



Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Voting intentions on social media and political opinion polls

Viktor Pekar^a, Hossein Najafi^b, Jane M. Binner^c, Riley Swanson^b, Charles Rickard^d, John Fry^{e,*}^a Aston Business School, Aston University, Birmingham B4 7ET, UK^b College of Arts and Sciences, University of Wisconsin, USA^c Birmingham Business School, The University of Birmingham, Edgbaston, Birmingham B14 2TY, UK^d Division of Technology Services, University of Wisconsin – River Falls, 410 S. 3rd Street, River Falls, WI 54022, USA^e Department of Physics and Mathematics, University of Hull, Hull HU6 7RX, UK

ARTICLE INFO

Keywords:

Behavioural intentions
Forecasting
LSTMs
Machine learning
Neural networks
NLP
Political polls
Social media

ABSTRACT

Opinion polls play an important role in modern democratic processes: they are known to not only affect the outcomes of elections, but also have a significant influence on government policy after elections. Recent years have seen large discrepancies between polls and outcomes at several major elections and referendums, stemming from decreased participation in polls and an increasingly volatile electorate. This calls for new ways to measure public support for political parties. In this paper, we propose a method for measuring the popularity of election candidates on social media using Machine Learning-based Natural Language Processing techniques. The method is based on detecting voting intentions in the data. This is a considerable advance upon earlier work using automatic sentiment analysis. We evaluate the method both intrinsically on a set of hand-labelled social media posts, and extrinsically – by forecasting daily election polls. In the extrinsic evaluation, we analyze data from the 2016 US presidential election, and find that voting intentions measured from social media provide significant additional predictive value for forecasting daily polls. Thus, we demonstrate that the proposed method can be used to interpolate polls both spatially and temporally, thus providing reliable, continuous and fine-grained information about public opinion on current political issues.

1. Introduction

Political polls have a long history, dating at least as far back as the initiation of scientific polling by George Gallup in the 1930's. Today, opinion polls do not simply measure the current popularity of politicians, they are a critical tool for governments to understand a nation's attitude to different political and social issues and, as such, are a major factor in policy-making. During election campaigns, polls play an important role in determining the eventual outcomes of elections by shaping the behaviour of both voters (Blais, Gidengil, & Neville, 2006; Larsen & Fazekas, 2020; Madson & Hillygus, 2020) and politicians (Faas, Mackenrodt, & Schmitt-Beck, 2008; Walther & Hellström, 2019). There is evidence that polls conducted during election campaigns have a lasting impact on subsequent government policy (Burstein, 2003). At the same time so-called issue polls, which measure the public opinion on current political issues outside of election campaigns, are known to have a strong effect on everyday policy making (Rothmayr & Hardmeier, 2001; Schaffer, Oehl, & Bernauer, 2021; Shapiro, 2011).

Recent evidence suggests that traditional polls no longer provide adequate means to gauge public opinion in modern political realities. This can be understood indirectly from the fact that at a number of national elections and referendums, observers registered large discrepancies between opinion polls and final vote outcomes. Prominent examples are the Scottish independence referendum in 2014, the US presidential elections and the Brexit referendum in 2016. However, there is broader international evidence of biased polls spanning the US (Kimball, 2019), UK (Fry & Brint, 2017), Canada (Clarke, Goodwin, & Whiteley, 2017) and Germany (Meffert & Gschwend, 2011). The large discrepancies between polls and election outcomes are attributed to unrepresentative sampling (Sturgis et al., 2018), declining response rates at the polls (Kennedy et al., 2018; Wang, Rothschild, Goel, & Gelman, 2015), and social desirability effects such as the Bradley effect (Payne, 2010) and the Shy Tory effect (Sturgis et al., 2018). Social desirability has also been identified as the reason why traditional polls misrepresent public opinion: social desirability may cause respondents to exaggerate their likelihood of voting (Whiteley, 2016) and introduce

* Corresponding author.

E-mail addresses: V.Pekar@aston.ac.uk (V. Pekar), Hossein.Najafi@uwrf.edu (H. Najafi), J.M.Binner@bham.ac.uk (J.M. Binner), riley.swanson@my.uwrf.edu (R. Swanson), Charles.E.Rickard@uwrf.edu (C. Rickard), J.M.Fry@hull.ac.uk (J. Fry).<https://doi.org/10.1016/j.giq.2021.101658>

Received 8 February 2021; Received in revised form 26 October 2021; Accepted 18 November 2021

Available online 25 November 2021

0740-624X/© 2021 Elsevier Inc. All rights reserved.

bias into responses on issues such as immigration (Janus, 2010), same-sex marriage (Powell, 2013) and votes involving liberal vs. conservative attitudes (Funk, 2016). There is thus a need for new tools to monitor public opinion on current political issues that can augment traditional polling.

One attractive opportunity is offered by the analysis of social media data using the emergent Big Data and Artificial Intelligence technologies. These methods promise to provide snapshots of public opinion that is unsolicited, measured at a much higher frequency, based on larger samples, and disaster management compared to traditional polls (Beauchamp, 2017). Social media data has been shown to contain valuable information that can be used to support policy-making in different areas of governance, including healthcare (De Choudhury, Gamon, Counts, & Horvitz, 2013; Zeemering, 2021), policing (Gerber, 2014), labor market flows (Antenucci, Cafarella, Levenstein, Re, & Shapiro, 2014), disaster management (Dufty, 2016; Nguyen et al., 2017; Pekar, Binner, Najafi, Hale, & Schmidt, 2020). There are important challenges with the use of social media for measuring public opinion. One has to do with the fact that the demographics of social media users do not have the same distribution as the general population of a country. According to Greenwood, Perrin, and Duggan (2016) the former tend to be higher-income, higher-education and younger people. Their political affiliation is also more likely to be left of center compared to the public at large. Nonetheless, the proportion of the general population using social media has been rising over the past few years. According to Greenwood et al. (2016) 68% of all US adults used Facebook, 28% used Instagram, 25% used LinkedIn, and 21% used Twitter. The most recent Pew Research Center report (Auxier & Anderson, 2021) estimated that these numbers had a gradual upward trend: 69% used Facebook, 40% used Instagram, 28% used LinkedIn, and 23% used Twitter. As this trend continues, the representativeness of social media for studies of public opinion is likely to increase as well. On the other hand, there is a growing body of research on the automatic induction of demographic characteristics of social media users (e.g., Sanders, de Gier, & van den Bosch, 2016). The information provided by these methods can be used to correct for demographic biases and thus help to create models of public opinion that are more representative of the population as a whole. Another challenge is self-selection bias. A sample of social media data is likely to be biased towards users who feel strongly on particular issues and who actively post corresponding opinions. A recent study suggests that 10% of the most prolific users account for 92% of Twitter content (Smith & Grant, 2020). In our study, we counteract this bias by identifying and removing unusually active users in the overall dataset and during particular periods of the campaign (see Section 3.1.3).

Despite these challenges, the new opportunities have gained a lot of attention from researchers over the past decade. A considerable amount of work has been focused on the use of social media data to forecast election results (O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Singh, Dwivedi, Kahlon, Pathania, & Swahney, 2020; Tumasjan, Sprenger, Sandner, & Welp, 2011; Vepsäläinen, Li, & Suomi, 2017; Williams & Gulati, 2008; Yaqub, Chun, Atluri, & Viadya, 2017). Some findings attest to the potential of the data for the study of electoral phenomena, but the initial work also attracted considerable criticism. In particular, methodological flaws in attempts to predict actual election results from Twitter have been observed and their overall feasibility questioned (see e.g. Gayo-Avello, 2012; Jungherr, Jürgens, & Schoen, 2012). As discussed in Jungherr, Schoen, Posegga, and Jurgens (2017) the two basic approaches taken by the literature are to forecast election outcomes or to forecast polls. In this paper, we focus on the second problem. Specifically, we address the task of using social media data to forecast daily opinion polls, a research topic that has so far received only scant attention.

Early measures of political support derived from social media include simple statistics such as the number of mentions, hashtags or followers (Bovet, Morone, & Malse, 2018; Tumasjan et al., 2011), and automatic sentiment analysis of posts (e.g., O'Connor et al., 2010; Smailović,

Kranjc, Grčar, Žnidaršič, & Mozetič, 2015; Yaqub et al., 2017). The success of these approaches has been mixed. Simple statistics may portray the amount of attention a political party attracts rather than actual levels of support (Jungherr et al., 2017). Further, measuring political opinion using automated analysis of texts remains an extremely challenging problem.

In this paper, we develop a new method of assessing public support for a political candidate from social media. Our method uses Natural Language Processing to extract and quantify voting intentions in public messages posted on social media platforms. In contrast to existing sentiment analysis work we reduce the problem to the more concrete task of detecting voting intentions in social media. These voting intentions are more directly related to eventual voting choices than the sentiment expressed towards different political parties. We conduct both an intrinsic and an extrinsic evaluation of the method. In the intrinsic evaluation we compare automatic labelling of Twitter posts for voting intentions with labels assigned manually by human annotators. We demonstrate that the method is capable of accurately identifying voting intentions both for and against candidates. In the extrinsic evaluation we use the detected voting intentions as a predictor in forecasting models of daily election polls. Whereas these polls cannot serve as the ultimate ground truth about the popularity of a political candidate, they nonetheless provide a quantitative representation of public support that is still useful for judging the relative quality of different models. We find that voting intentions extracted from social media improve forecasting models of election polls. Voting intentions on social media thus contains information about candidate popularity over and above that contained in past observations of traditional polls.

The importance of our contribution is twofold. Firstly, we propose a novel method to measure support for political parties by detecting expressions of voting intentions on social media. This contrasts with previous work that relied on mention counts or sentiment analysis of social media content. Secondly, we demonstrate that voting intentions identified on social media have a significant relationship with measures of support derived from traditional polls, and can be used to improve their forecasts. Our method can thus potentially improve the coverage of traditional opinion polls by interpolating both across geographical regions and time periods, thereby providing continuous and fine-grained information about the public opinion on a particular subject. Ultimately, this line of research promises to facilitate evidence-based decision-making in public administration, increase its efficiency, and eventually translate into enhanced public value.

The layout of this paper is as follows. Section 2 reviews the literature. Section 3 presents our new method used to detect voting intentions via automatic linguistic analysis of social media messages and supervised classification and then formulates a model to forecast opinion polls. Section 4 provides an empirical evaluation. Section 5 explores governmental decision-making and governance implications. Section 6 concludes and proposes opportunities for further work.

