of social media in general on the public's participation in politics online and offline has been observed to be higher among youth (Tariq et al., 2022). Moreover, analysis of user-generated political content on social media indicates that the more frequently individuals observe others in their network engaging in politics, the more they participate in the political discourse on social media; this effect is stronger in networks with similar people (Kim & Ellison, 2022).

Twitter (now X)¹ has been prominently used during elections by users with political affiliations, political parties and candidates (Shahi & Majchrzak, 2022b). During the present study, platforms such as Facebook restricted the data availability for research quite strictly; however, Tweets were accessible through Twitter Application Programming Interfaces (APIs). Twitter's importance during elections is increasing due to the potential outreach of Tweets and the possibility of following the discussion based on hashtags. This enables users to tag their Tweet with a hashtag relevant to a specific event and to start a discussion or to follow and join debates already in progress (Bruns & Burgess, 2015).

The 2019 European Parliament Election in Germany is analyzed in the present study based on online communication of German-speaking users on Twitter (Shahi et al., 2022). The study focuses on agenda formation and voting prediction for three major political parties during the 2019 election (Wiliarty, 2023): *Christian Democratic Union of Germany (CDU/CSU)*, *Social Democratic Party of Germany (SPD)*, and *a merger of Alliance 90, The Greens and The Green Party (Alliance 90/The Greens)*. The CDU/CSU is considered a conservative party closely aligned with Christian values, the Alliance 90/The Greens is a left-leaning party that prioritizes environmental protection and climate change, and the SPD is a center-left political party dedicated to workers' rights and a welfare state. Some important topics that were discussed during the election phase, were European integration, European fundamental values concerning multiculturalism, LGBTQ rights, and gender equality (Galpin & Trenz, 2019). In the literature, the role of social media has already been analyzed in US Elections and Brexit (Hall et al., 2018). The 2019 European Parliament Election was the first election of its kind after Brexit, that raised several uncertainties about future of European Union. The 2019 European Parliament Election was conducted between 23 and 26 May 2019, and the voting day for Germany was 26 May 2019. The election was held for 705 Members of the European Parliament (MEPs), with 96 candidates from Germany. Germany has the highest number of MEPs among the 27 member states of the European Parliament.

¹ In the manuscript, we used the old name Twitter because data were collected in 2019, i.e. before Twitter changed its name to X

Social media platforms provide an additional and direct channel to reach a broad audience and potentially set the agenda of topics to the public and specific groups in virtual environments such as echo chambers (Shahi et al., 2022). In previous research, prediction approaches not relying on social media data, for example, to predict the behavior of customers, were applied (Chen et al., 2005); however, user opinions expressed on social media platforms have also been leveraged to forecast sales (Benthaus & Skodda, 2015) and make predictions for the stock market (Nann et al., 2013; Nofer & Hinz, 2015), televised media events (Stieglitz et al., 2020), box office results (Feng et al., 2024) and even the intention of social media users to vote in elections (Maldonado & Sierra, 2015). When examining the approaches that were taken in the previous literature to predict the election or voting tendency of an audience, machine learning approaches relying on data acquired from Twitter focused on making predictions based mainly on sentiment and data directly available from Twitter (Budiharto & Meiliana, 2018; Kristiyanti et al., 2019).

With the advancement of generative artificial intelligence (GAI), large language models (LLM) are being discussed as a way to replace human annotation and prediction models (Nguyen & Rudra, 2024). Currently, open domain chatbots can answer a multitude of questions (Kocoń et al., 2023), which is why we tested ChatGPT² provided by OpenAI to determine the voting tendencies of users based on individual Tweets. We provided the concept of voting tendency and asked ChatGPT to determine the voting tendencies in Tweets. The following prompts were used:

Prompt 1 (To provide the task description): Based on the given Tweet, can you predict voting tendency for the 2019 European parliament election in Germany?

Prompt 2 (To predict the voting tendency from the given Tweets): also kann man die spd nicht wählen (Translation-So you can not vote for the SPD).

However, ChatGPT answered that the given question violates their content policy and cannot provide explicit decisions because of German laws. Overall, the current ChatGPT cannot predict voting tendencies reliably because it is not able to recognize the context of Tweets (Kocoń et al., 2023). Hence, even with the advancement of LLMs, we need human-annotated data for identifying the speech_act and building prediction models for predicting voting tendency.

Agenda-setting theory was used as the theoretical framework of the current study. Using data enrichment and manual coding, several other pieces of information were gathered for building the prediction model, such as the speech type within Tweets, for example calling to elect a certain candidate or forecasting how a candidate will perform. Utilizing extracted data to train machine learning algorithms could improve their

² <https://chat.openai.com/>

results (Chauhan et al., 2021). In addition, information such as gender and account age of users, whether the user is a bot, which was looked at through the preprocessing of raw Tweets, was used to train the model used in the present study (Shahi & Kana Tsoplefack, 2022a). The approach used combines social, network, and context-based features to build the prediction model. A combination of the above features has not been done in research before. Thus far, election predictions have been made using sentiment analysis or textual data. Furthermore, a combination of social, network and context-based features from user-generated social media content has not been used in research to our knowledge to analyze and determine their influence on the public's political agenda per the agenda-setting theory. Thus, the present study aims to address this research gap by answering the following research question:

RQ1: How did different actors influence agenda formation during the 2019 European Parliament Election in the German Twittersphere?

RQ2: How does adding different features improve the performance of the prediction model in predicting voting tendencies?

To answer these research questions, first, the literature on predictive analytics and agenda-setting theory was reviewed. Afterwards, Tweets during the 2019 European election in Germany were collected using the Twitter API. Texts were further split to find variables such as object name, party, and speech act, which is further described in Section 3.2.

To answer RQ1, the role of different actors was analyzed (role_comp and role_spec described in Section 3.2) in agenda formation. The election manifestos of major political parties (CDU/CSU, SPD, and Alliance 90/The Greens) were studied. The speech act discussed by experts (role_act) and users with political affiliation (role_comp) were filtered, and, using frequent words, the important discussions by those users were gathered. Frequent words were further mapped with the election manifesto to find the agenda formation by role_spec and role_comp.

To answer RQ2, the extracted data were transformed to gather additional information, such as social and network-based features, which is discussed in Section 3.3. Using extracted features, a prediction model (mentioned in Section 3.4) was built using a variety of combinations of these features, and their impact on the result was determined. Then, an error analysis was conducted to evaluate the cause of errors, which is discussed in Section 3.4.1.

By addressing these research questions, a contribution to the literature is made per the type 2 contribution definition of machine learning in information systems research for understanding phenomena using machine learning for causal inference instead of contributing a new ML method (Padmanabhan et al., 2022). This is done by determining the most effective of our novel variables that are not directly available

through Twitter on the prediction of the event under consideration. This provides researchers with an approach to identify the political agenda formation as part of agenda-setting theory and more accurately predict the results of different events by leveraging the most effective variables in predicting the voting tendencies of the public in the analyzed case. These variables can be directly included in prediction algorithms for better results as well as be specifically monitored when analyzing agenda-setting, specifically the political agenda, elections, and the response to policies for policy-making because of them being the most influential predictors for the public's voting tendencies within the set of variables. Their influence on the prediction of voting tendencies highlights their potential in analyzing and predicting the setting of political agendas on social media platforms. These variables and the prediction approach are novel contributions to research and also contribute to practice by enabling different actors, such as political parties or news outlets, to predict event results more accurately and act or prepare accordingly based on the identified variables and prediction approach.

The remainder of the present paper is organized as follows: Section 2 presents the literature on predicting public events based on social media data analysis and agenda-setting theory as the theoretical background and framework. Section 3 discusses the research method used, which includes the steps involved in the study, Section 4 outlines the results and implementation. Finally, Section 5 presents the observations made based on the result of the present study, and Section 6 provides the conclusion and scope for future work.

2 Related Work and Theoretical Background

This section discusses the related work done in predictive analytics by focusing on elections and agenda-setting theory, which is done by presenting the relevant literature in these domains. Furthermore, propositions developed based on the referenced literature are presented to explain the phenomenon more objectively (Kar et al., 2023; Kar & Dwivedi, 2020; Andersen et al., 2009; Kushwaha et al., 2021).

The referenced literature was identified through keyword searches on literature databases such as the AIS eLibrary and Scopus and checking relevant papers, their references, and papers referencing the identified papers further for their relevance. The keywords used were 'predictive analytics' and 'agenda-setting theory' together with 'election', 'politics', or 'voting tendencies'.

2.1 Predictive Analytics

'Predictive analytics' is a term used to describe an approach that uses data from the past to predict a specific outcome and, in doing so, forecast the future by combining mathe-

matics, statistics, and machine learning (Zakir et al., 2015). Predictive analytics provides insights into what can be done to increase the probability of a desired outcome and, in general, practical value in the decision-making process (Shi-Nash & Hardoon, 2017). This is valuable for different areas such as business, education, medicine, entertainment, finance, marketing, communication, public decision-making, and politics, where the ability to proactively make decisions and develop strategies based on forecasts grounded in previous data is useful (Poornima & Pushpalatha, 2016). For these approaches, the data used and amount of data available to train models to make predictions based on hidden patterns are of the utmost importance. The increasing popularity of different social media platforms indicates that these platforms are a powerful source of data and information because of an extraordinary amount of daily content generated by users. Users utilize these platforms to express their opinions and thoughts about events and topics, allowing for the analysis of the communication around a certain event by leveraging these data to predict a certain event based on user opinions. These opinions have been shared by many researchers who have utilized social media data with machine learning techniques to make predictions about the future (Asur & Huberman, 2010; Birmingham & Smeaton, 2011; Jaidka et al., 2019; Kim et al., 2021; Kumpulainen et al., 2020; Makazhanov & Rafiei, 2013; Tumasjan et al., 2010). Moreover, researchers are constantly trying to improve the prediction models. Furthermore, research has also shown that the inclusion of social media data for prediction purposes, such as the prediction of box office revenue in addition to traditional data such as movie characteristics like movie genre (Bogaert et al., 2021) or social media network data (Bogaert et al., 2016) significantly increase the prediction accuracy, emphasizing the potential of social media for prediction purposes.

