

# Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach

Ruowei Liu, Xiaobai Yao, Chenxiao Guo & Xuebin Wei

To cite this article: Ruowei Liu, Xiaobai Yao, Chenxiao Guo & Xuebin Wei (2021) Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach, *Annals of GIS*, 27:1, 43-56, DOI: [10.1080/19475683.2020.1829704](https://doi.org/10.1080/19475683.2020.1829704)

To link to this article: <https://doi.org/10.1080/19475683.2020.1829704>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group, on behalf of Nanjing Normal University.



Published online: 22 Oct 2020.



Submit your article to this journal [↗](#)



Article views: 1691





View related articles [↗](#)



View Crossmark data [↗](#)

# Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach

Ruowei Liu <sup>a</sup>, Xiaobai Yao <sup>a</sup>, Chenxiao Guo <sup>b</sup> and Xuebin Wei <sup>c</sup>

<sup>a</sup>Department of Geography, University of Georgia, Athens, GA, USA; <sup>b</sup>Department of Geography, University of Wisconsin-Madison, Madison, WI, USA; <sup>c</sup>Department of Integrated Science and Technology, James Madison University, Harrisonburg, VA, USA

## ABSTRACT

Forecasting political elections has attracted a lot of attention. Traditional election forecasting models in political science generally take preference in poll surveys and economic growth at the national level as the predictive factors. However, spatially or temporally dense polling has always been expensive. In the recent decades, the exponential growth of social media has drawn enormous research interests from various disciplines. Existing studies suggest that social media data have the potential to reflect the political landscape. Particularly, Twitter data have been extensively used for sentiment analysis to predict election outcomes around the world. However, previous studies have typically been data-driven and the reasoning process was oversimplified without robust theoretical foundations. Most of the studies correlate twitter sentiment directly and solely with the election results which can hardly be regarded as predictions. To develop a more theoretically plausible approach this study draws on political science prediction models and modifies them in two aspects. First, our approach uses Twitter sentiment to replace polling data. Second, we transform traditional political science models from the national level to the county level, the finest spatial level of voting counts. The proposed model has independent variables of support rate based on Twitter sentiment and variables related to economic growth. The dependent variable is the actual voting result. The 2016 U.S. presidential election data in Georgia is used to train the model. Results show that the proposed model is effective with the accuracy of 81% and the support rate based on Twitter sentiment ranks the second most important feature.

## ARTICLE HISTORY

Received 3 January 2020  
Accepted 17 September 2020

## Keywords

election forecasting;  
sentiment analysis; Twitter;  
political science; machine  
learning; location-based  
social media data

## Introduction

Forecasting the presidential election has attracted a lot of attention in academia and the public. In the literature of election forecasting, two threads of distinct and popular methods are found. One thread started in the field of political science. From the 1980s, political scientists have taken an effort to develop a variety of election prediction models to examine the relationships between predicted voting results of one presidential election candidate, usually from the incumbent party, and several predictive variables, usually support rate from election poll survey and economic growth. The second thread is from the field of computer science. Since the 2010s, as the popularity of social media big data increases, a strand of research has started to predict elections based on sentiments expressed on social media platforms such as Twitter. The main goal is usually to calculate the sentiment scores from related social media posts as accurately as possible. If the sentiment is positive towards a candidate, they predict this candidate will win in the election.

Despite all the scientific merits and practical values, both types of methods suffer drawbacks. The traditional election forecasting models require supporting rates from poll surveys. However, conducting a poll survey is typically very expensive and time-consuming. There are also limitations in the new thread of research based on social media sentiments. This type of study was criticized as failing to establish a relationship between Twitter sentiment and voting results in elections (Gayo-Avello 2012b), but only focusing on sentiment calculations in a data-driven fashion. Furthermore, studies in both fields usually forecast elections nationwide while barely conduct the prediction at a finer geographic level such as the county level.