2. Background and literature review

Previous work on measuring public support on social media relied on Machine Learning Natural Language Processing methods. In this section, we provide an overview of machine learning (Section 2.1) and Natural Language Processing (Section 2.2) and present terminology that will be used throughout the paper. Next, we discuss how these methods were previously applied to political forecasting (Section 2.3) and behavioural intention detection (Section 2.4).

2.1. Machine learning

The process of systematic examination of observations with the goal of finding patterns and using the discovered patterns to make future predictions has long been at the core of scientific enquiry. The recent explosion in the volume of data captured and stored by humanity has

given a huge impetus to the development of computational algorithms which help automate the process of finding patterns in the data. The field of study that aims to accomplish this is referred to as Machine Learning. Today Machine Learning finds application in disciplines and industries as diverse as agriculture, bioinformatics, economics, humanities, manufacturing, sports and arts.

Machine Learning is commonly defined as a field of study that gives computers the ability to learn without being explicitly programmed. Learning takes place by optimization algorithms that crunch through massive amounts of data with the goal of finding patterns, which can be applied to predict future events as well as obtain insights into factors affecting them. In this study, we follow a particular variety of Machine Learning methods, called *supervised algorithms*. In supervised settings, a machine learning method is presented with data which is labelled, i.e. where each observation is provided with a *target variable*, such as a categorical label or a numerical value representing the outcome of the observation. The goal of supervised methods is to estimate a *model* of a statistical phenomenon that will describe a mapping from different characteristics of known observations to their outcome and that can thus be used to predict the outcome of any new observation. The model is trained on a *training dataset*, and, then evaluated on a separate *test dataset*. During testing, measures of the *model quality* are calculated, based on a comparison of the true values of the target variable and the values predicted by the trained model. Because a model is developed on one subset of the data, but evaluated on another, one can ensure that the evaluation results will reflect the robustness of the model, i.e., how well it will apply to new, previously unknown observations drawn from the same population. Specifically, test-set performance will reveal if and how much the model has been affected by any noise present in the training data, if the models suffer from any bias, or if the model has overfitted the training data, and therefore will likely fail to apply to new observations. Machine Learning methods often involve *hyperparameters* that need to be fine-tuned to adapt the model to the specifics of the problem at hand. Hyperparameters are adjusted experimentally during model training, by holding out a part of the training data as a *validation dataset*. Once a model has been fine-tuned and evaluated, it can be deployed in real-world settings, where it will be able to generate predictions for previously unseen observations.

There are two types of supervised learning methods: *classification* and *regression*. A classification method will learn an *automatic classifier*. This is a classification model capable of outputting a categorical label for a given observation, e.g. a topical category label for a text document. A regression method will train a *regression model*, which will generate predictions on a continuous scale, such as the demand for a certain product or service. In this paper, we will first use classification methods to classify social media messages in terms of the authors' intention to vote at a forthcoming election. We will then use regression methods that operate on time series data to forecast voting intentions expressed at daily election polls. For an in-depth introduction to Machine Learning see James, Witten, Hastie, and Tibshirani (2021).

2.2. Natural language processing

Natural Language Processing (NLP) is an area of Artificial Intelligence concerned with the computational treatment of human language: it aims to enable computers to understand and generate human language. NLP is a highly multi-disciplinary area drawing primarily upon linguistics, computer science, and statistics. A specific NLP application is typically constructed as a pipeline of processing steps. It starts with simple problems analyzing the surface characteristics of the text, such as the *tokenization* of running text into separate words and punctuation symbols. It is then followed by intermediate steps such as *part-of-speech tagging* and *syntactic parsing*, to more difficult tasks such as *named entity recognition*. This includes identification of multi-word entities mentioned in the text, their classification, or *semantic role labelling* and *coreference resolution*. The aim is to detect semantic relationships between these

different entities. Modern NLP solutions, especially solutions to complex tasks involving grammatical and semantic interpretation of text, are based on Machine Learning methods. For example, a named entity recognition task is approached as a classification problem, where the goal is to assign a category label, such as Person, Location or Organization, to a named entity, given the grammatical and lexical characteristics of the surrounding words. As such, these NLP solutions are developed and evaluated within the usual Machine Learning workflow: a classification model is constructed on a set of manually labelled training documents and evaluated on a set of test documents.

Over the past two decades, a number of distinct research areas within NLP have emerged, corresponding to different practical applications, such as *text summarization* (automatic generation of a summary describing the gist of a document), *question answering* (generation of answers to questions formulated in natural language), and *sentiment analysis* (detection of the polarity of the sentiment that the author of a text expresses towards entities and events they describe). For a more detailed introduction to the field of NLP, see Jurafsky and Martin (2009).

2.3. Political forecasting using social media data

A large body of literature on political forecasting using social media has evolved on predicting election outcomes, either in terms of the elected party or the vote shares of the competing parties (see, e.g., Tumasjan et al., 2011; Di Grazia, McKelvey, Bollen, & Rojas, 2013; Wang et al., 2015). Early work presented evidence of the value of signals found on social media about public support for a political candidate. However, subsequent work has emphasized numerous flaws with the underlying methods. Gayo-Avello (2012), highlighted the fact that evaluation of these predictive models is done post-hoc and based on a single election outcome. Amid concerns over poor out-of-sample forecasting performance Jungherr et al. (2012, 2017) report negative results for German elections characterized by relatively uneventful election campaigns and a low rate of Twitter adoption. In this paper, we address the related problem of forecasting opinion polls throughout an election campaign, a task which is much more amenable to analysis and evaluation within well-established statistical learning procedures.

Previous research has explored a broad range of possible signals extracted from social media as potential predictors of political support. This included simple statistics such as the number of mentions (Sang & Bos, 2012; Tumasjan et al., 2011), the number of hashtags, retweets, likes, and followers on Twitter (Bovet et al., 2018) and the number of search engine queries (Mavragani & Tsagarakis, 2019). Despite apparent success stories, the use of raw counts has been criticized as indicating public attention paid to a party but not necessarily actual support (Jungherr et al., 2017).

A more sophisticated line of research has been to measure the "public mood" from posts and derive measures of sentiment that can be used to predict polls. To measure the polarity of tweets, previous political applications have implemented lexicon-based approaches or supervised learning methods. Lexicon-based approaches rely on predefined lists of words and phrases expressing positive and negative sentiment (O'Connor et al., 2010). Supervised learning methods construct sentiment classifiers from labelled examples of positive and negative posts (Bermingham & Smeaton, 2011; Neogi, Garg, Mishra, & Dwivedi, 2021; Smailović et al., 2015). Both approaches have had only limited success. Firstly, much of this work used existing sentiment analysis systems as generic tools, i.e. using systems developed and evaluated on other types of texts, thereby ignoring the fact that interpretation of sentiment is very much domain-dependent. Political discourse is characterized by particular ways of expressing support or disagreement, such as sarcasm, memes, allusions, implications and references to facts and events external to the current debate. Such linguistic phenomena are very hard to interpret using modern NLP tools, even in a system which has been specifically tailored to political applications. Secondly, as discussed

above, a further complication is that sentiment expressed on Twitter may only be loosely related to political support (Jungherr et al., 2017).

To address deficiencies associated with sentiment analysis applied to political texts, stance detection seeks to conduct lexical analysis within the context of social interactions to gauge the writer's attitude standpoint and judgement towards a proposition (Biber & Finegan, 1988). Political stance thus captures a person's affiliation within a particular cohort of other social network users with respect to a certain political issue (ALDayel & Magdy, 2021). Stance detection is therefore typically associated with linguistic features such as adjectives, adverbs and lexical terms (Jaffe, 2009). Several studies identify clusters of Twitter users based on their stance towards issues such as immigration and gun control in the US, reporting recognition of the political affiliation of a Twitter account with accuracy of over 90% (Darwish, Stefanov, Aupetit, & Nakov, 2020). A related strand of work aims to use social media data in order to study polarization and acculturation of voter preferences, by performing a geographical analysis of tweets, users and hashtags, as well as social network analysis (Grover, Kar, Dwivedi, & Janssen, 2019). However, it is not clear that these methods are well suited to the problem of forecasting opinion polls. Indeed, our own forecasting experiments (not reported) found no compelling evidence that stance detection methods can be used to predict polls. Political affiliation is known to be very stable over time, and polarization and acculturation are gradual processes. Techniques used to measure political stance, polarization and acculturation may not reflect short-term fluctuations in the voting behaviour of people who change their voting intentions more readily. This inability to detect short-term fluctuations leads to poor forecasting performance of voter preferences at fine-grained levels. Thus, this serves as additional motivation for the voting-intention method discussed in Section 3.1.

In this paper, we follow several previous studies that examine the forecasting of daily opinion polls. O'Connor et al. (2010) model the daily presidential job approval rating for Barack Obama over the course of 2009, and opinion polls during the 2008 U.S. presidential election cycle. They use a moving average forecasting model, where the exogenous predictor is the daily sentiment index, constructed from Twitter messages on the topic using a lexicon-based sentiment analyzer. Bermingham and Smeaton (2011) predict both the conventional opinion polls and the final election result at the 2011 Irish General Election, using a machine learning sentiment analyzer trained specifically for the task at hand, along with volume-based predictor features to construct a linear regression model. Ceron, Curini, Iacus, and Porro (2014) follow a similar approach, predicting political polls using a semi-supervised sentiment analyzer trained using a limited amount of hand-labelled relevant Twitter posts as well as volume-based features. Most of the early predictive models used are based on linear regression models operating on cross-sectional data created by averaging results of daily polls, and hence display strong bias. A more advanced approach is followed by Beauchamp (2017) who combines principled out-of-sample testing with both classical linear and nonlinear machine learning regression models. Predictor variables include Twitter word counts, state-level effects and a linear time trend. Beyond the raw forecasting challenge, the text analysis in Beauchamp (2017) identifies key themes associated with shifts in the levels of support for Democrats and Republicans.

2.4. Behavioural intentions detection

Detection of behavioural intentions from user-generated online content has been the focus of studies conducted in many areas of application. Queries submitted to search engines have been studied as a potential indicator of a purchase intention for different types of products, see e.g. Fantazzini (2014) for a review. Purchase intentions extracted from Twitter have also been investigated as a leading indicator of consumer demand (Najafi & Miller, 2015; Pekar, 2020).