A popular approach in research where social media data are leveraged to predict the outcome of an event is by mining the sentiment on social media platforms from user-generated content (Shahheidari et al., 2013). This technique has previously been employed to forecast the outcome of several political elections (Tumasjan et al., 2010; Nawaz et al., 2022). Prior research by Tumasjan et al. (2010), where a content analysis on Tweets related to political parties or candidates around the German Federal Election in 2009, was conducted, and voting behavior based on the frequency of Tweets mentioning a party and their sentiment was predicted—both studies showed that Twitter data can reflect the results of an election as well as that of the sentiment of Tweets, and the volume of positive Tweets can be leveraged to predict the outcome of political events. These results inspired further studies in this field (Asur & Huberman, 2010; Birmingham & Smeaton, 2011; Makazhanov & Rafiei, 2013), of which some did not support their notion of social media data being a reliable basis for predicting political events because of the

incorrect prediction of results (Jungherr et al., 2012). This was attributed to different factors, such as the Twitter users, on average, being young, well educated, and liberal (Barberá & Rivero, 2015; Mellon & Prosser, 2017). Nonetheless, it has been argued that Twitter volume and sentiment are good predictors; however, additional predictors such as the public role of a user could improve predictive models (Shi et al., 2012). Moreover, the findings by Bogaert et al. (2016) indicate that the prediction accuracy of different prediction models utilizing social media data is improved by including additional input variables, for example, network variables.

Proposition 1 *The inclusion of novel social media variables not directly available through the social media platform, such as whether an account belongs to a competitor in that event, improves the prediction results.*

The approach used in the present study identifies and applies additional factors that can be extracted from Tweets to improve the prediction model and predict the voting tendencies of the public more accurately. Table 1 shows a brief summary of the features used in the literature that are referenced in the present paper for the predictions of various events and scenarios.

2.2 Agenda-Setting Theory

Agenda-setting theory is focused on the notion that media consumption is not the only source of learning about a certain topic; instead, people also learn about its importance by evaluating the place and space of a specific topic (Baran, 2015). Furthermore, this theory proposes that the media does not dictate what people think; instead, it motivates what they think about (Cohen, 2015). This theory is used to describe how the media attempts to influence their viewers, and it differentiates between three types of agendas that are inter-related and influence each other: the public agenda, the media agenda, and the policy agenda. All three types are influenced by personal experience, the 'real world', and interpersonal communication (Dearing & Rogers, 1996). The focus of the present study lies on the policy agenda. This type of agenda is concerned with the government's agenda and includes the agendas of political parties, bureaucracies, presidents, committees, and the Lower House and Upper House (Soroka, 2002).

In politics, it is important that a candidate or party identifies important topics for the public and gives them more visibility, for which they get broader support from voters (Colomer & Llavador, 2012). The media and policy agenda are closely connected, influencing each other, which means policymakers are not independent of the media, and vice versa (Wolfe, 2012). This can be seen through political decisions that are then reported and topics in media that

Table 1 Summary of the features used in the literature for prediction of different events

Source	Data Source	Features of Data	Event/Scenario
Election			
Tumasjan et al. (2010)	Twitter	Tweet volume and Sentiment	German federal election results
Maldonado and Sierra (2015)	Twitter	Sentiment	2012 Dominican Republic Presidential Election voting intentions
Budiharto and Meiliana (2018)	Twitter	Sentiment & Tweet volume	Indonesia Presidential election
Kristiyanti et al. (2019)	Twitter	Sentiment	Indonesia Presidential election 2019-2024
Bermingham and Smeaton (2011)	Twitter	Sentiment & Tweet volume	Irish General Election
Jaidka et al. (2019)	Twitter	Sentiment, Tweet volume & social network information (e.g. page rank)	Election Malaysia, India, and Pakistan
Makazhanov and Rafei (2013)	Twitter	Sentiment and network features (e.g. retweet count, followers count)	Alberta 2012 general election
Shahheidari et al. (2013)	Twitter	Text	News, finance, job, movies, & sports sentiment
Shi et al. (2012)	Twitter	Tweet volume & Sentiment	American republican presidential election
Other Events			
Benthhaus and Skodda (2015)	Twitter	consumer information search behaviour and consumer emotions	Car sales
Nann et al. (2013)	Twitter & trading websites	messages and sentiment	Stock market
Nofer and Hinz (2015)	Twitter & Google consumer search information	Sentiment	Car sales
Stieglitz et al. (2020)	Twitter	Sentiment and Tweet volume at different times	Eurovision Song Contest
Asur and Huberman (2010)	Twitter	rate of Tweets are created & Sentiment	Movie box office revenue
Kim et al. (2021)	Reddit	Tweet volume and Sentiment	Academy award for best picture
Kumpulainen et al. (2020)	Twitter	Sentiment & Eurovision televoting scoring system	Eurovision song contest televoting
Bogaert et al. (2021)	Movie data, Facebook & Twitter	Tweet volume at different times and Sentiment	Movie box office sales
Bogaert et al. (2016)	Facebook	Social & Network features	Attendance soccer team games

politicians, in return, discuss. Nowadays, campaign messages representing the agenda of a politician and their party are not only exclusively disseminated through traditional media but also through social media platforms such as Twitter. The functionality of Twitter to follow accounts such as those of politicians ensures that the target group of a politician is up to date with their agenda, and by sharing the agenda of politicians, the reach of their message is amplified to individuals who are not directly following the politician. This makes Twitter a valuable tool in the context of the agenda-setting theory. Traditional media has lost some of its power in the agenda-setting process because of its prior power being divided between media generated by the public and traditional media (Meraz, 2009). Moreover, Twitter has become a frequently used tool by politicians, especially during their election campaigns (Vergeer, 2015), to communicate with journalists and the public (Barberá & Zeitzoff, 2018) and engage with political opponents (Russell, 2018). Twitter gives them a less-restrictive platform to express their opinions and agenda compared with traditional platforms, such as speeches shown in the news (Proksch & Slapin, 2015). Social media platforms have become essential in forming public opinion during elections because they allow the interaction between politicians, media, and the public as part of the so-called hybrid media system and traditional media such as television (Chadwick, 2017). Furthermore, the formation of individuals' opinions is influenced by the interactions and content available on social media platforms (Burbach et al., 2019).

Social media platforms are changing the role of the voter from passively consuming the content and opinions presented on traditional media to actively participating in the discussion around an election through creating their own content, disseminating new information and expressing a variety of opinions (Bakshy et al., 2012). Further, focusing on echo chambers and herd behavior political social media induced opinion polarization can be driven by exposure and participation in political discussions on social media platforms, further highlighting the role of social media and user-generated content on political agenda-setting (Kushwaha et al., 2022). Moreover, these individual interactions on social media in sharing and consuming opinions and other content result in information around an election being personalized for the individuals consuming them, amplifying their effect on the consumer (DeVito, 2017). This also gives users the ability to find any information, meaning information that not only supports their own opinions but also contradicts their opinions, further influencing the formulation of opinions on social media through the ability to either only seek out supportive information such as information from echo chambers or contradicting information for one's own opinion. Therefore, these conditions must be considered as possible factors of influence when setting the agenda, such

as the policy agenda and polarizing belief systems and virtual environments, because the effectiveness of the policy agenda expressed by political parties and politicians can be increased by considering these factors when communicating the agenda through social media.

Proposition 2 *The novel social media variables have an influence on the public's voting tendencies and, therefore, the setting of the political agenda.*

Therefore, Twitter is a useful tool for studying agenda-setting, and it was applied in this study to identify and analyze the agenda of different user groups on Twitter during the European election 2019 to acquire insights into the political agendas pushed by different groups and how it is done.

3 Research Methodology

In this section, the steps involved in the data collection, data preprocessing and enrichment, manual annotation, agenda formation, generation of the prediction model, and feature analysis are described. A graphical overview of the research methodology is presented in Fig. 1 and further explained in the below sections.