To address the existing problems mentioned, this study aims to learn from both fields of research and integrate the two threads of approaches. On the one hand, Twitter data are relatively more accessible and less expensive comparing with the polling survey, and people can share opinions on a more voluntary basis. On the other hand, traditional election forecasting models in political science are built on a more robust theoretical

background. Integrating the two approaches may forecast the presidential election results in a more comprehensive way (Gayo-Avello 2013). Learning from the fields of political science, a generalization of presidential election forecasting models can be expressed as  $V(E, P)$ , in which  $V$  represents the voting results,  $E$  represents economic growth, and  $P$  represents the support rate in the poll survey. This study subscribes to the same theoretical framework but makes two innovative improvements. First, existing studies of the traditional approach are the macro-level nationwide models. This study develops models at the county level to accommodate variations across geography. Second, the study replaces supporting rates in the traditional model with sentiments that can be learned from social media data. Specifically, we calculate sentiments towards each candidate from geo-tagged tweets in each county. These estimations of sentiment are used to substitute  $P$  at the county-level. Finally, following the underlying theory of the traditional political science approach, this study builds a sentiment-based election forecast model to establish the relationship between voting results and predictive factors such as economic growth and support rate at the county-level. We use the 2016 U.S. presidential election in Georgia for a case study.

The paper is organized as follows. Section 2 reviews the literature on traditional election forecasting models in political science, election prediction based on Twitter sentiments as well as the comparison between Twitter sentiments and poll data. Section 3 introduces the data used in this study and discusses data preprocessing methods. The data include Twitter posts from September 26<sup>th</sup> (the first debate) to November 8<sup>th</sup> (election day) in 2016 and socio-economic statistics at the county level. Section 4 presents data analysis, modelling methods, as well as results with interpretation.

Finally, this paper concludes with a discussion of findings and future research directions in section 5.

### Literature review

To conduct this research, we briefly look through the literature from these relevant fields, including election forecasting models in political science, election prediction based on Twitter sentiments, comparison between Twitter sentiments and poll data. Figure 1 shows the connection between our research and the current literature.

### Election forecasting models in political science

In the past decades, political scientists have proposed a series of election forecasting models. Lewis-Beck and Rice (1982) built the first presidential election forecasting model which is rooted in the political science theory. The model treats the job approval rating for the president in the July Gallup poll and the gross national product (GNP) growth rate in the first two quarters of the election year as two predictive factors. Based on this model, Campbell and Wink (1990) developed the **trial-heat** model. The model also consists of two predictive variables, namely the incumbent party’s candidate support on Gallup poll in early September and the second-quarter growth rate in the real GDP of the election year. In 1992, the group developed another model named the **convention-bump** model which considers three predictors including the incumbent party’s candidate support of the pre-convention polls, the net change of the incumbent party’s candidate support after both conventions are completed, and the second-quarter GDP growth rate in the economy (Campbell, Cherry, and Wink 1992).

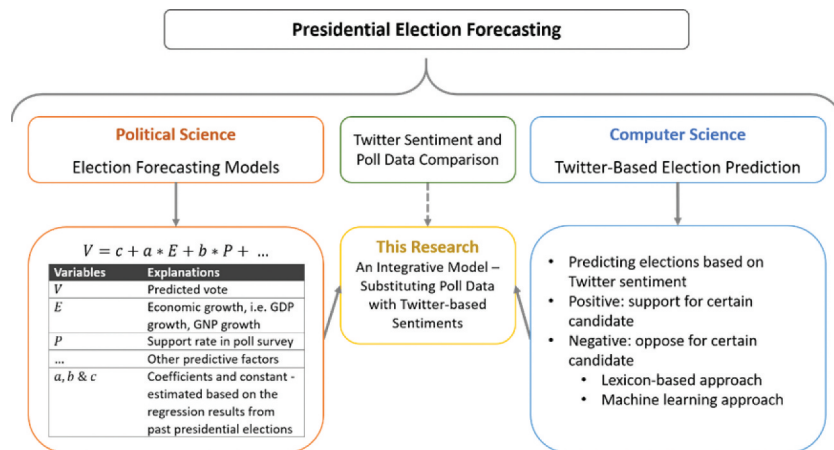


Figure 1. The structure of literature review and research gap.