Purchasing intentions outlined above constitute one of the key examples of behavioural intentions detection using social media data.

However, behavioural intentions detection extends far beyond purchasing intentions. Other examples include studies of criminal intentions (Resende de Mendonça, Felix de Brito, de Franco Rosa, dos Reis, & Bonacin, 2020), cyberbullying (Bastiaensens et al., 2014) and suicide ideation (Coppersmith, Leary, Crutchley, & Fine, 2018). Within this context, our contribution is therefore noteworthy as one of the first applications of behavioural intention detection using social media data to politics and governance.

Addressing the problem of voting behaviour, Jamal, Kizgin, Rana, Laroche, and Dwivedi (2019) argue that social networking sites provide a natural setting for studies of political engagement and use survey-based methods to compare online and offline forms of political participation with voting intentions. One of their striking findings is that online political participation may actually be negatively correlated with voting intentions. Misinformation (Aswani, Kar, & Ilavarasan, 2019) and fake news (Nasir, Khan, & Varlamis, 2021) in social media remain other important long-standing complications. This necessitates the development of new sophisticated techniques to detect voting intentions in social media data.

3. Methodology

Section 3.1 presents the proposed method to analyze social media data for indications of voting intentions. Section 3.2 describes a method to incorporate the extracted data on voting intentions into a model forecasting election polls.

3.1. Voting intentions on twitter

To obtain tweets containing election-related information from the stream of posts continuously published on Twitter, we created a set of search terms that are used as queries to the Twitter API. As search terms, we used the names of all the election candidates and associated hashtags. Next, we perform linguistic analysis of filtered tweets and apply supervised machine learning methods to detect three types of information in each tweet:

- (1) Whether or not an explicit expression of intention to vote was present;
- (2) Whether or not it was an intention to vote either in favor or against a particular candidate;
- (3) The name of the election candidate mentioned in relation to the voting intention.

We aim to register both the level of support as well as dislike of candidates. This latter issue may be particularly important if both candidates are relatively unpopular, as was the case with the 2016 US presidential election. The dislike for a certain candidate may have a major impact on election outcomes (Misch, Ferguson, & Dunham, 2018), and we therefore build this information into our forecasting model.

We develop and evaluate three automatic classification models, which correspond with the three types of information in each tweet listed above. The best-tuned classifiers were then used to process the full collection of election-related tweets and to construct an index of voting intention throughout the election campaign. The proposed method is described in further detail below.

3.1.1. Data collection, filtration and labelling

A collection of about 386 million election-related tweets were continuously collected from December of 2015 to a few days after the US presidential Election Day on November 8th of 2016. The data was obtained via Twitter API using a complete list of the names of all presidential candidates, along with their spelling variants and hashtags, as search terms. The overall workflow for processing the data is shown in Fig. 1.

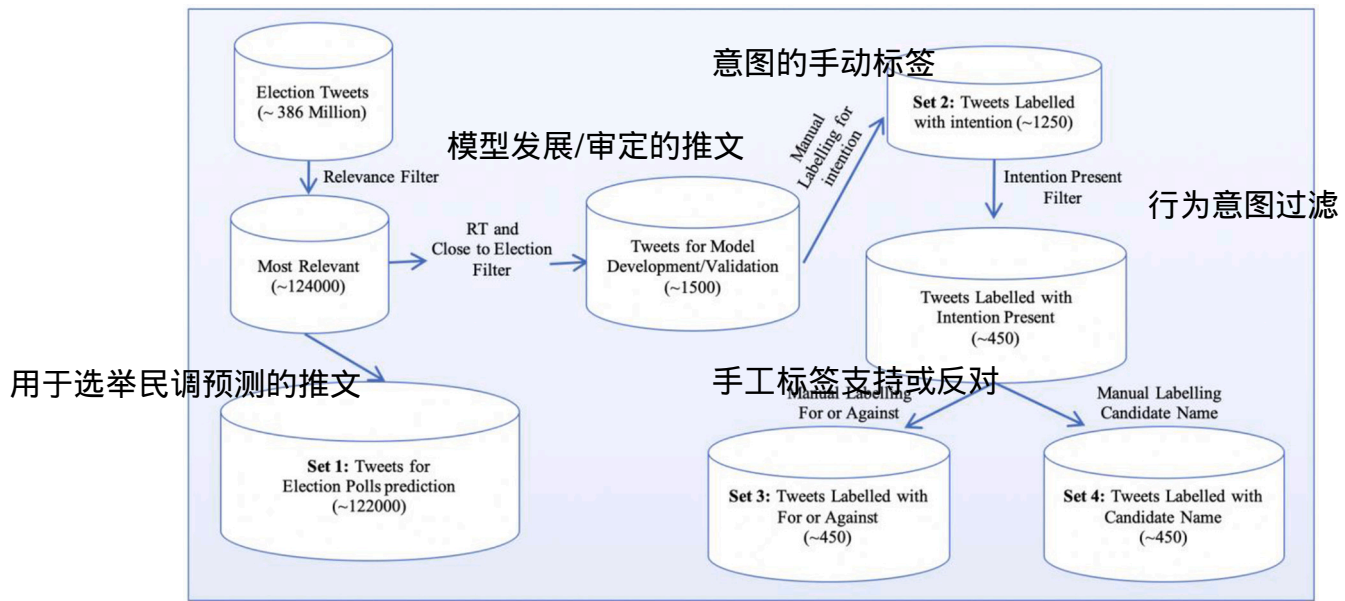


Fig. 1. Tweet collection, filtration and labelling process.

First, to generate a reliable data set for developing classification models, the complete set was further filtered to include only tweets that contained a first-person pronoun (“I”, “we”, “my”, “our”, “me”, “us”) followed by the words “vote”, “votes”, “voted” or “voting” within several words of the pronoun.

To reduce the self-selection bias in the data, and to increase the number of tweets containing a genuine expression to vote, we removed those tweets that were (1) exact duplicates of a previously published tweet, (2) written by all users who expressed an intention to vote more than ten times during the campaign and (3) all but the first tweet of the same author, if the author wrote more than one intention-related tweet on the same day. After this filtering step about 124,000 tweets were left. From these, a set of 1500 tweets was selected for hand-labelling and used for model development and validation. The remaining tweets were put aside for election polls prediction, shown as Set 1 in Fig. 1.

The selected tweets were independently labelled by two annotators in terms of the three classification tasks. To measure the agreement between the annotators, we used Cohen’s Kappa, a popular statistic that is used to measure inter-rater reliability for qualitative judgements (Landis & Koch, 1977). Table 1 provides Kappa scores and the size of each of the three sets of labels. According to guidelines in Landis and Koch (1977) these results point to “substantial” reliability for voting intention labelling ($\kappa > 0.61$) and “almost perfect” reliability for the other two tasks ($\kappa > 0.81$). Discarding tweets where annotators disagreed on the label, we obtained a “gold standard” set of 1254 tweets, Set 2 in Fig. 1. Of these tweets, 457 tweets were labelled as containing an intention, and were therefore used to create a set labelled in terms of for vs. against voting intention (Set 3) and a set labelled in terms of the name of the candidate, Set 4.

3.1.2. Voting intention models

To develop classification models for identification of voting intentions in the text of messages, we followed the common process of extracting features from tweets for content-based classification: after

removing all non-textual contents such as URLs and hashtags, we used unigrams (single words) and bigrams (all possible two-word sequences) as classification features. Next, we created custom features for the problems at hand, such as the ratio of positive to negative words in the tweet using the VADER sentiment lexicon (Hutto & Gilbert, 2014). This is a manually compiled general-purpose list of positive and negative English words, which has been used in a number of previous related studies to analyze the sentiment of a text. Intention-related words were detected following the method by Najafi and Miller (2015), based on the Harvard General Inquirer lexicon, which encodes various semantic categories of the core English vocabulary.

We trained and evaluated several models for each of the three classification problems: K-Nearest Neighbor (Mucherino, Papajorgji, & Pardalos, 2009), Gradient Boosting (Friedman, 2001), AdaBoost (Freund & Schapire, 1996), Random Forest (Breiman, 2001), Maximum Entropy, and Support Vector Machine (Cortes & Vapnik, 1995). Maximum Entropy classifiers performed best on all three tasks on the training data, and so in the following sections we present the results of experiments on the detection of voting intention using this method.

3.1.3. Voting intention index

The best-performing classifiers were used to detect voting intentions in the entire set of tweets collected during the election campaign. The full collection of tweets was pre-processed following the same procedure described in Section 3.1.1. After that, the three classifiers were arranged into a pipeline, where each tweet was first classified for the presence of voting intention. Then, positively labelled tweets were further classified for whether they express an intention to vote in favor or against a candidate. Finally, the name of the candidate was extracted. Since the two candidates who have reached this stage of the general election were by far the most represented in the dataset, we limited the analysis to tweets where either Clinton or Trump was mentioned. Tweets that were not found to contain an intention to vote were removed from further processing. These steps produced 48,881 tweets written by 41,029 unique authors, i.e. 1.19 tweets per author. The low number of tweets per author indicates that there is little bias in the data introduced by overly active supporters for either candidate.

Using this information as input, we calculate a daily *Social Media Voting Intention* (SMVI) index. At day i , the index for candidate c is given by:

Table 1
Size and Kappa statistics for classification data sets.

	Intention to Vote	Candidate	Voting For or Against
Size (# of tweets)	1254	453	457
Kappa	0.63	0.91	0.91

$$SMVI_{i,c} = \frac{f_{c,i} + \sum_{c' \neq c} a_{c',i}}{\sum_{k \in C} f_{k,i} + a_{k,i}}, \quad (1)$$

where C is the set of competing candidates, $f_{c,i}$ is the number of tweets published at i and containing an intention to vote for c , $a_{c,i}$ is the number of intentions to vote against c . The SMVI for a given candidate, therefore, ranges between 0 and 1 and increases with the number of intentions to vote for the candidate and the number of intentions to vote against other candidates.

3.2. Forecasting models

In this section, we incorporate the voting intentions data into a model to forecast the outcomes of traditional political polls. We use conventional autoregressive models that include past observations of polls as well as the Twitter data constructed in Section 3.1 as

$$y_t = \sum_{i=1}^p \beta_i y_{t-i} + \sum_{j=1}^p \omega_j x_{t-j} + e_t, \quad (2)$$

where y_t is the share of the vote from opinion polls at day t and the x_{t-j} is the exogenous variable, i.e., the SMVI index, at lag j .