3.1 Data Collection and Feature Extraction

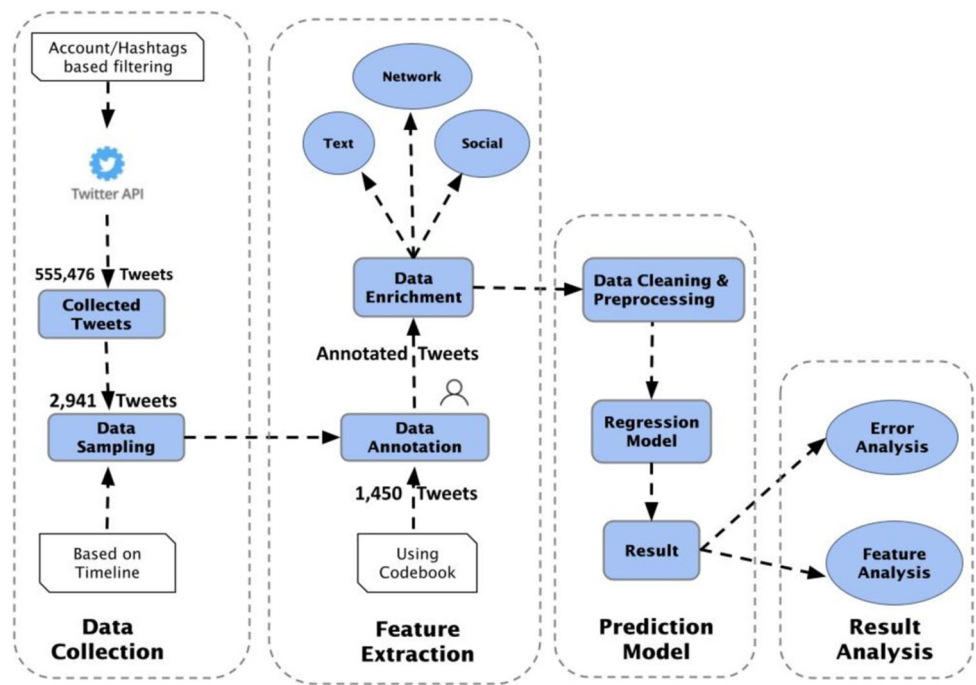
In this section, we explain the process of data collection, Tweet sampling, data annotation (manual feature extraction), and data feature extraction (text, social, and network-based). A short summary of all features used in the regression model is mentioned in Table 3.

3.1.1 Data Collection

For the analysis of the 2019 European Parliament Election in Germany, Tweets were collected during the election period using Twitter API V1.³ The data were collected through a self-developed crawler using the Python library Tweepy (Shahi et al., 2021; Diaz Ferreyra et al., 2023). Approximately three months of Twitter discussion were collected from February 28, 2019 (0:00 UTC), to May 26, 2019 (23:59 UTC). Trending and party-specific hashtags and Twitter accounts of candidates were used to collect Tweets referring to three major parties that competed for votes in the European Elections 2019 in Germany: the CDU/CSU, the SPD, and Alliance 90/The Greens. The hashtags were selected based on their usage frequency during the election period, the official Twitter account handle of the political party, and other nonpolitical organizations (e.g., media). For

³ <https://developer.x.com/en/docs/twitter-api/v1>

Fig. 1 Methodology used for the prediction of voting tendencies



each party, a list of different words was compiled consisting of party names, their top national candidates in Germany, and the names of their top candidates from each party family in Europe; details are presented in Table 2. Some top candidates are electable as a duo, so “Alliance 90/The Greens” included two names in the top candidate columns of Table 2. Collected Tweets contained at least one of the following predefined hashtags or Twitter account: #europawahl, #euw19, #EP2019, #EUWahl, #euelection, #Europawahl2019, #diesmalwähleich, #eleccionesUE2019, #EstaVezVoto, #elecciones, #eleccioneseuropas, #EE2019, #euelections2019, @PSOE, @EPinDeutschland, @spdde, @SPDEuropa, @CDU, @cducsubt, @CDU_CSU_EP, @EPP, @Europarl_ES, @PE_Espana, @EquipoEuropa, @populares, @ahorapodemos, @EuropeElects, @Europe-

Decides, @EPElections2019, @Europarl_EN, @TheProgressives, @ee_stats.

In total, 555,476 Tweets, including original Tweets, retweets, and replies, were collected in German. Retweets were excluded because the goal was to focus on original posts for public opinion formation. Overall, we retrieved 234,757 original Tweets and replies from May 12 to May 26, 2019, which were further sampled for analysis.

- CDU/CSU

- Party names: CDU, Christlich Demokratische Union, Christdemokraten; CSU, Christlich-Soziale Union, Christsoziale; CDU/CSU, Union, Unionsparteien; EVP, Europäische Volkspartei

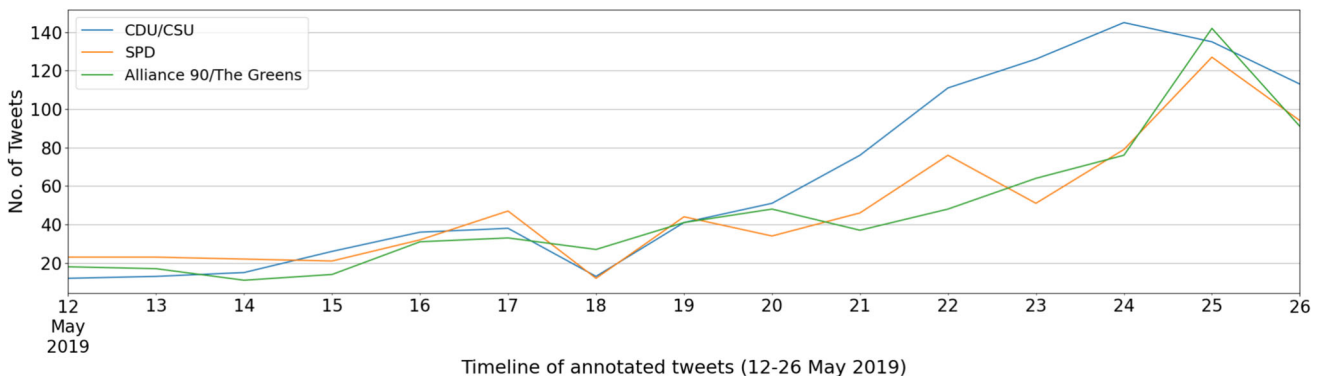


Fig. 2 Number of Tweets used in the study for each political party

Table 2 Different keywords and Twitter handles are used to collect Tweets for each party

Political Party	Variants of political party names	EU Level top candidate(Twitter handle)	National top candidates (Twitter handle)
CDU/CSU	CDU, Christlich Demokratische Union, CDU/CSU,CSU, Christdemokraten,Christsoziale, Unionsparteien,EVP, Christlich-Soziale Union,Union, Europäische Volksparte	Manfred Weber (@ManfredWeber)	Manfred Weber (@ManfredWeber)
SPD	SPD, Sozialdemokratische Partei Deutschlands, Sozialdemokratie, Sozialdemokraten, SPE, S&D, Sozialdemokratische Partei Europas, Progressive Allianz der Sozialdemokraten	Frans Timmermans (@TimmermansEU)	Katarina Barley(@katarinabarley), Udo Bullmann (@UdoBullmann)
Alliance 90/The Greens	Bündnis 90/Die Grünen, Die Grünen, Grüne, Grüne Partei	Ska(Franziska) Keller,Bas Eickhout (@BasEickhout)	Ska(Franziska) Keller, Sven Giegold (@sven_giegold)

- European top candidate: Manfred Weber, Twitter handle @ManfredWeber
- National to the candidate in Germany: Manfred Weber, Twitter handle @ManfredWeber

- SPD

- Party names: SPD, Sozialdemokratische Partei Deutschlands, Sozialdemokratie, Sozialdemokraten; SPE, Sozialdemokratische Partei Europas, Progressive Allianz der Sozialdemokraten, S&D
- European top candidate: Frans Timmermans Twitter handle @TimmermansEU
- National top candidates: Katarina Barley Twitter handle @katarinabarley, Udo Bullmann Twitter handle @UdoBullmann

- Alliance 90/The Greens

- Party names: Alliance 90/The Greens, Die Grünen, Grüne, Grüne Partei
- European top candidates: Ska (Franziska) Keller, Bas Eickhout Twitter handle @BasEickhout
- National top candidates: Ska (Franziska) Keller, Sven Giegold Twitter handle @sven_giegold

Based on Table 2, collected Tweets were searched using different hashtags and Twitter handles of each party. After filtering, CDU/CSU comprised $n(\text{Tweets, CDU/CSU}) = 116,767$, the subpopulation mentioning the SPD comprised $n(\text{Tweets, SPD}) = 47,154$, and the subpopulation mentioning Alliance 90/The Greens comprised $n(\text{Tweets, Alliance 90/The Greens}) = 5,975$. From the collected Tweets, we sam-

pled 2,000 Tweets for each party and used them for data annotation. A detailed description is provided in the data annotation section.

3.1.2 Data Annotation

First, from the filtered Tweets, the coding of statements was performed for each political party manually. The order of Tweets was randomized to ensure that specific Tweet clusters were not oversampled (e.g., date clusters). Tweets were manually coded resulting in a subsample of $n(\text{Tweets}) = 2,000$ for each party. Second, the three subsamples were merged, and further, duplicate Tweets were checked and removed. This procedure provided a sample of $n(\text{Tweets}) = 2,940$ Tweets, mentioning at least one of the three parties of interest. The codebook for extracting the important information from the tweets was defined. While developing the codebook, an emphasis was placed on Tweets predicting the voting tendency and background of users to examine the agenda used during the election. Below is a list of the variables with their description that was used for annotation (Table 3).

Obj_name is used to determine if a statement contains the politician's name or the politician's related content from the CDU/CSU, SPD, and the Alliance 90/The Greens.

Obj_party is coded in reference the objective party mentioned in the statement e.g., CDU/CSU as 1, SPD as 2 and Alliance 90/The Greens as 3.

Speech_act is an open category which is determined based on the text of tweets. Basically, it is a token (word) or, phrase, or hashtag based on which a speech act relevant to a party was identified.