While the previous models are in the form of mathematical functions applicable at the national level, Holbrook and DeSart (1999) presented a simple forecasting model, dubbed as the **long-range** model, for state-level presidential outcomes based on statewide preference polls and a lagged vote variable. This model considers four predictive variables, including the statewide voting results from the previous election, the national polls taken in October before the election, whether the state is the home state for Democratic or Republican candidate, and the number of terms the Democratic currently occupying. Abramowitz (2001) presented the **time-for-change** model. It is based on three predictors: the incumbent party candidate's mid-year approval rating in the Gallup poll (late June or early July), the growth rate of real GDP in the second quarter of the election year, and whether the incumbent president's party has held the White House for one term or more than one term. A generalized equation of all the models mentioned above is shown below. Table 1 explains each variable in this generalization model.

$$V = c + a * E + b * P + \dots \quad (1)$$

### **Election prediction based on Twitter sentiments**

With the advent and increasing popularity of big data in the current century, researchers started to incorporate Twitter data as the assistance in election predictions (Birmingham and Smeaton 2011; Chung and Mustafaraj 2011; Ahmed, Jaidka, and Skoric 2016). A stream of studies has suggested that Twitter data are powerful for the prediction of political election outcomes by using sentiments extracted from tweets in the U.S. (Ceron, Curini, and Iacus 2015; Paul et al. 2017; Swamy, Ritter, and de Marneffe 2017; Grover et al. 2019) and other countries (Sang and Bos 2012; Ceron et al. 2014; Burnap et al. 2016; L. Wang and Gan 2017). Tweets are typically collected using keywords related to certain elections through the Twitter application programming interface (API). Sentiment analysis is conducted on these data to extract sentiments towards a candidate or party.

"Sentiment analysis, also called opinion mining, aims to analyse people's opinions, sentiments, evaluations,

appraisals, attitudes, and emotions from written language towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" (Liu 2012). The objective of sentiment analysis is to identify or categorize the attitude expressed in a piece of text. Specifically, the attitude may be positive (favourable), negative (unfavourable), or neutral towards a subject (Nasukawa and Yi 2003).

There are mainly two types of approaches for sentiment analysis on election-related Twitter data. One is the lexicon-based approach and the other is the machine learning approach. The lexicon-based approach for sentiment analysis relies on a pre-defined sentiment lexicon and compares the presence or frequency of words in the given text with the words in the lexicon. For example, Ahmed, Jaidka, and Skoric (2016) predicted elections in four countries and compare the quality of predictions and the role of different technological infrastructures and democracies setups in these countries. For sentiment analysis, they applied a sentiment lexicon called SentiStrength to assign a positive score and a negative score to all the tweets relevant to a party. Because of the different internet connectivity and political environment, the prediction accuracy is different in the four countries.

The machine learning approach for sentiment analysis is generally divided into two sub-categories, supervised learning methods and unsupervised learning methods. Past work usually employed supervised learning methods that need good pre-labelled training datasets. For instance, Wang et al. (2012) developed a system for real-time analysis of tweet sentiment towards presidential candidates in the 2012 U.S. election. They trained a Naïve Bayes model on unigram features to classify the sentiment. Paul et al. (2017) collected geotagged tweets from a period of 6 months leading up to the U.S. presidential election in 2016 and classified the tweets towards either democratic or republic based on their sentiment at the county level. They trained the Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), Recurrent Neural Network (RNN), and FastText models by 1.6 million tweets from the Stanford Twitter Sentiment (STS) corpus.

**Table 1.** Variables in a generalization of presidential election forecasting models.

Variables	Explanations
<i>V</i>	Predicted vote
<i>E</i>	Economic growth, i.e. GDP growth, GNP growth
<i>P</i>	Support rate in the poll survey
...	Other predictive factors
<i>a, b &amp; c</i>	Coefficients and constant – estimated based on the regression results from past presidential elections

Despite the various kinds of work about predicting elections based on Twitter sentiment, Gayo-Avello pointed out several drawbacks of these works (Gayo Avello, Metaxas, and Mustafaraj 2011; Gayo-Avello 2011; Metaxas, Mustafaraj, and Gayo-Avello 2011; Gayo-Avello 2012a, 2012b). He criticized that these works were not predictions at all because their analyses were post hoc and past positive results of one election in one

country would not guarantee generalization. We should not assume all the tweets are trustworthy and ignore the rumours, propaganda, and misleading information. What is more, demographic information should be considered since the Twitter population is not a representative and unbiased sample of the voting population. Future work can take advantage of the well-established forecasting models from political science and integrate them with Twitter-based election prediction (Gayo-Avello 2013).