To build forecasting models, we ran initial experiments with LSTM deep neural networks and several machine learning regression methods. We found that AdaBoost, Gradient Boosting and LSTMs produced the most promising results on validation data, consistent with a number of comparative studies of supervised learners, see e.g. Caruana and Niculescu-Mizil (2006), hence they were chosen for further development. We provide brief descriptions of these methods below.

AdaBoost (Freund & Schapire, 1996) is a boosting algorithm that combines an ensemble of multiple “weak learners”, such as decision trees or decision stumps. Each weak learner is trained successively and after one weak model is built, the algorithm identifies the most difficult instances and computes their weights to exaggerate their effect on the training of the next model. The goal of this step is to “teach” the next model to correctly predict the test instances on which errors were made. Initially, all instances have the same weight and hence have the same impact on the training of the initial model. After each iteration, the weights of instances are adjusted, while the weights of instances with accurate predictions are decreased. Furthermore, each model is assigned a weight based on its overall accuracy. During the testing phase, the forecast values and the weights of the models are taken into account to produce a weighted average value. To find the best hyper-parameter combination for AdaBoost models, we used a grid search over the number of weak learners (between 20 and 1000), learning rate (0.1 to 10) and the loss function (linear and exponential).

Gradient Boosting (Friedman, 2001) is a gradient descent ensemble algorithm, which, similar to other boosting methods, operates by sequential training of weak models, which would collectively form a strong model. This is accomplished by training successive regression models on the residuals of the previous model, computed from errors it made. With each training round, Gradient Boosting improves the previous model by adding to it a new model that is trained only on the residuals, thus gradually improving upon errors made in the previous steps. To prevent overfitting, we used an early stopping technique: stopping the training of the model if the validation loss has been increasing in four consecutive iterations. During training, we fine-tuned the following hyper parameters of the algorithm: the number of weak models, the learning rate, the maximum tree depth, the maximum number of features to consider before making a split and the minimum samples required to make a split. All remaining parameters were set to a default value.

Long-Short Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) are a type of Recurrent Neural Network specifically

designed to capture short-term as well as long-term dependencies in sequential data. This feature of LSTMs has recently been shown to be beneficial in many applications involving time series forecasting. RNNs are able to model the temporal dependencies in a sequence of events by using the weights on layers from previous observations as an additional input for the next observation. During training, we experimented with three-layer LSTMs: an input layer, one hidden layer with either five or ten cells, and the output layer with one cell. We also evaluated the effect of dropout, using a dropout rate of 0.2, as well as different bias regularization values between 0.0 and 0.1. We used 2000 epochs and the ReLU activation function for each model and we optimized learning rates with the Adam optimizer.

4. Empirical evaluation

This section presents the results of an empirical evaluation of the proposed method. Section 4.1 describes an intrinsic evaluation of the method for voting intention detection. Section 4.2 describes its extrinsic evaluation, where voting intention data are used to forecast pre-election polls.

4.1. Voting intention detection

As discussed in Section 3.1, our method for detecting voting intentions represents a pipeline of three classification models:

- (1) Intention Classifier, which detects the presence of a voting intention in a tweet;
- (2) For/Against Classifier, which recognizes if the intention is to vote for or against a candidate;
- (3) Candidate Classifier, which recognizes the candidate, i.e., the named entity to whom the expressed voting intention relates.

Maximum Entropy models were trained and evaluated using the manually labelled dataset described in Section 3.1.1. The dataset was split into a training-validation set and a test set, in the 80–20% proportion for each experiment conducted. The training set was used to identify parameters of a model via ten-fold cross-validation. Hyper parameters of the models were tuned via an exhaustive grid search: for each combination of hyper parameter settings a separate model was built on the training set and evaluated on the validation set. The optimized model was then evaluated on the test set, in order to detect any bias present in the model and to ensure the model’s robustness against noise. The results of this evaluation are reported below.

To have a point of reference for assessing the quality of the models, we used majority classifiers as baselines, which assigned all test instances to the majority class in the data. To evaluate the accuracy of the models, we used precision and recall (commonly used to measure the quality of automatic classifiers; Jurafsky & Martin, 2009). The precision with respect to class c is defined as the proportion of true positives for c , i.e. instances that were correctly classified as c , to the sum of true positives and false positives, the latter being the instances of other classes that were incorrectly classified as c :

$$P_c = \frac{tp_c}{tp_c + fp_c} \quad (3)$$

The recall of the classifier with respect to class c is defined as the proportion of true positives to the sum of true positives and false negatives, i.e., instances that should have been classified as c , and were not.

$$R_c = \frac{tp_c}{tp_c + fn_c} \quad (4)$$

In addition, we used the F1 score, which is a harmonic mean of precision and recall:

$$F_1 = 2 \frac{PR}{P+R}$$

首先开发单一特征的模型，之后将在多数极限上提高分类精度的特征组合起来，创建最终的分类特征集，并在此基础上训练最终模型

To obtain a single measure across all classes we used macro-averaged precision and recall rates, i.e., the averages of class-specific rates.

To select informative classification features in each of the three problems, we first developed models that used a single feature: the quality of each model thus indicated the informativeness of each feature. Those features that improved the classification accuracy over a majority

baseline were combined to create the final set of classification features, on which the eventual model was trained. Fig. 2 describes the precision, recall and F1 scores of the created models for the three tasks.

As can be seen from these results, for the easier task of candidate classification, an F1 score of 97.67 was achieved compared to the baseline model which achieved 41.0 for the same task. The F1 score of the For/Against Candidate classification was at 90.18 compared to the baseline score of only 36. For the harder task of intention classification, the F1 score was 75.65, an improvement of 174% over the 27.61 achieved by the respective baseline model. Table 2 summarizes the improvement for each model and each performance measure

The results demonstrate that the optimized models outperformed the majority baseline by a large margin, showing an improvement of anywhere from 43% to 220%. For two of the problems, Candidate and For/Against classifications, the F1 scores were over 90. On the third problem, Intention classification, the F1 score was somewhat lower, around 75. However, it should be emphasized that the Intention Classifier is the first classifier in our pipeline, and its precision of the positive class plays a more significant role than its recall. Its goal is not to identify all the tweets with voting intention, but rather to achieve a greater classification accuracy for tweets that are classified as having the intention to vote. In terms of precision, the Intention Classifier also achieved a high rate of 86.02, almost double the precision of the majority baseline.

It is interesting to compare these results with the level of agreement that human annotators displayed when manually labelling tweets relative to these three problems. Annotators tended to have a strong agreement on the tasks of For/Against classification and detection of candidate name: the kappa statistic was above 0.9. Automatic classifiers attained a very high level of accuracy on these problems, too: the F1 rates were very high, over 90. On the other hand, on the problem of Intention detection, annotators showed noticeably less agreement, the kappa being 0.63. Similarly, the problem was much more difficult for the automatic classifiers, where the F1 rate was only around 75.

The results of these experiments thus show that using a series of machine learning classifiers that operate over text-based features extracted from the body of the tweets, it is possible to detect voting intentions with a level of accuracy comparable to that which can be expected from human annotators. Next, we evaluate the utility of this data for predicting traditional opinion polls, by incorporating it into forecasting models.

评价结论

4.2. Forecasting applications

4.2.1. Target variable

To construct the target variable, i.e. the variable to be predicted, we used the 2016 US presidential elections polls data published by the FiveThirtyEight website (<https://fivethirtyeight.com/>). This data has been popularised by Silver (2013) and previously used in scientific work on electoral phenomena (Fry & Burke, 2020). This dataset contains results of state-level and nationwide voting intention surveys conducted by pollsters from Nov. 16th 2015 to Nov. 7th 2016. In this study we used only the nationwide surveys, as state-level geolocation data is not directly available in the Twitter dataset we use. This dataset also includes pollster ratings ranging from C- (the worst) to A+ (the best). Here, we included the survey results from all categories of pollsters. We, therefore, used data from 1106 nationwide surveys conducted by 54 pollsters. Table 3 shows examples of pollsters in each category as well as the total number of survey participants in each category. In each survey,

Table 2

Performance improvement of constructed models relative to the baseline model.

	F1, %	Precision, %	Recall, %
Intention Classifier	+174	+178	+43
For/Against Classifier	+150	+220	+83
Candidate Classifier	+138	+180	+96

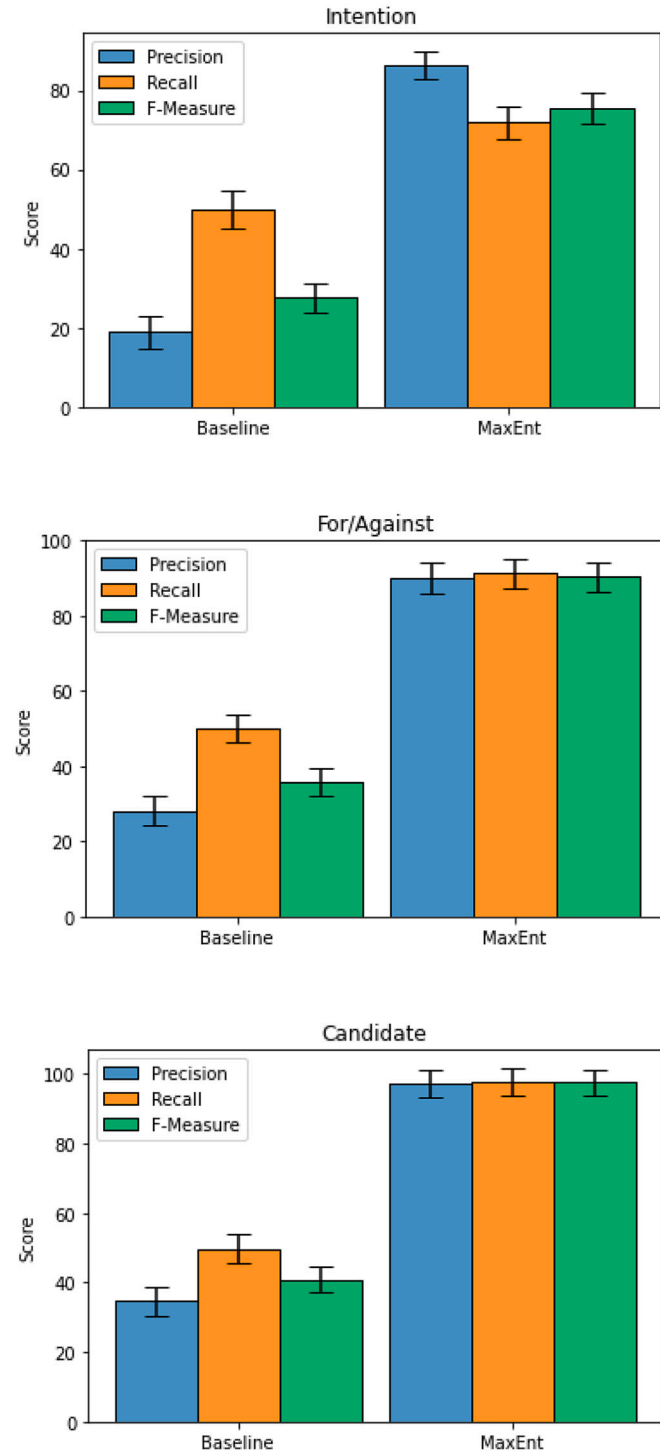


Fig. 2. Performance of best classifiers compared to the majority-class baseline, for each classification task. Top panel: intention detection. Middle panel: for/against classification. Bottom panel: candidate name detection.