Table 3 Summary of different features used in the study

Feature	Description	Type
Annotated Features		
Obj_name	To highlight a politician's name or the politician's office	Textual
Obj_party	Referenced to a political party	Categorical
Speech_act	open category for signal words representing statements	Textual
Speech_type	The speech type shows the speech acts in nature	Categorical
Praed	To identify how the party referenced in the statement is evaluated	Categorical
Rel_state	political party to which the speech acts refer	Categorical
Role_comp	describes whether the Twitter account is a competitor or not	Categorical
Role_spec	Twitter account belongs to an expert in politics	Binary
Role_org	Twitter account belongs to an organization	Binary
Data Enrichment		
<i>Social Features</i>		
Verified Account	Verified accounts provided by Twitter(before paid verification)	Binary
Account Age	Time difference(in days) of account creation date and the day before election	Numeric
Gender	Gender of Twitter handle	Binary
<i>Text Features</i>		
Hashtags & Mentions	Hashtags or keywords mentioned in Tweets	Textual
Sentiment	Sentiment of Tweets	Categorical
Hate speech	If some hate word is mentioned in tweet	Binary
<i>Network Features</i>		
Likes	Number of likes of Tweet	Numerical
Retweet	Number of retweets of Tweet	Numerical
Friends & Followers	Number of friends and followers of account	Numerical
Popularity	Ratio of follower and following of user	Binary

Speech_type shows the speech acts in nature. It has four categories: 1: evaluation, 2: persuasion attempt, 3: forecast of results and 4: reporting results.

Praed is used to identify how the political party referenced in the statement is evaluated. We categorised them into three categories 1: negatively(Call not to vote for the party; Forecast election defeat; Election defeat reported); 2: positively (Call to vote the party; Forecast election win; Election win reported); 0: not clear

Rel_state references the political party to which the speech act contains, such as 1: CDU/CSU, 2: SPD; 3: Alliance 90/The Greens; and 4: others.

Role_comp describes whether the Twitter account is a competitor or not; one category for each political party was defined: 1: Other role: Not a competitor; 2: Competitor: CDU/CSU; 3: Competitor: SPD; 4: Competitor: Alliance 90/The Greens.

Role_spec denotes if the Twitter account belongs to an expert in politics as follows: 1: Other roles: Not an expert; 2: Expert

Role_org denotes if the Twitter account belongs to an organization and which kind of organization. 1: Organisation connection to media; 2: Organisation connection to APO/social and civil society or movements; 3: Organisation connection to economy; 4: Organisation connection to trade unions, welfare organisations, and consumer protection; 5: Politically interested without organization connection

The coding process was completed by three students with master's degree who have a background in political science. Each annotator followed the above-discussed codebook to annotate variables independently. The reliability of the annotators was measured, resulting in Cohen's kappa of 0.776, which signals an acceptable to good intercoder reliability. After having the manual coding results analyzed and

checked by a senior research associate not directly involved in the annotation for correctness and objectivity to reduce biases, coding was done for 2,940 Tweets, and only 1,450 Tweets were found relevant based on our codebook. From 1,450 Tweets, $n(\text{statements}) = 2,627$ withstood data annotation criteria, consisting of $n(\text{statements, CDU/CSU}) = 1,011$ statements on the CDU/CSU, $n(\text{statements, SPD}) = 814$ statements on the SPD, $n(\text{statements, Alliance 90/The Greens}) = 802$ statements on Alliance 90/The Greens. A distribution of annotated Tweets for each political party used is shown in Fig. 2. Some examples of annotated `speech_act(statements)` are shown in Table 4 and overall descriptive information in Table 5.

3.1.3 Data Enrichment

This section provides details about the data analysis conducted to obtain more information about the collected Tweets. After the manual coding of the Tweets, an automatic analysis was performed to fetch additional features using Python code. These features were categorized into social, textual, and network-based features, as described below.

Social Features Social features are information extracted from Twitter handles that are related to accounts, which are explained as follows:

Verified Account Before account verification was purchasable (November 2022), the verified account badge was one of the most important criteria to indicate an authentic and official Twitter handle. Usually, these accounts belong to individuals in the public eye, such as politicians and journalists. In the context of this research, if political or journalist accounts were verified, they were considered genuine expert accounts. Therefore, the verified status of each account was fetched using data collected with the Twitter API.

Account Age Account age is another feature for the credibility of the account (Shahi et al., 2021). Usually, older accounts are considered to be genuine because they are not

Table 5 Descriptive analysis of annotated Tweets and statement

Parameter	Value
Tweets	
Number of Tweets	1,450
Unique Account	1,093
Verified Account	44
Popularity of Account	398
Mean Retweet Count	2.05
Mean Favourite Count	8.32
Median Followers Count	259
Median Friends Count	352
Median Account Age (days)	1,530
Unique Hashtags	38
Unique mentions	38
Unique Emoji	7
Gender(Male/Female/Unknown)	422/130/541
Statements	
CDU/CSU	1011
SPD	814
Alliance 90/The Greens	802
Sentiment(positive/negative/neutral)	270/1,799/558

removed by the Twitter bot detection tool or reported as spam and are not created to spread information about elections or events. The account creation dates were fetched through the Twitter API to calculate the account age. The account age was calculated as the difference between the cutoff and account creation dates. The cutoff date was the day before the voting, that is May 25, 2019.

Gender Each person has a different opinion regarding political issues or parties; it might vary with gender. Therefore, the gender of Tweets using a name-based list were identified, and the Tweets were analyzed (Mejova & Suarez-Lledó, 2020). A dictionary-based list (Mejova &

Table 4 Examples of annotated statements of major political parties

Political Party	Example	English Translation	Voting Behaviour
CDU/CSU	als gesamtpacket entspricht mir zur zeit als grün konservativer die union am meisten	As a whole, as a green-conservative, the union suits me the most at the moment	positive
CDU/CSU	nie wieder cdu/ csu	never again cdu/ csu	negative
SPD	jetzt am sonntag spd wählen	vote now on sunday spd	positive
SPD	spd ist nicht die antwort	spd is not the answer	negative
Alliance 90/The Greens	die grünen überzeugen fast überall	the greens are convincing almost everywhere	positive
Alliance 90/The Greens	sperrt die grünen wegen gezielten und falschen wahlkampf - fehlerhauptungen	block die grünen for targeted and false election campaign - misstatements	negative

Suarez-Lledó, 2020) was used to get the gender of the author as male, female or unknown. We matched the name of the Twitter handle with a list to identify their gender.

Text Features Tweets were processed based on different parameters as part of the content analysis to look into them intensely, resulting in the identification of different features from them, as discussed below.

Hashtags and Mentions Using hashtags on Twitter is a trend that involves joining a group of people who discuss particular issues. The hashtags used in the Tweets of each political party were fetched. Users also utilize hashtags for particular agendas to spread information, so grouping the hashtags together helps figure out the political agenda. In comparison, mentions are used to include particular users while posting Tweets. Usually, politicians or political parties are tagged in Tweets. The hashtags mentioned in the Tweets were fetched and grouped together for each political party. Regex-based Python code was used to find the hashtags and mentions from the speech act and count their occurrence for textual features, hashtags, and mentions.

Sentiment Analysis The sentiment of the Tweets is an important factor in deciding the polarity of the message. Often, political parties share messages against their opponent, which might have a negative sentiment, so in the present study, the sentiment of each Tweet was identified. A Bidirectional Encoder Representations from Transformers (BERT)-based approach was used for the German language to analyze the sentence (Guhr et al., 2020).

Hate speech Often, politicians or users use hate words for the opposition; a previous study presents the role of hate speech in the election discourse (Jacobs & Van Spanje, 2020; Nandini & Schmid, 2022; Shahi & Kana Tsoplefack, 2022a). Following this, hate words were identified from the German hate speech dataset (Bassignana et al., 2018; Shahi & Majchrzak, 2024) and fetched in the annotated corpus using text matching. The presence of hate words shows the style of speech used by users for a political party during the European election.

Network-Based Feature Network-based features are generated because of the interaction of Tweets and users within a network. This section describes the network-based features to understand the diffusion and user reaction of Tweets.

Likes The favorite count of each Tweet was included because of likes indicating that the user agrees with the content. Therefore, Tweets with more likes are more popular and reach more users (Stieglitz & Dang-Xuan, 2013).

Retweet Retweet refers to sharing a Tweet, which indicates the diffusion of posts among other users. More retweets mean it reaches more users on their timelines in the network (Stieglitz & Dang-Xuan, 2013), which is why retweets were identified and used in the model of the present study.

Friends and Followers Twitter shares the information of the number of accounts as well as those who follow an account, which is indicated as a follower. Simultaneously, if the same user follows the following user back, they are considered friends. Then, each posted Tweet is displayed on all their followers' and friends' networks, increasing the visibility of Tweets.

Popularity of Account The popularity of an account is another feature discussed by Shahi et al. (2021). The Tweets posted by popular accounts gather more attention and diffuse faster. So, if a popular account shares a message, it reaches a larger crowd. Account popularity was computed as the ratio of followers and follows; if there were more followers than follows, then the account was considered a popular account.

3.2 Data Cleaning and Preprocessing

After the data annotation and feature extraction, information from Tweets that were irrelevant to our purpose was removed. This included removing hyperlinks, special characters, punctuation marks, and special characters mentioned in the Tweets. The Python package Natural Language Toolkit (NLTK) (Loper & Bird, 2002) was used for data cleaning and removing unwanted features in preprocessing. From the cleaned Tweets, we used the extracted annotated features mentioned in Table 3 as a reference. The cleaned text of Tweets were converted into sentence embeddings for the regression model. The sentence embeddings were created using the XLM-BERT sentence model (Kazemi et al., 2022). After cleaning the data, we received 1,011 Tweets for CDU/CSU, 814 Tweets for the SPD, and 802 Tweets for die grünen party.