Besides the drawbacks discussed above, researchers also criticized that more than half of these types of studies are just 'analysis' but not 'prediction' (Gayo-Avello 2013; Le et al. 2017; Yaqub et al. 2017). Even in the studies that did predict election based on Twitter sentiment, they barely established robust models for the relationships between sentiments and predicted election results. Instead, positive sentiment was simply treated as voting for a certain candidate or party, vice versa. Furthermore, almost all of the previous studies conducted predictions nationwide instead of county level. However, the presidential election actually runs county by county and the geographical context does play important roles. Therefore, this study aims to substitute poll in traditional election forecasting models with Twitter-based sentiment and establish a more robust model. This model will examine the relationship between voting results and several predictive variables including support rate calculated by Twitter sentiment, economic growth at the county level.

### Twitter sentiments and poll data comparison

Past studies have shown the feasibility of substituting poll with the Twitter sentiment. O'Connor et al. (2010) found that the Twitter sentiment had correlations with public opinion by analyzing several surveys on consumer confidence and political opinion from 2008 to 2009. Beauchamp (2017) modeled state-level polls during the 2012 presidential election as a function of political tweets and found that Twitter-based measures can predict opinion polls. Anuta, Churchin, and Luo (2017) found that both the polls and Twitter were biased in the 2016 U.S. election. The poll had a small bias towards Hillary Clinton while Twitter had a slightly larger bias towards Donald Trump. Bovet, Morone, and Makse (2018) showed that the Twitter opinion trends in the 2016 U.S. presidential election followed the aggregated New York Times polls with remarkable accuracy by comparing them based on their proposed method.

## Data

### Twitter data

The Twitter data is obtained from (Poorthuis and Zook 2017) in their research about social media big data. The original data consists of all geotagged tweets sent from a bounding box around the state of Georgia from September 26<sup>th</sup>, 2016 (first presidential debate) to November 8<sup>th</sup>, 2016 (election day). Before using it in our research, we preprocess the data and extract the information including tweet id, user id, latitude, longitude, created time, tweet content.

By keyword search, this study selects only the tweets that mentioned either of the two presidential election candidates, Donald Trump and Hillary Clinton. Table 2 shows the keywords used to filter the tweets. Valid tweets are classified as Trump-related or Clinton-related. We leave out the tweets which mention both Trump and Clinton because of the potential ambiguity as to which candidate the sentiment of the tweet is about.

Because the tweets are sent from a bounding box, we remove the tweets outside the extent of Georgia and assign the county to each tweet based on its latitude and longitude. However, we notice that more than 2000 tweets are sent from a county named Wilkinson. After a careful examination, those tweets are recognized to be sent from the centroid of Georgia. Because when people geotag their tweets just in Georgia instead of a very specific location, Twitter will automatically assign the coordinates of the centroid of Georgia to their tweets. This situation cannot represent the actual Twitter users in Wilkinson county, therefore we remove those tweets for the sake of accuracy.

Figure 2 shows the tweet count by date. Four peaks can be found in it, which are right after the dates of the three presidential election debates (September 26<sup>th</sup>, October 9<sup>th</sup>, October 19<sup>th</sup>) and the final election day (November 8<sup>th</sup>). This makes sense that people tend to tweet more right after the events happen. In Figure 3, we can see that more than half of the tweets (approximately

**Table 2.** Keywords for filtering the tweets.

Presidential Election Candidates	Keywords
Donald Trump	realdonaldtrump donaldtrump donald trump republican gop
Hillary Clinton	hillaryclinton hillary clinton democratic democrat

60%) are sent during the night from 8:00 pm to 8:00 am in local time.

### Economic data

The theoretical framework of election forecasting in political science models takes economic growth as one of the two significant predictors. Past studies (Eisenberg and Ketcham 2004; Lacombe and Shaughnessy 2007) have shown that per capita personal income growth and unemployment rate influence vote shares at the county level. Therefore, in this study, in addition to the GDP growth rate, we also use per capita personal income growth rate, unemployment rate growth rate, to represent the economic growth at the county-level in the current administration of the incumbent party. They are obtained or calculated from the original data downloaded from official websites of the Bureau of Economic Analysis (BEA), U.S. Bureau of Labour Statistics (BLS). From the definitions on the BEA website, Gross Domestic Product (GDP) is the value of the goods and services produced by counties in Georgia (<https://www.bea.gov/resources/learning-center/what-to-know-gdp>), personal income is income people get from wages, proprietors' income, dividends, interest, rents, and government benefits (<https://www.bea.gov/resources/learning-center/what-to-know-income-saving>). From the definition on the BLS website, unemployment rate data comes from the Current Population Survey (CPS), the household survey, and is estimated based on a building-block approach for counties (<https://www.bls.gov/lau/lauov.htm>). In the original political science election forecasting models, economic growth, for instance, GDP growth, is the second-quarter growth in the election year. Since data at the county level are released annually, we use the growth rate from 2015 to 2016 to represent the local economic growth.