Table 3

Examples of pollsters belonging to different grade categories.

Grade	Example pollsters	#Participants
A+	ABC News/Washington Post, Monmouth University, Selzer and Company	30,102
A	Fox News/Anderson Robbins Research/Shaw and Company Research, SurveyUSA	37,174
A-	Angus Reid Global, CBS News/New York Times, CNN/Opinion Research Corporation, Ipsos	514,043
B+	GfK Group, Pew Research Center, Princeton Survey Research Associates International	438,304
B	George Washington University (Battleground), YouGov, Google Consumer Surveys	53,666
B-	Gravis Marketing, Penn Schoen Berland, Schoen Consulting	71,466
C+	Rasmussen Reports/Pulse Opinion Research, American Research Group, CVOTER International	1000
C	TargetPoint	220,013
C-	McLaughlin and Associates, SurveyMonkey, Zogby Interactive/JZ Analytics	564,415

we obtained raw counts of polls participants who expressed an intention to vote for either Clinton or Trump. Table 4 shows descriptive statistics of raw daily counts of intentions to vote for Clinton and Trump registered at the polls.

Daily raw counts for either candidate were then added up across pollsters and converted into percentages. On a few dates during the campaign, e.g., public holidays, no surveys were undertaken. Values for these dates were obtained by linear interpolation. The raw counts of expressed intentions were converted to percentages and used as the target variable.

Fig. 3 compares the target variable and the SMVI index. The SMVI index fluctuates more widely over the period in question, which suggests it is more sensitive to underlying opinions in circulation. In contrast, the traditional opinion polls are rather stable over time, which may be a manifestation of herding behaviour resulting from a conformity bias, as suggested in past research on the topic (Payne, 2010; Sturgis et al., 2018). Further observation of Fig. 3 reveals that, whilst in the traditional polls, the fraction of intentions to vote expressed for Trump generally stays just below the 50% mark, the Twitter index indicates a substantially larger share, often around 60%, of Trump supporters. Considering the eventual Trump victory at the 2016 election, this observation seems to indicate that Twitter data may be a good source to mine for candidates' current popularity and has the capacity to improve the accuracy of traditional survey techniques.

4.2.2. Stationarity and lag selection

Before a time series can be used to estimate a forecasting model, one needs to ensure it is stationary, i.e. that its mean, variance and autocorrelation are constant over time. Any non-stationary time series can be made stationary through some form of transformation, such as differencing, whereby the original series is transformed into a series of period-to-period differences (Gujarati & Porter, 2009). To verify that the target and the predictor variables in our dataset are stationary, we performed two different tests for stationarity: the Augmented Dickey-Fuller test (Dickey & Fuller, 1979) and the KPSS test (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). The results are shown in Table 5. The original values of all three variables were found to be non-stationary and hence were made stationary by first differencing.

Next we selected the optimal number of lags for autoregressive

Table 4

Descriptive statistics of the total raw counts of voting intentions registered at opinion polls.

Candidate	Observations	Mean	St.Dev.	Minimum	Maximum
Clinton	340	3244.1	5891.2	95	85,245
Trump	340	2945.2	5346.2	102	76,708

models, i.e., the number of previous observations to be used as parameters in the models, by using three sets of Information Criteria (IC): Akaike IC (AIC), Bayesian IC (BIC) and Hannan-Quinn IC (HQIC). Since BIC and HQIC both suggest five lags whilst AIC suggests six; we use five lags in the autoregressive models. This selection is intuitively appealing as the five lags roughly correspond to the five days in the previous working week.

4.2.3. Autoregressive baselines

To construct forecasting models, the available data was divided into the training, validation and test parts, in the proportion 60%–20%–20%. All models were estimated and fine-tuned using the training and validation sets. After that, each model was evaluated on the test set, in order to detect any bias and check its robustness against noise. Since we use five-day lags to create endogenous variables, there are five-day gaps between the training and validation sets as well as between the validation and test sets, thereby ensuring that no training data is used for validation or testing.

Having trained on the training set and optimized the resulting parameters on the validation set, one-step-ahead forecasts were obtained from the test set. As evaluation metrics, we use the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), which are commonly used to measure the quality of fit of regression models (James et al., 2021). Both are calculated based on the differences between the model-predicted and ground truth values, but RMSE gives greater emphasis to large, albeit rare errors than MAE. The metrics are defined as:

$$RMSE := \sqrt{\frac{1}{T} \sum_{n=1}^T (y_n - \hat{y}_n)^2} \quad (6)$$

$$MAE := \frac{\sum_{n=1}^T |y_n - \hat{y}_n|}{T} \quad (7)$$

where T is the set of test instances, y_n is the observed value and \hat{y}_n is the forecasted value.

We then consider LSTM, AdaBoost and Gradient Boosting models built using only autoregressive variables. As a benchmark, to judge the models' quality, we use a persistence baseline to generate predictions by using the preceding day's value as the forecast for the following day. Results in Table 6 show that autoregressive models produce considerably more accurate predictions than the persistence baseline. The RMSE and MAE criteria both drop by between 40 and 50% in the case of LSTM and AdaBoost, whilst for Gradient Boosting, RMSE decreased by 30% and MAE by 16%. Results demonstrate that purely autoregressive models can lead to significant forecasting improvements when compared to the simple benchmark.

4.2.4. Models incorporating the SMVI index

Next, we train and evaluate LSTM, AdaBoost and Gradient Boosting models where autoregressive variables are augmented with the SMVI index. Fig. 4 compares the RMSE and MAE rates achieved using the SMVI with those of earlier autoregressive models. The confidence intervals were calculated by building and evaluating 30 models with the same hyper parameter combination, but using different random seed numbers as the initialization values for the learning algorithms. Table 7 reports the RMSE and MAE differences, alongside their significance, tested with an independent-sample *t*-test of the means over the 30 models. The significance of the differences between predicted test-set values was assessed using the Diebold-Mariano test (Diebold & Mariano, 1995), which tests the null hypothesis of equal forecasting accuracy of two methods. The Diebold-Mariano statistics for each model are shown in the last column with statistically significant differences indicated by asterisks. Significant results across the board for the Gradient Boosting model demonstrates that voting intentions can

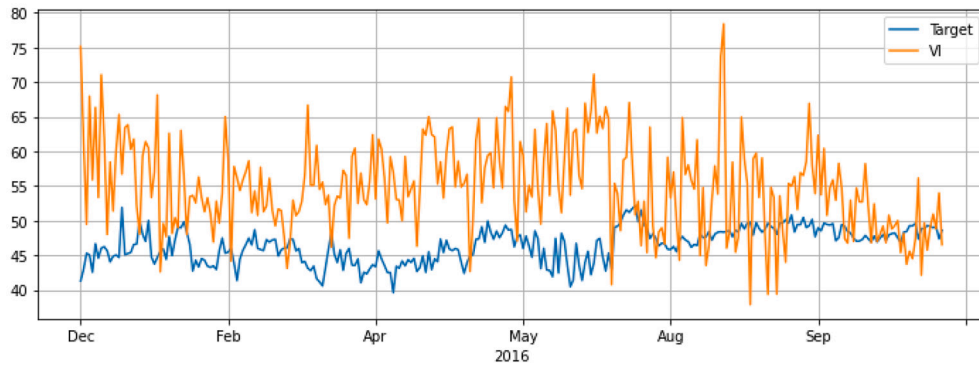


Fig. 3. The target variable and the SMVI index during the election campaign.

Table 5

Stationarity tests on the levels and first differences of the target and SMVI (asterisks indicate significance at the *0.1, **0.05 and ***0.01 significance levels).

	Levels		First difference	
	ADF	KPSS	ADF	KPSS
Target	0.046	1.066**	−5.818***	0.042
SMVI	−0.755	0.466**	−10.683***	0.07

Table 6

RMSE and MAE rates of the persistence baseline, AdaBoost, Gradient Boosting and LSTM models trained on only autoregressive variables.

Model	RMSE	MAE
Persistence baseline	1.898	1.402
AdaBoost	0.941	0.712
Gradient Boosting	1.18	0.974
LSTM	0.914	0.73

capture short-term fluctuations in polls. Amid mixed results for the LSTM model differences between forecasted values are not significant according to the Diebold-Mariano test. There is no evidence that the inclusion of SMVI improves the AdaBoost model. The actual predicted test-set values are shown in Fig. 5 (AdaBoost), Fig. 6 (Gradient Boosting) and Fig. 7 (LSTM).

In summary, these experiments have found that the voting intention index we construct using Twitter data has a long-run relationship with the voting intentions expressed by polls' participants. The inclusion of the SMVI index yields statistically significant improvements in the forecasting accuracy of Gradient Boosting regression models. The increased predictive power of the models demonstrates that voting intentions on social media contain information about candidate popularity amongst the general public. This information is supplementary to that contained in the poll responses registered on previous days.