3.3 Opinion Formation

The opinions of experts and users with political stands or preferences (such as political candidates) were analyzed; categories are mutually exclusive. During data annotation, the role of the Twitter handle was identified, for example, whether the Tweet was coming from an expert(role_spec) with a background organization (role_org), such as from the category economy or users with political affiliations (role_comp). The agenda formation was analyzed by looking into role_spec(experts) and role_comp(user with political affiliations), and the differences were compared. The overall data were as follows: 478 speech_acts from 256 Tweets posted by 208 expert authors and 237 speech_acts from 169 Tweets posted by 118 regular users. Because of the small sample size, methods such as word clouds or topic modeling were not applicable; hence, word frequency was used to analyze the agenda formation. The speech_act posted by an

expert was tokenized, and the word frequency was counted. There are 1,199 and 737 unique tokens for `role_spec` and `role_comp`, respectively. The words related to the political party were manually analyzed. From the frequent words, topics discussed by `role_spec` and `role_comp` for each political party were identified. Along with frequent words, the profiles of `role_spec` and `role_comp` and the reactions of users to Tweets while forming an agenda were compared. The `speech_type` of Tweets posted by `role_spec` and `role_comp` regarding evaluation, persuasion attempt, and forecast of the results for each political party were analyzed.

3.4 Prediction Model

The next step was to build a prediction model using annotated data and features extracted in Section 3.1. After data were processed and additional features were identified, we used a regression model that predicts voting behaviour. In the present work, different regression models were used, including linear regression, stochastic gradient descent regression, logistic regression, support vector regression (RBF kernel, linear kernel), Bayesian regression, automatic relevance determination (ARD) regression, multinomial NB regression, Adaboost regression, and random forest regression. For the prediction of the voting behavior, different regression models were implemented by using the scikit-learn machine learning (Pedregosa et al., 2011). Finally, the results and interpretations of the features were compared for the different models.

3.4.1 Error Analysis

The prediction model faces criticism regarding its error rate (Nandini & Schmid, 2022). In the implementation, there is a possibility that the predicted results are not uniform across all political parties. Therefore, we manually analyzed the results obtained from the prediction model. The goal of the error analysis is to find the prediction error responsible for the incorrect prediction of voting tendencies. We sampled 100 annotated data `speech_act` entries equally distributed for each political party. We identified that 30% were predicted accurately, and around 64 % were predicted with an error of less than 10%, and the remaining had an average error of around 20%, which is equivalent to the average error of the best prediction model. The errors were mainly due to short `speech_acts` such as `#zurueckzudengruenen(#backtothegreen)`, `grüne 18%(green 18%)` or `speech_acts` that did not talk explicitly about any political party for instance, `ihr habt kein profil mehr und keine glaubwürdigkeit(You no longer have a profile and no credibility)`. So, with the longer `speech_act` referring to the political par-

ties, the prediction model can predict the voting tendency of the users better.

3.4.2 Feature Analysis

Once we had result from the prediction model, we performed feature analysis, looking deeper into the role of features in predicting the voting tendencies. We looked at the coefficient of variables in the prediction model to determine the feature importance. We presented a bar chart and coefficient to represent feature importance and then provided an analysis to identify the role of different features in the prediction model.

4 Research Results

This section describes the implementation and results regarding the agenda-setting theory, predictions, and prediction models with its error rates objectively.

Agenda Formation First, `role_spec` (whether the account is an expert) and `role_comp` (whether the account is a competitor) accounts, which set the political agenda in the public discourse, were analyzed. More verified accounts in `role_comp` than `role_spec` were identified; altogether, there were 33 verified accounts out of 44 in the collected datasets. There was no clear difference regarding followers and friends; however, for some, `role_spec` had more followers while `role_comp` had more friends. Therefore, some experts had a huge user base to spread their opinions. Gender distribution and popularity in both categories were similar.

Agenda formation was analyzed for each political party; both `role_spec` and `role_comp` keep their agenda specific to a political party. Regarding content, `role_spec` posts 69% of Tweets with negative sentiment compared to 44% posted by `role_comp`.

Both `role_spec` and `role_comp` raise a similar agenda in the discourse; however, in the communication style, `role_comp` was too direct while `role_spec` discussed different forms of the same issue. For CDU/CSU, experts discussed climate change and poverty, for the SPD, experts talked about the economy and the environment, for the Alliance 90/The Greens, experts talked about humanity and green energy. All three political parties discussed 'poverty', 'pensions', 'environmental destruction', and 'refugees'. Apart from the agenda formation, `role_spec` urged users to vote in the European Parliament election 2019, reminding them of the voting date, i.e., May 26. In terms of `speech_type`, both `role_spec` and `role_comp` discussed evaluation, followed by a persuasion attempt, but `role_comp` formed an agenda about their party winning to motivate the users to vote for them.

The agenda set by both *role_spec* and *role_comp* for Alliance 90/The Greens successfully got more user reactions in terms of likes and retweets; these Tweets got around double the retweets compared with the average retweets, so users were willing to share their posts and get involved in spreading the agenda to their followers in the network. For the CDU/CSU, the likes were less than average with few retweets, while the SPD got around half the likes of the average likes and few retweets. This showed that users believed in the topic discussed by *role_spec* and *role_comp* for the Alliance 90/The Greens compared with the CDU/CSU and SPD. The examples of *speech_act* discussed by each political party for agenda formation are presented in Table 6. Tweets related to each political party contained hate speech, while discourse involving the Alliance 90/The Greens used the most hate words, followed by CDU/CUS and SPD. The commonly used hate words in the dataset were '*verbrecher*' (criminal), '*betrüger*' (fraudster), '*dumm*' (stupid), '*müll*' (trash), '*narzissmus*' (narcissism), '*quatsch*' (nonsense), '*ungebildete*' (uneducated).

Prediction Model For the prediction model, the features were converted into vectors using sentence embedding. BERT-based embedding for the sentiment, called German sentiment (Guhr et al., 2020), was used.

For implementing the regression model, data were split for training and testing in a ratio of 70 and 30, respectively. The numerical features were normalized using min-max normalization. *Speech_act*, hashtags, and mentions were converted into vectors using sentence embedding model (Reimers & Gurevych, 2019; Kazemi et al., 2022). Finally, all features were merged for the prediction model. The mean absolute error (MAE) of the different regression models is shown in Table 7. We explain the implementation of the best regression model, that is logistic regression. Fivefold cross-validation was used for training and grid search for the hyperparameter optimization. For the best-performing model, that is logistic regression, the mean absolute error was around 18%. The best hyperparameters used for the training was *bootstrap: True*, *max_depth: 80*, *max_features: 'sqrt'*, *n_estimators: 300*. The feature relevance for the model performance was also computed using the scikit-learn library; the most important features were *role_comp*(0.24), *speech_type* (0.20), *check_union* (0.17), *gender*(0.13) *check_spd*(0.10), *role_spec* (0.09), *check_grün* (0.08), *user popularity* (0.06), and *account age*(0.07). The important features of the best regression model are shown in Fig. 3. The results show that the model performs well by combining all three feature categories-textual, social features, and network features, that is text feature (*speech_type* and *role_comp*) and social features (account age), and network-based feature (user popularity) play a significant role in the

prediction model. In comparison, some of the features, such as sentiment, likes, and retweets, are less important.

5 Discussion

The present study shows the research conducted to predict voting tendencies of Twitter users during the 2019 European Parliament Election in Germany. In terms of writing, few emojis and hashtags were used in the Tweets, showing users were relatively neutral and were not following a common discussion on Twitter, however, some Tweets used hateful words in the discourse. Tweets were used for the prediction of users' voting tendencies by utilising numerous features for the prediction model. Several of them were identified through the previous research of predictive analytics as shown in Table 1. The Tweets were further split into *speech_acts* and annotated to identify the voting tendency of a user; afterwards, features such as network and social features were extracted to train the prediction model. The best prediction result is obtained using logistic regression. The Observations showed that the combination of feature categories works well for the prediction model, which goes against some of the previous work (Maldonado & Sierra, 2015; Nofer & Hinz, 2015; Budiharto & Meiliana, 2018; Kristiyanti et al., 2019) that simply used a sentiment-based prediction model.

The actual results of the 2019 European Parliament Election in Germany⁴ indicate that the prediction of the voting tendency of users also lean towards the three major parties focused on in this research. The results in Fig. 3 show that the most important variables acquired from social media for the prediction model were *role_comp* and *speech_type* meaning, whether a competitor made a statement and what type of speech was used, for example, it a persuasion attempt or a reporting of results. This indicates that these variables are not only the most important ones when it comes to predicting the voting behavior of the public but also for setting the political agenda. The belief systems in virtual communities can be polarized by experts in future elections in accordance with the agenda-setting theory as they are most indicative in determining how the audience might vote. Furthermore, these results are in support of the findings by Bogaert et al. (2021, 2016) and consistent with proposition 1. They indicate that the inclusion of additional data that is not directly available on social media platforms through means such as their API leads to the improvement of prediction results. These variables extend the previous research shown in Table 1 by building on multiple variables.