Table 3 shows information of the economic variables. All growth rates are calculated with Equation (2), in which  $GR$  represents a specific growth rate,  $E_{new}$  represents the end-year (e.g., 2016) value of the relevant economic variable,  $E_{old}$  represents the starting year (e.g., 2015) value of the relevant economic variable.

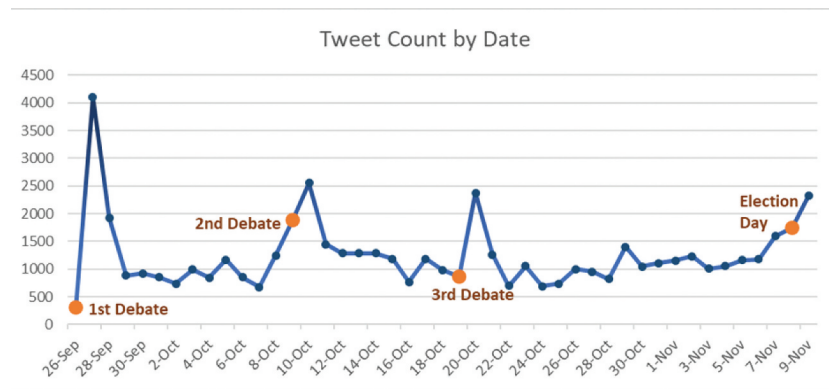


Figure 2. Tweets count by date.

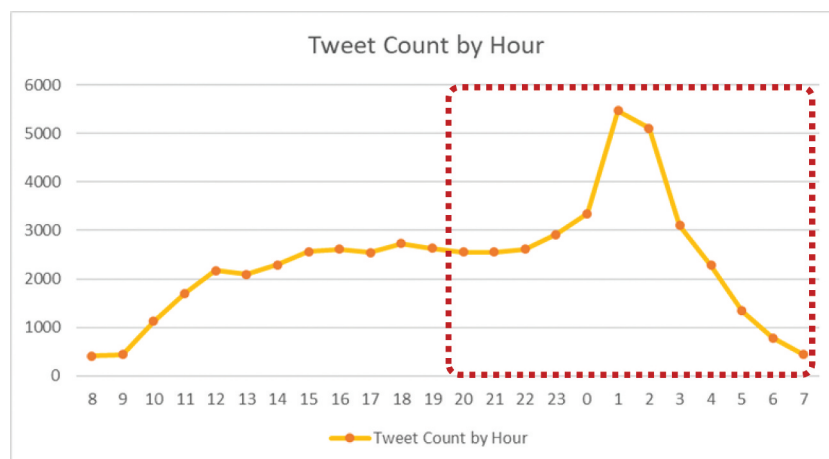


Figure 3. Tweets count by hour.

**Table 3.** Selected economic variables.

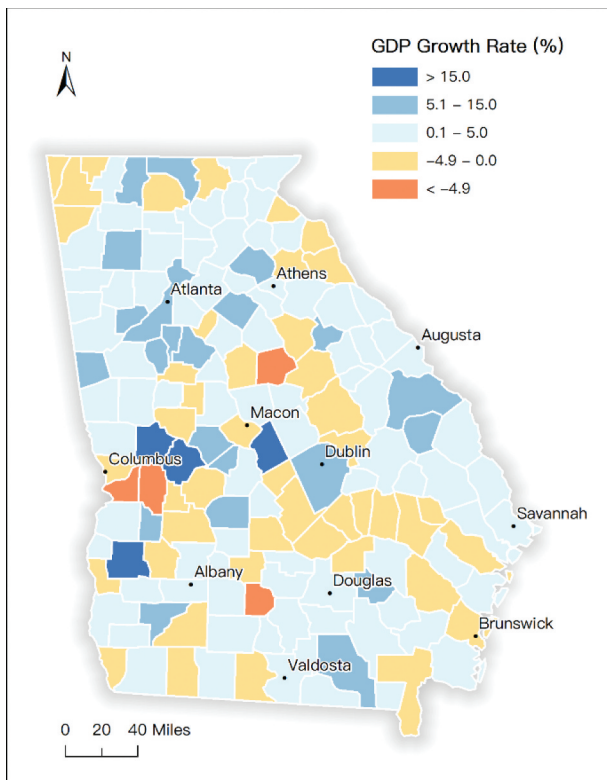
Variable Name	Descriptions	Time	Sources
GDP_GR	GDP Growth Rate	2015–2016	Bureau of Economic Analysis
PC_PI_GR	Per Capita Personal Income Growth Rate	2015–2016	Bureau of Economic Analysis
UnemployR_GR	Unemployment Rate Growth Rate	2015–2016	U.S. Bureau of Labour Statistics