5. Implications for governance

5.1. Implications for practice

Government agencies and public sector organizations are increasingly looking to make use of social media to enhance the quality of government services and enable greater citizen engagement, improve the feedback mechanism between the government and the public and reduce the costs of its operation (see e.g. Fry & Binner, 2016). Political polling, in particular, which is widely recognized to be in need of improvement, can benefit from innovations involving Artificial Intelligence technologies applied to social media data. This paper has proposed a new method to measure public support for a candidate on social media. The method was evaluated in the context of the problem of forecasting political opinion polls, and has thus presented a proof of concept both in terms of a new application of the intention-detection method and in the principled use of social media data in political forecasting applications. Our results show intentions expressed on social media may enable more accurate and timely forecasts of political opinion polls.

The method potentially offers a number of important practical advantages over the polling methods currently in operation. Firstly, research suggests that participation in traditional opinion polls has been decreasing (Kennedy et al., 2018; Wang et al., 2015). Given the growing

Table 7

The RMSE and MAE rates of the persistence baseline, AdaBoost, Gradient Boosting and LSTM models trained on only autoregressive variables, and their significance, for LSTM, AdaBoost and Gradient Boosting regressions.

	RMSE	MAE	Δ, RMSE, %	Δ, MAE, %	DM, t score
AdaBoost	0.942	0.709	+0.12	−0.45	−0.11
Gradient Boosting	1.12	0.901	−5.09***	−7.54***	2.43**
LSTM	0.95	0.728	+2.31**	−1.82**	−1.34

与目前正在使用的轮
询方法的比较优势

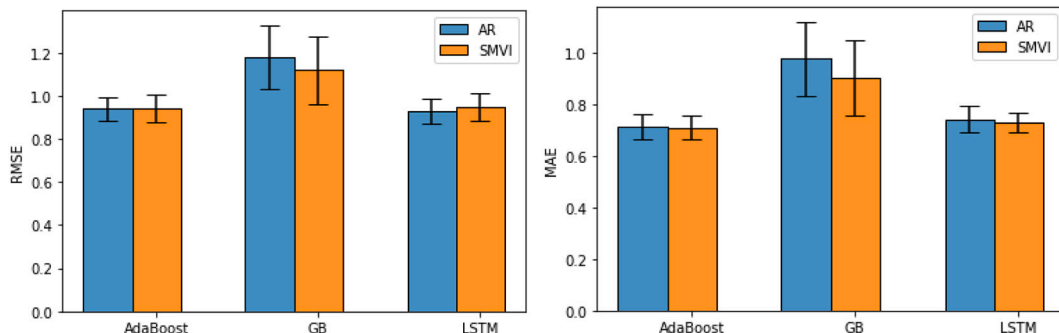


Fig. 4. RMSE (left) and MAE (right) of models built with autoregressive (AR) variables and SMVI variable.

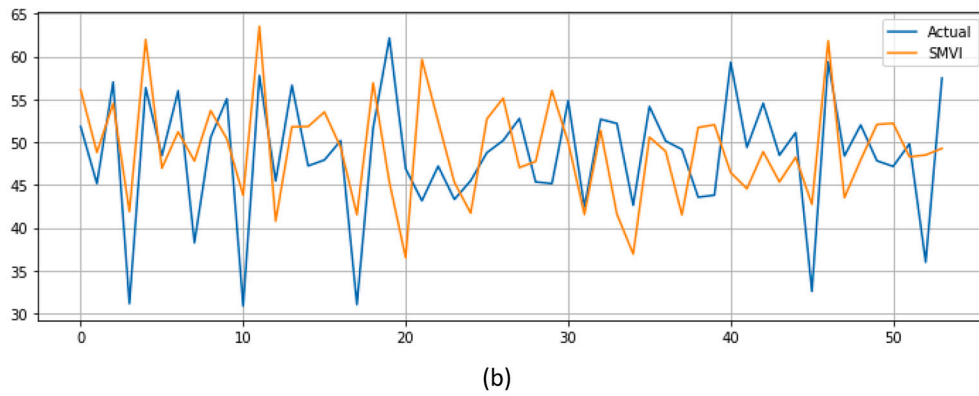
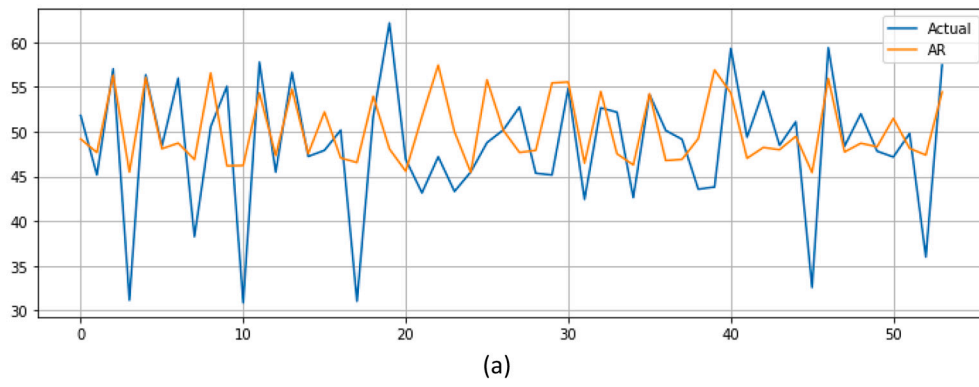


Fig. 5. Test-set forecasts produced with AdaBoost models trained on AR (a) and SMVI (b) variables.

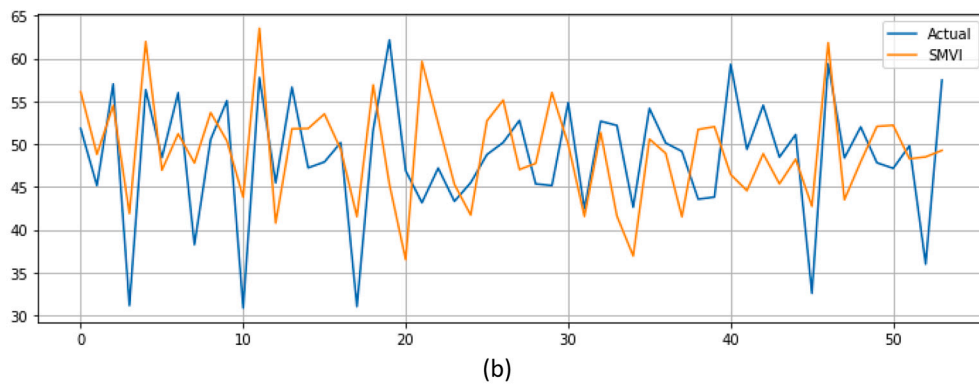
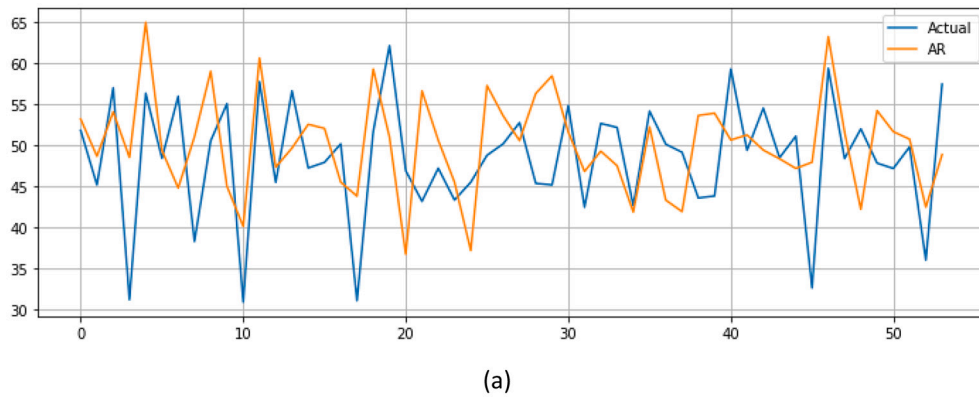


Fig. 6. Test-set forecasts produced with Gradient Boosting models trained on AR (a) and SMVI (b) variables.

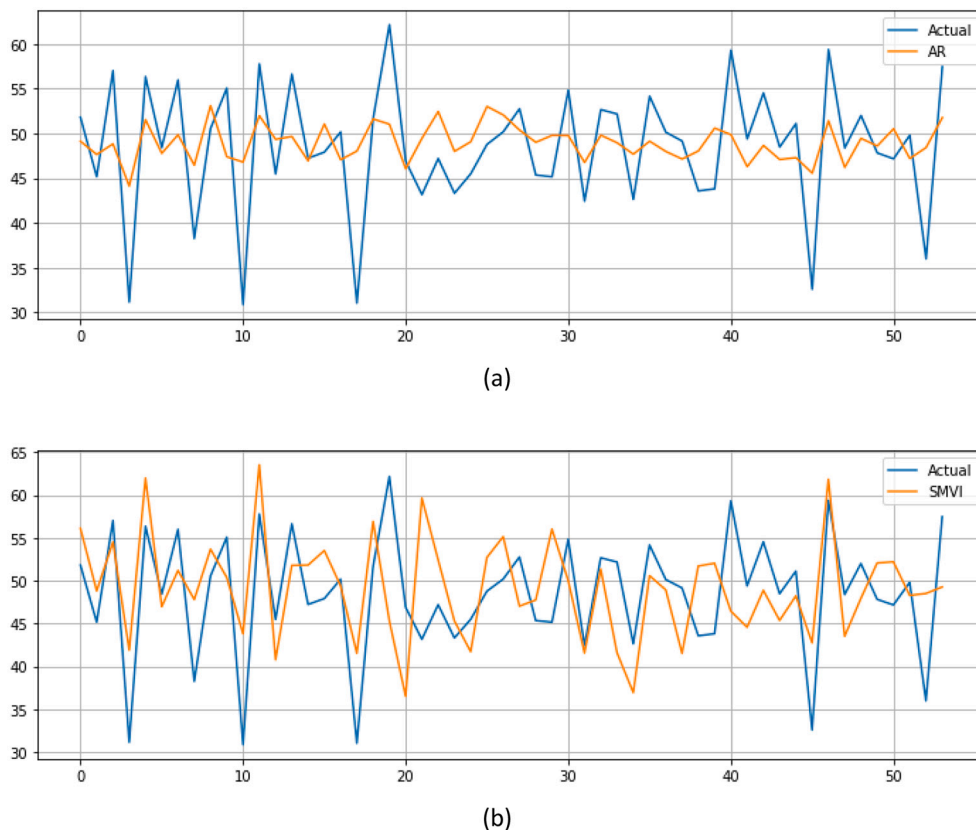


Fig. 7. Test-set forecasts produced with LSTMs trained on AR (a) and SMVI (b) variables.

popularity of social media in modern society, our proposed method can collect vast samples of voting intentions from which public opinion can be estimated, thereby compensating for the declining response rates at traditional polls. Secondly, the automatic nature of data collection now enables the measurement of public opinion at a much higher frequency and with a far greater geographic coverage. In this way, we are able to alleviate the problem of unrepresentativeness of samples that traditional polls suffer from (Sturgis et al., 2018). Thirdly, expressions of voting intentions on social media are unsolicited, and therefore less likely to be affected by social desirability bias (Payne, 2010; Sturgis et al., 2018). While voting intentions on social media may be affected by self-selection and deliberate manipulation, recent research has been developing ways to counteract these effects. In this paper, we have used a range of filtering techniques to minimize these effects and future work will be able to develop such methods further.