Social features played an important role in the performance of prediction models. Social features such as account

⁴ <https://www.bundeswahlleiterin.de/en/europawahlen/2019/ergebnisse/bund-99.html>

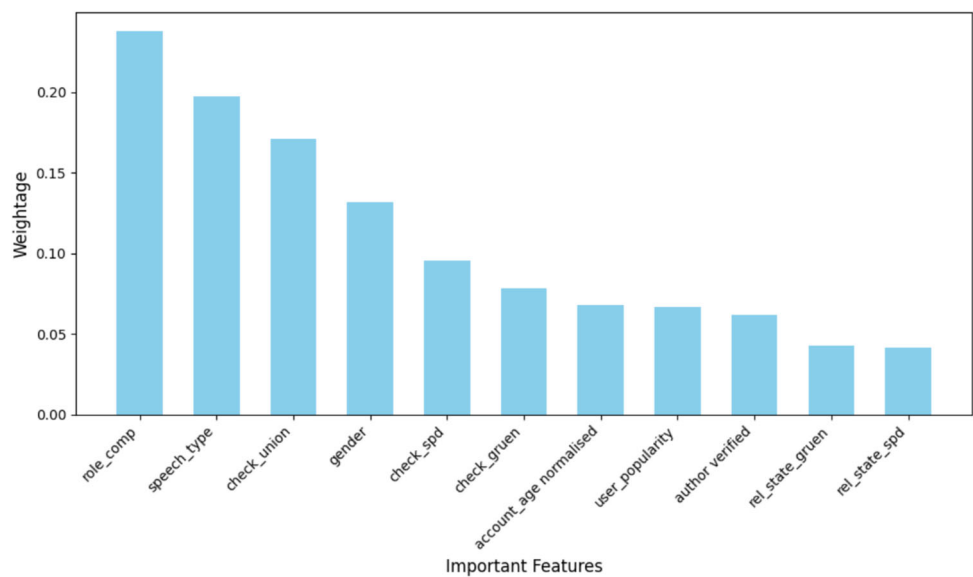
Table 6 Examples of speech_act for setting an agenda of different political parties by experts (Role_spec) and competitors (Role_comp)

Political party	Speech_act	English Translation
Role_spec		
CDU/CSU	“klimaschutz humanität transparenz vertrauen”; “wäre wenn jetzt hinsichtlich klimas ihre arbeit machen würden statt sich über quellenangaben versehens video aufzuregen”	“climate protection humanity transparency trust”; “would be if they would now do their work with regard to climate instead of getting upset about the video provided with source information”
SPD	“politik machen gegen eigene volk hungerlöhnen armuts renten hart armut tafeln suppenküchen flaschen müll suchen überleben”;“wahlprogramm überzeugt mich nicht prozent umfragewerte europawahl definitiv viele viele andere wähler auch nicht”	“making politics against one’s own people starvation wages poverty pensions hard poverty eating soup kitchens bottles garbage looking for survival”;“election program doesn’t convince me percent poll values European elections definitely many many other voters don’t either”
Alliance 90/The Greens	“verteidigung menschenrechte”;“sind in anderen ländern im bezug auf kernenergie deutlich aufgeklärter”	“defense of human rights”;“are much more enlightened in other countries with regard to nuclear energy”
Role_comp		
CDU/CSU	“werden leider immer gewählt” ; “gehe jetzt meine stimme abgeben damit manfred weber kommissionpräsident wird”	“are unfortunately always chosen” ; “I’m going to cast my vote now so that Manfred Weber becomes Commission President”
SPD	“franz timmermans wählen” ; “teambarley”	“choose frans timmermans” ; “team barley”
Alliance 90/The Greens	“votegreen2019” ; “europawahl2019 darum grün”	“vote green 2019” ; “european election 2019 date green”

Table 7 Result of Different Regression models for prediction of voting tendency(Selected results are reported)

Regression Model	Error(MAE)
Linear Regression	0.25
Support Vector Regression (SVR)	0.24
Random Forest	0.20
Automatic Relevance Determination Regression (ARD)	0.24
AdaBoost	0.28
Logistic (Text)	0.22
Logistic (Text+Social)	0.21
Logistic (Text+Social+Network)	0.18

Fig. 3 Feature importance in the prediction model



age and gender were important criteria; on average, accounts were four years old and quite confident in their statements. There were only 44 accounts verified by Twitter (based on their old policy).

With the error analysis, we identified that the errors were mainly due to short or irrelevant no speech_acts, which do not include much information about the user's opinion on political parties. The key takeaway is that, with the longer and more relevant speech_acts, we can predict the voting tendency efficiently. In the future, for data annotation, we have to choose meaningful and longer speech_acts that present the user's tendency to vote for a political party.

The findings of (Colomer & Llavador, 2012) using agenda-setting theory suggested that, to get support from voters or polarise belief systems or virtual communities, a candidate or party should identify important topics for the public and give them more visibility, which can result in even more effective agenda-setting when extended with the findings of the present study. Our results shows the importance of different forms of actors, such as based on role_spec and role_comp, the candidates could leverage their position with different types of speech and form of actors, such as an appeal to vote.

In this regard, the findings through the speech_act shown in Table 6 can be used as an indication of how certain topics related to the manifestos of political parties, such as 'Klimaschutz' (Climate protection) for the *CDU/CSU*, 'Kernenergie' (Nuclear energy) for the *Alliance 90/The Greens* and 'Armut' (Poverty) for the *SPD*, can and should be framed within agenda-setting theory to set the policy agenda to win as many votes as possible. By identifying the sentiment of the conversation around these topics, the parties can position themselves and their party's stance on the key issues included in the manifesto of the party, hence aligning with the policy agenda-setting theory in a favourable light to gain more votes. Therefore, by combining these key talking points with the different variables, such as role_comp and role_spec, the policy agenda can be set more effectively and result in a more favorable public opinion towards a certain candidate or party; ultimately, this can lead to a win for the candidate or party that is setting the agenda leveraging our results in alignment with proposition 2. Moreover, these findings help address the theoretically established research gap by proposing that these novel variables are related to the public's voting tendencies and therefore, valuable in analyzing the political agenda per agenda-setting theory. Further echo chambers can be utilized to strengthen the position of a party and its candidate by adopting a favorable opinion towards that party or candidate and spreading it to their followers and other individuals susceptible to that opinion. Moreover, spreading process could be further advanced by the freedom

to express one's opinion anonymously, which is proposed as one of the principles of the Bright Internet and is provided by several social media platforms through their anonymous usernames (Lee et al., 2018).

Apart from the prediction model, in the Twitter discourse, information diffusion can be measured by the number of retweets and likes, which shows the engagement of users with others content on the platform (Stieglitz & Dang-Xuan, 2013). In the collected data, there were fewer retweets and likes that showed the rate of diffusion of those Tweets that was slow and was hardly shared over the platform. However, for the agenda formation, the diffusion of Tweets for the *Alliance 90/The Greens* had more than the average number of retweets of the dataset, indicating the relevance of several variables from previous prediction research, which is shown in Table 1 for agenda formation.

Information Systems research involving a data-driven approach is still in its early stages (Kar & Dwivedi, 2020). The present study contains several building blocks for theoretical development; we provide an extensive description of data collection, sampling, annotation, data enrichment, and quality checks. With the prediction model, we also explain the role of different features in calculating the results. Overall, user-generated Tweets helped predict election results that matched the actual outcome of the election. Our prediction models show the importance of annotated data in predicting election results, mainly the value of role_comp, speech_type, and speech_act. Tweets including different speech_types, posted by competitors (role_comp) to for example convince users to vote (persuasion attempt) or share the possibility of a winner (forecast of results) improve prediction models. Our findings support the proposed propositions. Therefore, the data-driven research approach discussed in the study can further be used to investigate theory-building for similar elections by proposing and testing strong hypotheses (Kar et al., 2023; Miranda et al., 2022) based on user-generated content. The proposed approach can accommodate data from different platforms and elections worldwide.

For practical implementation, an approach based on agenda-setting theory can be used to monitor the formation of political agendas among online users; it can be useful to break the negative agenda formation of a group of users to conduct a fair election Monitoring agenda formation also allows a political party to gain political advantages in terms of gaining votes. The positive side of monitoring agenda formation is to motivate users to vote or highlight the need for action to address sensitive issues like human rights or climate change. Another implication could be that, post-election, the proposed approach and findings can be used to predict the voting tendency of users and used to analyse and to predict the winning party instead of regular opinion polls.

6 Conclusion and Future Work

In our research, we used a Twitter dataset based on communication during the 2019 European Parliament election, with this we are able to build a prediction model utilizing different features acquired from Tweets and users within that dataset. The dataset used in the current research to train the prediction model to forecast one of three parties as the winner of the 2019 European Election in Germany, this was tracked with keywords related to the three parties. We used agenda-setting theory as our theoretical framework to analyze the formation of the political agenda and polarisation of belief systems and virtual communities during the discourse around the election. Agenda formation and the role of which additional features from Tweets improve the performance of the prediction model on the 2019 European Parliament Election in Germany were studied. The agenda formation was established mostly by `role_comp` and `role_speech` for each political party. The rate of diffusion of Tweets for Alliance 90/The Greens was higher compared with those for the CDU/CSU and SPD.

Additional features were extracted from Tweets, such as account details, network-based features, text features, manually coded Tweets, and user account details. A machine learning model was trained based on these features for a better prediction model. The findings implied that combining additional features for predictions improves the prediction model's performance. Further, it was identified that variables such as `role_comp` and `speech_type` play an important role in improving the prediction model, successfully forming a political agenda for spreading political manifestos and requesting votes for their party. The spreading process could be enhanced by the freedom to express one's opinion anonymously, as proposed as one of the principles of the Bright Internet through the anonymity provided by social media platforms through, for example, usernames (Lee et al., 2018).