Figures 4–6 reveal the uneven spatial distribution of the economic growth rates in the case study state.

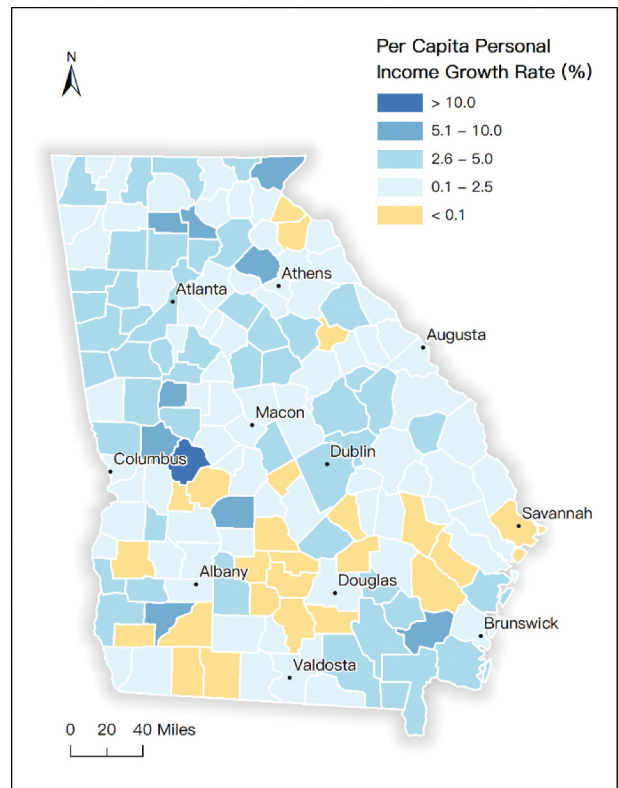
$$GR = \frac{E_{new} - E_{old}}{E_{old}} \quad (2)$$

**Voting data**

The actual voting data are used to train the models and to evaluate the model accuracy. Figure 7 shows the actual voting results for Clinton by county in Georgia. Following the same theoretical framework of the traditional political science model which predicts the voting outcome for the incumbent party, the models in this study also use the voting results for the candidate of the incumbent party (Clinton in this case) as the dependent variable.



**Figure 4.** GDP growth rate in Georgia from 2015 to 2016.



**Figure 5.** Per capita personal income growth rate in Georgia from 2015 to 2016.

**Method**

Figure 8 below illustrates the research design of this study, with colour-coded data and method components.

**Sentiment analysis of tweets**

After the data preparation and preprocessing process, the database contains a total of **57,912** tweets. The next step is to calculate the sentiment score for each tweet based on its content. The URLs, hashtag symbols (#), and mention symbols (@) are removed as they may be meaningless noise. The sentiment analysis in this study applies a Natural Language Processing (NLP) method. The method uses a deep learning technique and is developed by the NLP Group at Stanford University. The method builds up a representation of whole sentences based on the sentence structure and computes the sentiment based on how words compose the meaning of longer phrases (Socher et al. 2013). The NLP sentiment analysis method classifies sentiment score in five categories: very negative (score: 0), negative (score: 1), neutral (score: 2), positive (score: 3) and very positive (score: 4). Figure 9 below shows the average tweet sentiments towards Trump and Clinton by date.