Although in this study the proposed method was evaluated on data on election polls, the application of this research is not limited to forecasting polls. Our new methods may also be applied to the public sector and can help improve government decisions and derive policy in many ways. In particular, the on-going coronavirus pandemic suggests that the forecasting of human-systems based on Twitter data is likely to remain very topical (see e.g. recent work in Zeemering, 2021). Intention to socially distance during a pandemic may be extracted from social media data using the methodologies proposed in this paper, providing crucial information to emergency services and first responders, and help to develop local and federal government responses accordingly. Similarly, the public's intention to switch to electric vehicles, use public transport or use solar panels may be extracted from social media and be used to lobby for/against policies for green energy and climate change. As such, further development and use of this line of research will, in the long run, facilitate situational awareness and evidence-based decision-making in governance, increasing its efficiency, and stimulating public engagement into the process of public administration.

5.2. Implications for theory

The theoretical implications of the study relate to the use of a new type of evidence to quantify political support on social media. Previous work explored a broad variety of signals extracted from text and user behaviour data on social media platforms in order to assess the support a political party is likely to have in the general public. These signals include counts of references to political candidates (Bovet et al., 2018; Tumasjan et al., 2011), the polarity of the sentiment of messages containing these references (O'Connor et al., 2010; Smailović et al., 2015; Yaqub et al., 2017), full-text lexical analysis of the messages (ALDayel & Magdy, 2021), follower-follower relationships and likes and repost counts (ALDayel & Magdy, 2021; Darwish et al., 2020).

In this paper, we examine expressions of voting intentions on social media as a possible predictor of political support. On the one hand, these expressions have specific semantics, pertain directly to intended activities of the electorate, and thus are less likely to be affected by language ambiguity that is known to harm approaches based on the analysis of sentiment or of individual word and phrases. On the other hand, given the variety of ways an intention to vote can be expressed, this is a challenging problem for automatic text analysis. Our findings show that it is feasible to identify voting intentions in social media with a very high accuracy, and that the voting intentions have a significant relationship to voting intentions expressed at traditional polls. Voting intentions on social media may thus be used to help improve forecasting models.

6. Conclusions and suggestions for future research

This paper contributes to the rapidly developing research area of political forecasting using social media data. We address two major concerns in the literature of social media data and their use in political forecasting: the appropriate handling and treatment of social media data and the comparison of models of political forecasting. We provide a detailed analysis of the data and the models, and we compare the results of the models with the results of the traditional polls. The results show that the proposed method is highly accurate and that the voting intentions on social media have a significant relationship to voting intentions expressed at traditional polls. The results also show that the proposed method is more robust to noise and outliers than the traditional polls. The results suggest that the proposed method can be used to improve forecasting models.

本研究不局限于预测民意调查，还可应用于公共部门

我们解决了文献中关于社交媒体数据的适当处理和比较模型的政治支持模型的正确预测基准的两个主要问题

Avello, 2012; Jungherr et al., 2012; Jungherr et al., 2017).

We develop a new method of measuring voting intentions by extending previous Machine-Learning techniques for detecting behavioural intention in text. Building on the results of a number of recent forecasting applications, we experiment with several different machine-learning and neural network techniques and construct nonlinear autoregressive models for opinion polls based on AdaBoost, Gradient Boosting and LSTM models. Diebold-Mariano tests confirm that inclusion of the Social Media Voting Intention Index yields significant improvements over a benchmark model constructed using either LSTM or Gradient Boosting methods.

We make two important contributions to the literature. Firstly, we extend previous work on behavioural intention detection from social media. Our approach goes beyond the mainstream literature on purchase intentions to detect the intention to vote and operationalize it to construct forecasting models of opinion poll data. By capturing both intentions to vote for and against candidates, the method takes into account both levels of support for as well as dislike of candidates. The latter may be a major factor in elections such as the 2016 US presidential election (Misch et al., 2018). Secondly, we demonstrate that our voting intention detection method yields both enhanced forecasting performance and greater political insights. Ultimately the successful application of our new method would yield both enhanced government policy and increased citizen satisfaction.

A number of theoretical and modelling challenges still need to be addressed to derive a fuller understanding of public opinion from social media data (Barberá & Rivero, 2015; Gayo-Avello, 2012). These challenges include the self-selection bias of social media users, the demographic bias present in social media and the deliberate spread of political (mis)information. In this paper, we were able to reduce the self-selection bias and the deliberate spread of information by applying different filters to remove data from overly active or otherwise suspect Twitter accounts. We have not explicitly addressed the issue of possible demographic biases in the data and this is a limitation of the current study. Participation in online political forums is associated with a myriad of complex cultural considerations (Jamal et al., 2019). The consequences of increased interactions between online and offline environments may also be hugely significant (Dey, Yen, & Samuel, 2020).


Future development of our method may refine the forecasting model by removing the inevitable biases present in social media. The effect of the demographic bias in a given dataset can be assessed and reduced by introducing variables that represent of different demographic groups, see Sanders et al. (2016), for early evidence on this approach. To minimize the effects of deliberate manipulation of social media content, suspect accounts and messages can be filtered out by drawing upon methods for identification of misinformation on social media (e.g., Aswani et al., 2019; Buntain & Goldbeck, 2017). In this paper, a method for assessing public support was evaluated by including its output into a model forecasting daily polls. A promising direction to extend this work is to use the method in models of election outcomes, where state-level geolocation information on social media would be used to forecast the share of the vote of different candidates across different constituencies.

Acknowledgements

The authors would like to acknowledge helpful and supportive comments from four anonymous reviewers, who helped us to significantly improve the manuscript. All remaining errors are our own.

References

- ALDayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing and Management*, 58, Article 102597.
- Antenucci, D., Cafarella, M., Levenstein, M., Re, C., & Shapiro, M. D. (2014). Using social media to measure labor market flows. In *Working paper 20010*. National Bureau of Economic Research.

- Aswani, R., Kar, A. K., & Ilavarasan, P. V. (2019). Experience: Managing misinformation in social media—Insights for policymakers from Twitter analytics. *Journal of Data and Information Quality*, 12, 1–18.
- Auxier, B., & Anderson, M. (2021). *Social Media use in 2021*. Pew Research Center.
- Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 33, 712–729.
- Bastiaenssens, S., Vandeboosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social networking sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behaviour*, 31, 259–271.
- Beauchamp, N. (2017). Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science*, 61, 490–503.
- Bermingham, A., & Smeaton, A. F. (2011). On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the workshop on sentiment analysis where AI meets psychology* (pp. 2–10). SAAIP.
- Biber, D., & Finegan, E. (1988). Adverbial stance types in English. *Discourse Processes*, 11, 1–34.
- Blais, A., Gidengil, E., & Neville, N. (2006). Do polls influence the vote? In H. E. Brady, & R. Johnston (Eds.), *Capturing campaign effects* (pp. 263–279). University of Michigan Press.
- Bovet, A., Morone, F., & Malse, H. A. (2018). Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. *Scientific Reports*, 8, 8673.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Buntain, C., & Goldbeck, J. (2017). Automatically detecting fake news in popular Twitter trends. In *IEEE conference on smart cloud* (pp. 208–215).
- Burstein, P. (2003). The impact of public opinion on public policy: A review and an agenda. *Political Research Quarterly*, 56, 29–40.
- Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning (ICML '06)* (pp. 161–168).
- Ceron,  (2014). Every Tweet counts? How sentiment and our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, 16, 340–358.
- Clarke, H. D., Goodwin, M., & Whiteley, P. (2017). *Brexit. Why Britain voted to leave the European Union*. Cambridge University Press.
- Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10, 117822618792860.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.
- Darwish, K., Stefanov, P., Aupetit, M., & Nakov, P. (2020). Unsupervised user stance detection on Twitter. In *Proceedings of the fourteenth international AAAI conference on web and social media (ICWSM 2020)* (pp. 141–152).
- De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media* (pp. 128–137).
- Dey, B. L., Yen, D., & Samuel, L. (2020). Digital consumer culture and digital acculturation. *International Journal of Information Management*, 51, Article 102057.
- Di Grazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behaviour. *PLoS One*, 8, Article e79449.
- Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74, 427–431.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–264.
- Duffy, N. (2016). Twitter turns ten: Its use to date in disaster management. *The Australian Journal of Emergency Management*, 31, 50–54.
- Faas, T., Mackenrodt, C., & Schmitt-Beck, R. (2008). Polls that mattered: Effects of media polls on voters' coalition expectations and party preferences in the 2005 German parliamentary election. *International Journal of Public Opinion Research*, 20, 299–325.
- Fantazzini, D. (2014). Nowcasting and forecasting the monthly food stamps data in the US using online search data. *PLoS One*, 9, Article e111894.
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *ICML '96: Proceedings of the thirteenth international conference on machine learning* (pp. 148–156).
- Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.
- Fry, J., & Binner, J. (2016). Elementary modelling and behavioural analysis for emergency evacuations using social media. *European Journal for Operations Research*, 249, 1014–1023.
- Fry, J., & Brint, A. (2017). Bubbles, blind spots and Brexit. *Risks*, 5, 37.
- Fry, J., & Burke, M. (2020). An options-pricing approach to election prediction. *Quantitative Finance*, 20, 1583–1589.
- Funk, P. (2016). How accurate are surveyed preferences for public policies? Evidence from a unique institutional setup. *The Review of Economics and Statistics*, 98, 442–454.
- Gayo-Avello, D. (2012). No you cannot predict elections using Twitter data. *IEEE Internet Computing*, 16, 91–94.
- Gerber, M. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115–125.
- Greenwood, S., Perrin, A., & Duggan, M. (2016). *Social Media update 2016*. Pew Research Center.
- Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2019). Polarization and acculturation in US Election 2016 outcomes - can Twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145, 438–460.

- Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735–1780.
- Hutto, C. J., & Gilbert, E. E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of Social Media text. In *Eighth international conference on weblogs and social media (ICWSM-14)* (pp. 216–225).
- Jaffe, A. (Ed.). (2009). *Stance: Sociolinguistic perspectives*. Oxford University Press.
- Jamal, A., Kizgin, H., Rana, N. P., Laroche, M., & Dwivedi, Y. K. (2019). Impact of acculturation, online participation and involvement on voting intentions. *Government Information Quarterly*, 36, 510–519.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer.
- Janus, A. L. (2010). The influence of social desirability pressures on expressed immigration attitudes. *Social Science Quarterly*, 91, 928–946.
- Jungherr, A., Jürgens, P., & Schoen, A. (2012). Why the pirate party won the German election of 2009 and the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G. & Welpel, I. M. In *Social science computing review: 30. Predicting elections with Twitter; what 140 characters reveal about political sentiment* (pp. 229–234).
- Jungherr, A., Schoen, H., Posegga, O., & Jurgens, P. (2017). Digital trace data in the study of public opinion: An indicator of attention towards politics rather than political support. *Social Science Computer Review*, 35, 336–356.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing* (2nd ed.). Prentice-Hall.
- Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., ... Wleizen, C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82, 1–33.
- Kimball, S. (2019). Presidential statewide polling – A substandard performance: A proposal and application for evaluating pre-election poll accuracy. *The American Behavioral Scientist*, 63, 768–788.
- Kwiatkowski, D., Phillips, P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54, 159–178.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Larsen, E. G., & Fazekas, Z. (2020). Transforming stability into change: How the media select and represent opinion polls. *International Journal of Press/Politics*, 25, 115–134.
- Madson, G. J., & Hillygus, D. S. (2020). All the best polls agree with me: bias in evaluations of political polling. *Political Behaviour*, 42, 1055–1072.
- Mavragani, A., & Tsagarakis, K. P. (2019). Predicting referendum results in the Big Data era. *Journal of Big Data*, 6, 3.
- Meffert, M. F., & Gschwend, T. (2011). Polls, coalition signals and strategic voting: An experimental investigation of perceptions and effects. *European Journal of Political Research*, 50, 636–667.
- Misch, A., Ferguson, G., & Dunham, Y. (2018). Temporal dynamics of partisan identity fusion and pro-sociality during the 2016 US presidential election. *Self and Identity*, 17, 531–548.
- Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). *Data mining in agriculture*. Springer.
- Najafi, H., & Miller, D. (2015). Comparing analysis of social media content with traditional survey methods of predicting opening night box-office revenues for motion pictures. *Journal of Digital and Social Media Marketing*, 3, 262–278.
- Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, 1, Article 100007.
- Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights*, 1, Article 100019.
- Nguyen, D., Mannai, K. A. A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional. *Neural Networks*, 1, 632–635.
- O'Connor, B., Balasubramanian, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international AAAI conference on weblogs and social media 2010* (pp. 122–129).
- Payne, J. G. (2010). The Bradley effect: Mediated reality of race and politics in the 2008 U.S. presidential election. *The American Behavioral Scientist*, 54, 417–435.
- Pekar, V. (2020). Purchase intentions on social media as predictors of consumer spending. In *Proceedings of the 14th international AAAI conference on web and social media* (pp. 545–556).
- Pekar, V., Binner, J., Najafi, H., Hale, C., & Schmidt, V. (2020). Early detection of heterogeneous disaster events using social media. *Journal of the Association for Information Science and Technology*, 71, 43–54.
- Powell, R. J. (2013). Social desirability bias in polling on same-sex marriage ballot measures. *American Politics Research*, 41, 1052–1070.
- Resende de Mendonça, R., Felix de Brito, D., de Franco Rosa, F., dos Reis, J. C., & Bonacin, R. (2020). A framework for detecting intentions of criminal acts in social media: A case study on Twitter. *Information*, 11, 154.
- Rothmayr, C., & Hardmeier, S. (2001). Government and polling: Use and impact of polls in the policy-making process in Switzerland. *International Journal of Public Opinion Research*, 14, 123–140.
- Sanders, E., de Gier, M., & van den Bosch, A. (2016). Using demographics in predicting election results with Twitter. In E. Spiro, & Y. Y. Ahn (Eds.), *10047. Social informatics. SocInfo 2016* (pp. 259–268). Springer. Lecture Notes in Computer Science.
- Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 Dutch senate election results with Twitter. In *Proceedings of the workshop on semantic analysis in social media (EACL '12)* (pp. 53–60).
- Schaffer, L., Oehl, B., & Bernauer, T. (2021). Are policymakers responsive to public demand in climate politics? *Journal of Public Policy*, 1–29.
- Shapiro, R. Y. (2011). Public opinion and American democracy. *Public Opinion Quarterly*, 75, 982–1017.
- Silver, N. (2013). *The signal and the noise: The art and science of prediction*. Penguin.
- Singh, P., Dwivedi, Y. K., Kahlon, K. S., Pathania, A., & Swahney, R. S. (2020). Can Twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections. *Government Information Quarterly*, 37, Article 101444.
- Smailović, J., Kranjc, J., Grčar, M., Žnidaršič, M., & Mozetič, I. (2015). Monitoring the Twitter sentiment during the Bulgarian elections. In *IEEE DSAA '2015: Proceedings of the 2015 IEEE international conference on data science and advanced analytics* (pp. 1–10).
- Smith, A., & Grant, A. (2020). *Differences in how democrats and republicans behave on twitter*. Pew Research Center.
- Sturgis, P., Kuha, J., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Lauderdale, B. E., & Smith, P. (2018). An assessment of the causes of the errors in the 2015 UK general election opinion polls. *Journal of the Royal Statistical Society A*, 181, 757–781.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpel, I. M. (2011). Election forecasts with Twitter: What 140 characters reveal about political sentiment. *Social Science Computer Review*, 29, 402–418.
- Vepsäläinen, T., Li, H., & Suomi, R. (2017). Facebook likes and public opinion. Predicting the 2015 Finnish parliamentary elections. *Government Information Quarterly*, 34, 524–532.
- Walther, D., & Hellström, J. (2019). The verdict in the polls: How government stability is affected by popular support. *West European Politics*, 42, 593–617.
- Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31, 980–991.
- Whiteley, P. (2016). Why do voters lie to the pollsters? *Political Insight*, 7, 16–19.
- Williams, C. B., & Gulati, G. J. (2008). The political impact of Facebook: Evidence from the 2006 midterm elections and 2008 nomination contest. *Politics and Technology Review*, 1, 11–21.
- Yaqub, U., Chun, S. A., Atluri, V., & Viadya, J. (2017). Analysis of political discourse on Twitter in the context of the 2016 US presidential election. *Government Information Quarterly*, 34, 613–626.
- Zeemering, E. S. (2021). Functional fragmentation in city hall and Twitter communication during the COVID-19 pandemic: Evidence from Atlanta, San Francisco and Washington DC. *Government Information Quarterly*, 38, Article 101539.

Viktor Pekar is a teaching fellow in Business Analytics at the Business School, Aston University, UK. He holds a PhD in Computational Linguistics, and previously worked as a post-doctoral researcher and a teaching fellow at the universities of Karlsruhe, Wolverhampton, and Birmingham, specialising in Artificial Intelligence, Data Science, Machine Learning, Computational Social Sciences and Natural Language Processing. He has over forty publications at major international conferences and journals on Natural Language Processing and Artificial Intelligence.

Hossein Najafi is a professor of Computer Science and Information Systems at the University of Wisconsin, River Falls. He holds a Ph.D. in Electrical Engineering from the University of Minnesota with specializations in Artificial Intelligence, Machine Learning, and Neural Networks. Hossein has over thirty years of experience researching with Neural Networks and their application to real-world problems. He has over thirty publications related to his specializations in major international conferences and journals.

Jane Binner joined the Accounting and Finance Department at Birmingham Business School as Chair of Finance in August 2013. Prior to this she worked as Head of the Accounting and Finance Division at Sheffield Management School and as Reader in Economics at Aston Business School for seven years. Jane has a PhD, MSc, PGCE and BA Hons in Economics from the University of Leeds. She has worked with a number of stakeholder groups such as the Home Office, Experian, the Boots Group plc and Wright Patterson Airforce Base. She brings expertise in analyzing the strategic investment decisions of large enterprises through econometric modelling. Jane has conducted research in econometrics for over twenty years and has extensive academic and commercial experience. Binner has attracted over £1000,000 in external research funding, including awards from the EPSRC/ESRC, the Leverhulme Trust, the National Science Foundation, the Jan Wallander Foundation as well as industrial funding from Boots and Experian. Binner has achieved international recognition for her work on the econometric performance of monetary aggregates and is world leading in her field of financial innovation in the construction of money. Jane has recently been appointed as a visiting professor at the College of Business and Economics Research Centre at the University of Wisconsin, USA and as an INDI Fellow at the Institute for Nonlinear Dynamical Inference. She has four books and over seventy publications in the area of Computational Finance and Economics.

Riley Swanson is a software engineer working for Baldwin Technologies. He has a Bachelors of Science in Computer Science and in Data Science from University of Wisconsin River Falls. While working full time he is also pursuing a Masters in Analytics from Georgia Institute of Technology as a part time student.

Charles Rickard graduated with a Bachelor of Science degree in Math and Computer Science from the University of Wisconsin – River Falls. He is currently a staff member at the university in the Division of Technology Services and is working towards his Masters in

Computers Science. His interests include software integration, and researching topics in data science and machine learning.

John Fry is Senior Lecturer in Applied Mathematics at the University of Hull. John has a degree in mathematics and statistics from the University of Newcastle-upon-Tyne and a

PhD in Mathematical Finance from the University of Sheffield. John has published in a number of leading academic journals including *Economics Letters*, *European Journal of Operational Research*, *European Physical Journal B*, *International Review of Financial Analysis*, *Journal of Business Research* and *Quantitative Finance*.