The practical contributions of the current research are that it introduces a prediction model with better performance compared with regular approaches that can be utilized by different actors, such as politicians or news agencies, to more accurately predict the outcome of elections that are discussed on social media platforms. In addition, these findings indicate that the type of speech and whether the individual who is sharing a message related to the event is a candidate are the most suitable variables in predicting the public's voting tendencies. This can be leveraged in future elections or events to shift the public's voting tendency in one party's favour and address echo chambers that potentially strengthen the position of a party and its candidate. Furthermore, the present research contributes to research by identifying additional features from data acquired from Twitter that, when used to train a prediction model and predict the voting tendency, improve the model and the prediction results. Moreover, the results

indicate the best features available through Twitter to predict the voting tendency of users during an election and to identify and set the political agenda per agenda-setting theory.

In future research, the approach can be extended to data from other platforms, such as Reddit and YouTube, to analyze voting tendencies. In addition, with the feature importance, an explainable model (Nandini & Schmid, 2022) can be applied to explain the prediction model. Furthermore, as a possible extension, the prediction model or agenda formation can be implemented for the analysis of the 2024 European Parliament Election.

Acknowledgements We acknowledge the role of previous colleagues in suggesting hashtags for data collection. We also recognize the role of data annotators who put effort into understanding the codebook and annotating the data.

Author Contributions Four authors prepared the manuscript; out of four authors, the first two are doctoral candidates. Stefan Stieglitz is a professor and PhD supervisor of Gautam Kishore Shahi & Ali Sercan Basyurt, and Christoph Neuberger is the professor and project partner. Gautam Kishore Shahi led the project, and the author was responsible for planning, idea formulation, and drafting the manuscript with a focus on introducing data collection, data analysis, implementing a machine learning model, results, error analysis, discussion, and conclusion. Ali Sercan Basyurt was also involved in ideas formulation, helping in the first draft of the paper, mainly the introduction, related work, theoretical foundation, discussion and conclusion, and future work. Stefan Stieglitz supervised the overall process, provided feedback, and proofread, mainly on data collection and machine learning models. Christoph Neuberger was responsible for supervising the overall process, providing feedback, and, mainly, on codebook development and data annotations.

Funding Open Access funding enabled and organized by Projekt DEAL. The study was carried out under the University of Duisburg-Essen, Germany, and Freie Universität Berlin with grant ID 210149318 funded by the Deutsche Forschungsgemeinschaft (DFG) to study transnational events in social media and result in predictions.

Availability of Data and Materials Once the manuscript is accepted, following the data-sharing policies of Twitter, we will share the Tweet IDs that are used in the study. We will also release the code and script, which can be used to replicate the study on other social media platforms. We will also share the codebook used in the data annotation so that it can be used in further research.

Declarations

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.

Ethics Approval and Consent to Participate While doing the research, we followed the ethical guidelines of the University of Duisburg-Essen, Germany for data collection, processing, and analysis. Overall, we have focused on the following points-

- While collecting the data, we followed the data collection policy of Twitter (now X).
- We collected publicly available data like profile information and Tweets; it is only used for research purposes, without looking at personal details or any commercialization.

- For the analysis, we have not provided any preference for data from any political party.
- The data is not considered to be sensitive or confidential in nature;
- Vulnerable or dependent groups are not included;
- Following the data-sharing policy of Twitter (now X), we will only share the Tweet ID and our method for the replication to other events or platforms.

Consent for Publication All co-authors have approved the content of the manuscript. All authors have given explicit consent to publish this manuscript. The work described in this manuscript (approximately 9246 words) is original work and prepared for submission to the special issue of the Information Systems Frontiers journal. This manuscript is not under consideration for publication anywhere else.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Allam, Z., & Dhunny, Z. A. (2019). On big data, artificial intelligence and smart cities. *Cities*, 89, 80–91.
- Andersen, P. H., Christensen, P. R., & Damgaard, T. (2009). Diverging expectations in buyer-seller relationships: Institutional contexts and relationship norms. *Industrial Marketing Management*, 38(7), 814–824.
- Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (vol. 1, pp. 492–499). IEEE
- Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 519–528)
- Baran, S. J. (2015). *Mass communication theory: Foundations, ferment, and future*.
- Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of twitter users. *Social Science Computer Review*, 33(6), 712–729.
- Barberá, P., & Zeitzoff, T. (2018). The new public address system: Why do world leaders adopt social media? *International Studies Quarterly*, 62(1), 121–130.
- Bassignana, E., Basile, V., Patti, V., et al. (2018). Hurltlex: A multilingual lexicon of words to hurt. In *CEUR Workshop Proceedings* (vol. 2253, pp. 1–6). CEUR-WS
- Benthaus, J., & Skodda, C. (2015). *Investigating consumer information search behavior and consumer emotions to improve sales forecasting*.
- Bermingham, A., & Smeaton, A. (2011). On using twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP 2011)* (pp. 2–10)
- Bogaert, M., Ballings, M., & Poel, D. (2016). The added value of facebook friends data in event attendance prediction. *Decision Support Systems*, 82, 26–34.
- Bogaert, M., Ballings, M., Poel, D., & Oztekin, A. (2021). Box office sales and social media: A cross-platform comparison of predictive ability and mechanisms. *Decision Support Systems*, 147, 113517.
- Bruns, A., & Burgess, J. (2015). Twitter hashtags from ad hoc to calculated publics. *Hashtag publics: The power and politics of discursive networks* (pp. 13–28)
- Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of indonesia presidential election from twitter using sentiment analysis. *Journal of Big Data*, 5(1), 1–10.
- Burbach, L., Belavadi, P., Halbach, P., Plettenberg, N., Nakayama, J., Ziefle, M., & Valdez, A. C. (2019). Towards an understanding of opinion formation on the internet: Using a latent process model to understand the spread of information on social media. In *Conference of the European Social Simulation Association* (pp. 133–145). Springer
- Chadwick, A. (2017). *The Hybrid Media System: Politics and Power*. Oxford University Press
- Chauhan, P., Sharma, N., & Sikka, G. (2021). The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, 12(2), 2601–2627.
- Chen, M.-C., Chiu, A.-L., & Chang, H.-H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4), 773–781.
- Cohen, B. C. (2015). *Press and foreign policy* (p. 2321). Princeton university press
- Colomer, J. M., & Llavador, H. (2012). An agenda-setting model of electoral competition. *SERIEs*, 3(1), 73–93.
- Dearing, J., & Rogers, E. (1996). *Agenda setting (communication concepts)*. Thousand Oaks: Sage.
- DeVito, M. A. (2017). From editors to algorithms: A values-based approach to understanding story selection in the Facebook news feed. *Digital Journalism*, 5(6), 753–773.
- Diaz Ferreyra, N. E., Shahi, G. K., Tony, C., Stieglitz, S., & Scandariato, R. (2023). Regret, delete,(do not) repeat: An analysis of self-cleaning practices on twitter after the outbreak of the covid-19 pandemic. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–7)
- Feng, N., Shi, Y., Li, Y., Li, D., Zhang, J., & Li, M. (2024). An exploration of the dynamics between social media and box office performance. *Information Systems Frontiers*, 26(2), 591–608.
- Galpin, C., & Trenz, H.-J. (2019). In the shadow of Brexit: The 2019 European parliament elections as first-order polity elections? *The Political Quarterly*, 90(4), 664–671.
- Gilardi, F., Gessler, T., Kubli, M., & Müller, S. (2022). Social media and political agenda setting. *Political Communication*, 39(1), 39–60.
- Guhr, O., Schumann, A.-K., Bahrman, F., & Böhme, H. J. (2020). Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1627–1632)
- Hall, W., Tinati, R., & Jennings, W. (2018). From Brexit to trump: Social media's role in democracy. *Computer*, 51(1), 18–27.
- Himmelroos, S., & Schoultz, Å. (2023). The mobilizing effects of political media consumption among external voters. *European Political Science*, 22(1), 44–62.
- Jacobs, L., & Van Spanje, J. (2020). Prosecuted, yet popular? Hate speech prosecution of anti-immigration politicians in the news and electoral support. *Comparative European Politics*, 18, 899–924.
- Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2019). Predicting elections from social media: a three-country, three-method comparative study. *Asian Journal of Communication*, 29(3), 252–273.
- Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A