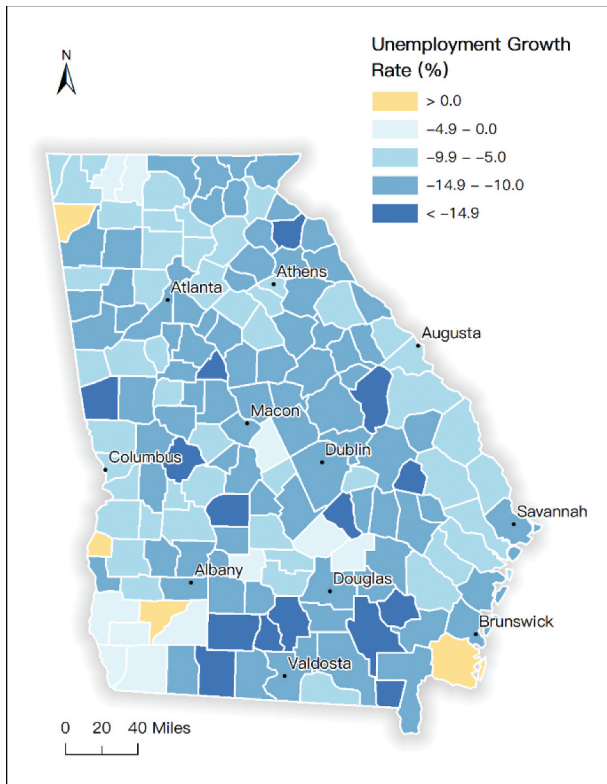


Figure 6. Unemployment rate growth rate in Georgia from 2015 to 2016.

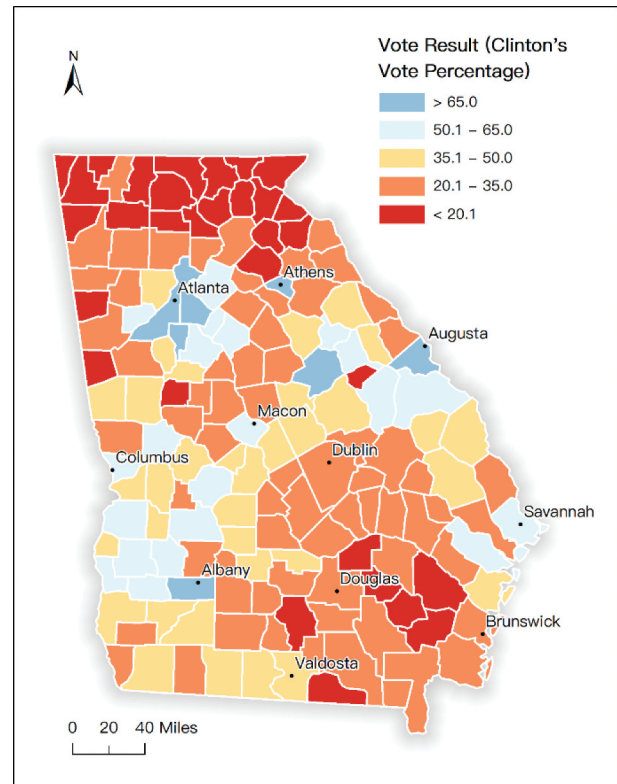


Figure 7. Percentages of votes for Clinton in Georgia in the 2016 presidential election.

To evaluate the performance of the sentiment analysis, we manually label a random sample of 100 tweets. The manual labels are obtained from six annotators and the average sentiment score for each tweet is taken as the ground truth. Table 4 is a confusion matrix of the sentiment analysis results. Table 5 presents the accuracy assessment of the sentiment analysis of the sample. For example, the precision of neutral sentiment is computed as:  $P(neu) = \frac{30}{5+30+22} = 0.5263$  which means the ratio of the number of tweets classified correctly as neutral to the total predicted neutral tweets; the recall of neutral sentiment is computed as:  $R(neu) = \frac{30}{5+30+6} = 0.7317$  which means the ratio of the number of tweets correctly classified as neutral to the number of known neutral

tweets; the F1 score is the harmonic mean between both precision and recall,  $F1(neu) = \frac{2 \cdot P(neu) \cdot R(neu)}{P(neu) + R(neu)}$ .

We can see that the model performs relatively better on negative and neutral sentiments than the positive sentiments. However, the accuracy of the model is not very good, with an overall accuracy of 57%. Because sentiment analysis is not the scientific contribution of this study, we leave for future research to employ more advanced methods for accuracy improvement.