- response to tumasjan, a., sprenger, to, sander, pg, & welp, im “predicting elections with twitter: What 140 characters reveal about political sentiment”. *Social Science Computer Review*, 30(2), 229–234.
- Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research-moving away from the “what” towards the “why.” *International Journal of Information Management*, 54, 102205.
- Kar, A. K., Angelopoulos, S., & Rao, H. R. (2023). *Guest Editorial: Big data-driven theory building: Philosophies, guiding principles, and common traps*. *International Journal of Information Management*, 71, 102661.
- Kazemi, A., Garimella, K., Shahi, G. K., Gaffney, D., & Hale, S. A. (2022). *Research note: Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 Indian general election on WhatsApp*. Harvard Kennedy School Misinformation Review.
- Kim, D. H., & Ellison, N. B. (2022). From observation on social media to offline political participation: The social media affordances approach. *New Media & Society*, 24(12), 2614–2634.
- Kim, J., Hwang, S., & Park, E. (2021). Can we predict the Oscar winner? a machine learning approach with social network services. *Entertainment Computing*, 39, 100441.
- Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniec, J., Gruza, M., Janz, A., Kanclerz, K., et al. (2023). Chatgpt: Jack of all trades, master of none. *Information Fusion*, 101861.
- Kristiyanti, D. A., Umam, A. H., et al. (2019). Prediction of indonesia presidential election results for the 2019-2024 period using twitter sentiment analysis. In *2019 5th International Conference on New Media Studies (CONMEDIA)* (pp. 36–42). IEEE
- Kumar, P., Kushwaha, A. K., Kar, A. K., Dwivedi, Y. K., & Rana, N. P. (2022). Managing buyer experience in a buyer-supplier relationship in msms and smes. *Annals of Operations Research*, 1–28.
- Kumpulainen, I., Praks, E., Korhonen, T., Ni, A., Rissanen, V., & Vankka, J. (2020). Predicting eurovision song contest results using sentiment analysis. In *Conference on Artificial Intelligence and Natural Language* (pp. 87–108). Springer
- Kushwaha, A. K., & Kar, A. K. (2021). Markbot-a language model-driven chatbot for interactive marketing in post-modern world. *Information Systems Frontiers*, 1–18.
- Kushwaha, A. K., Kar, A. K., & Dwivedi, Y. K. (2021). Applications of big data in emerging management disciplines: A literature review using text mining. *International Journal of Information Management Data Insights*, 1(2), 100017.
- Kushwaha, A. K., Kar, A. K., Roy, S. K., & Ilavarasan, P. V. (2022). Capricious opinions: A study of polarization of social media groups. *Government Information Quarterly*, 39(3), 101709.
- Lee, J. K., Cho, D., & Lim, G. G. (2018). Design and validation of the bright internet. *Journal of the Association for Information Systems*, 19(2), 3.
- Loper, E., & Bird, S. (2002). Nltk: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics-Volume 1* (pp. 63–70)
- Makazhanov, A., & Rafiei, D. (2013). Predicting political preference of twitter users. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 298–305)
- Maldonado, M., & Sierra, V. (2015). *Can social media predict voter intention in elections?_x000d_the case of the 2012 Dominican Republic presidential election*.
- Mejova, Y., & Suarez-Lledó, V. (2020). Impact of online health awareness campaign: case of national eating disorders association. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12* (pp. 192–205). Springer
- Mellon, J., & Prosser, C. (2017). Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4(3), 2053168017720008.
- Meraz, S. (2009). Is there an elite hold? traditional media to social media agenda setting influence in blog networks. *Journal of Computer-mediated Communication*, 14(3), 682–707.
- Miranda, S., Berente, N., Seidel, S., Safadi, H., & Burton-Jones, A. (2022). Editor’s comments: Computationally intensive theory construction: A primer for authors and reviewers. *MIS Quarterly*, 46(2),.
- Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11), 4366–4383.
- Nandini, D., & Schmid, U. (2022). Explaining hate speech classification with model-agnostic methods. In *Joint Proceedings of Workshops, Tutorials and Doctoral Consortium Co-located with the 45rd German Conference on Artificial Intelligence (KI 2022) Virtual Event, Trier, Germany, September 19-20*. <https://ceur-ws.org/Vol-3457/paper2tmg.pdf>
- Nann, S., Krauss, J., & Schoder, D. (2013). *Predictive analytics on public data-the case of stock markets*.
- Nawaz, A., Ali, T., Hafeez, Y., Rehman, S. U., & Rashid, M. R. (2022). Mining public opinion: a sentiment based forecasting for democratic elections of pakistan. *Spatial Information Research*, 1–13.
- Nguyen, T. H., & Rudra, K. (2024). Human vs ChatGPT: Effect of Data Annotation in Interpretable Crisis-Related Microblog Classification. In *Proceedings of the ACM on Web Conference 2024* (pp. 4534–4543)
- Nofer, M., & Hinz, O. (2015). Using twitter to predict the stock market. *Business & Information Systems Engineering*, 57(4), 229–242.
- Padmanabhan, B., Fang, X., Sahoo, N., & Burton-Jones, A. (2022). Machine learning in information systems research. *MIS Quarterly*, 46(1),.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Poornima, S., & Pushpalatha, M. (2016). A journey from big data towards prescriptive analytics. *ARPJ: Journal of Engineering and Applied Sciences*, 11(19), 11465–11474.
- Proksch, S.- O., & Slapin, J. B. (2015). *The Politics of Parliamentary Debate*. Cambridge University Press
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992)
- Russell, A. (2018). Us senators on twitter: Asymmetric party rhetoric in 140 characters. *American Politics Research*, 46(4), 695–723.
- Shahheidari, S., Dong, H., & Daud, M. N. R. B. (2013). Twitter sentiment mining: A multi domain analysis. In *2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems* (pp. 144–149). IEEE
- Shahi, G. K., & Kana Tsoplefack, W. (2022a). Mitigating harmful content on social media using an interactive user interface. In *International Conference on Social Informatics* (pp. 490–505). Springer
- Shahi, G. K., & Majchrzak, T. A. (2022b). Amused: an annotation framework of multimodal social media data. In *Intelligent Technologies and Applications: 4th International Conference, INTAP 2021, Grimstad, Norway, October 11–13, 2021, Revised Selected Papers* (pp. 287–299). Springer
- Shahi, G., & Majchrzak, T. (2024). Hate speech detection using cross-platform social media data in english and german language. In

Proceedings of the 20th International Conference on Web Information Systems and Technologies (pp. 131–140)

- Shahi, G. K., Clausen, S., & Stieglitz, S. (2022). Who shapes crisis communication on twitter? An analysis of German influencers during the COVID-19 pandemic. In: *55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event/Maui, Hawaii, USA, January 4-7, 2022* (pp. 1–10). ScholarSpace
- Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An exploratory study of covid-19 misinformation on twitter. *Online Social Networks and Media*, 22, 100104.
- Shi, L., Agarwal, N., Agrawal, A., Garg, R., & Spoelstra, J. (2012). *Predicting us primary elections with twitter* (p. 4). <http://snap.stanford.edu/social2012/papers/shi.pdf>
- Shi-Nash, A., & Hardoon, D. R. (2017). Data analytics and predictive analytics in the era of big data. *Internet of things and data analytics handbook* (pp. 329–345)
- Soroka, S. N. (2002). Issue attributes and agenda-setting by media, the public, and policymakers in Canada. *International Journal of Public Opinion Research*, 14(3), 264–285.
- Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29(4), 217–248.
- Stieglitz, S., Meske, C., Ross, B., & Mirbabaie, M. (2020). Going back in time to predict the future—the complex role of the data collection period in social media analytics. *Information Systems Frontiers*, 22(2), 395–409.
- Tariq, R., Zolkepli, I. A., & Ahmad, M. (2022). Political participation of young voters: Tracing direct and indirect effects of social media and political orientations. *Social Sciences*, 11(2), 81.
- Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media* (vol. 4, pp. 178–185)
- Vergeer, M. (2015). Twitter and political campaigning. *Sociology Compass*, 9(9), 745–760.
- Wiliarty, S. E. (2023). In Campbell, R., & Davidson-Schmich, L. K. (Eds.), *The CDU/CSU and the 2021 Federal Election* (pp. 81–100). Springer, Cham
- Wolfe, M. (2012). Putting on the brakes or pressing on the gas? media attention and the speed of policymaking. *Policy Studies Journal*, 40(1), 109–126.
- Zakir, J., Seymour, T., & Berg, K. (2015). Big data analytics. *Issues in Information Systems*, 16(2),.

Gautam Kishore Shahi is a PhD student at the University of Duisburg-Essen, Germany. His research interests are online harmful content, fact-checking, and Social Media Analytics. He has a background in computer science engineering with a focus on data science. His PhD is focused on the diffusion of harmful content on social media. Gautam received a master's degree from the University of Trento, Italy, and a bachelor's Degree from BIT Sindri, India.

Ali Sercan Basyurt is a research associate and PhD student in the chair for Business Information Systems and Digital Transformation at the University of Potsdam, Germany. His research interests are data science, machine learning, cybersecurity, and social media analytics. He has a background in Applied Cognitive and Media Science, focusing on computer science. Ali received his bachelor's and master's degrees from the University of Duisburg-Essen in Germany.

Stefan Stieglitz is a full professor of Business Information Systems and Digital Transformation at the University of Potsdam, Germany, and director of the Competence Center Connected Organization. Previously, he served as an assistant professor at the University of Münster and as a professor at the University of Duisburg-Essen, Germany. Furthermore, he held positions as a visiting professor and as an honorary professor at the University of Sydney, Australia. In his research, he investigates the impact of digital transformation on organizations, individuals, and society. His work has been published in reputable journals such as the *Journal of Management Information Systems*, the *European Journal of Information Systems*, the *Journal of Information Technology*, and *Business & Information Systems Engineering*. His articles have been awarded, among others, with the 'AIS Senior Scholars Best IS Publications Award' and the 'Stafford Beer Medal'.

Christoph Neuberger is a professor at the Institute for Media and Communication Studies at the Free University of Berlin. He is also the Director, scientific managing director, and Principal Investigator of the research group Digital News Dynamics at the Weizenbaum Institute in Berlin. He is a regular member of the Bavarian Academy of Sciences and Humanities (BAW) and the National Academy of Science and Engineering. His research focuses on the digital transformation of the public sphere, media, and journalism.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.