### Twitter user extraction

A Twitter user can have multiple tweets with different sentiment scores in different locations. The assumption in this study is that one Twitter user represents one

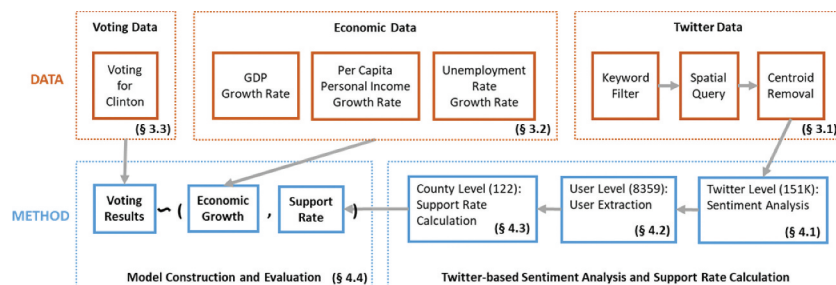


Figure 8. Research design.



**Table 4.** Confusion matrix of tweet sample sentiment analysis.

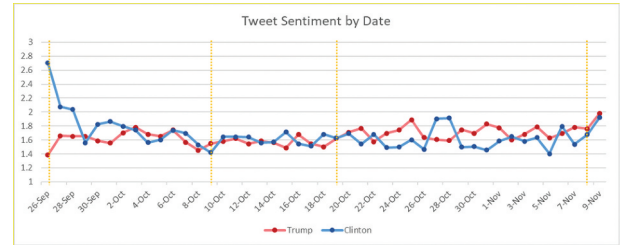
		Predicted		
		Positive	Neutral	Negative
Actual	Positive	4	5	1
	Neutral	5	30	6
	Negative	4	22	23

**Table 5.** Accuracy, precision, recall and F1-score of tweet sample sentiment analysis.

	Precision	Recall	F1
Positive	30.77	40.00	34.78
Neutral	52.63	73.17	61.22
Negative	76.67	46.94	58.23
Macro	53.36	53.37	51.41
Accuracy		57.00	

person in Georgia who should only have one vote in one county. Therefore, we need to extract the overall sentiment of a Twitter user towards a candidate from all valid tweets by this user. This step aggregates the sentiments from the tweet level to the user level. The resulting dataset records data for each Twitter user including a unique Twitter id, the user's home county, and the overall sentiment towards each candidate.

To accomplish this task, two independent scenarios are considered: 1) A Twitter user sends multiple tweets, some are Trump-related, some are Clinton-related; 2) A Twitter user sends tweets in multiple locations so that it is uncertain which is the home location. For the first scenario, the average of all the Trump-related sentiment is calculated and denoted as  $T_{sen}$ , the average of all Clinton-related sentiments is calculated and denoted as  $C_{sen}$ . The difference between the two,  $C_{sen} - T_{sen}$ , implies the likelihood of favouring each candidate. The greater the difference, the more likely that the user will vote for Clinton. The smaller the difference, the more likely that the user will vote for Trump. If the difference is 0, the sentiments towards both candidates are comparable, thus each candidate has 50% of the chance if we simplify the situation by ignoring the possibility of voting for the independent candidate. Instead of assuming the oversimplified binary relationship between the difference and voting results, the study applies fuzzy logic to model the relationship to account for uncertainties in the human decision-making process. Figure 10 is the fuzzy membership function used in this study. If the difference value is larger than 0.4 or smaller than  $-0.4$ , the support rate of this user for Clinton is 100% or 0; if the difference value is within the range  $[-0.4, 0.4]$ , a linear relationship is assumed and the possibility of supporting Clinton is calculated based on the equation:

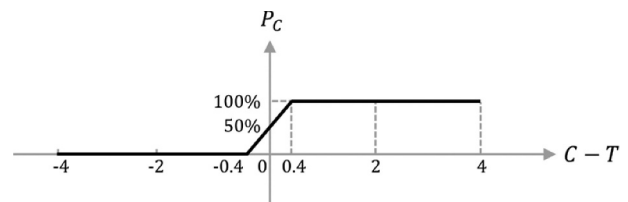
**Figure 9.** Change of average sentiments towards each candidate by date.

$$P_c = 1.25 \times (C_{sen} - T_{sen}) + 0.5 \quad (3)$$

The approach is generally applicable to other election forecasts, although the specific membership function can be flexible and should be decided with a good understanding of the problem at hand and the context of it.

For the second scenario, we assign a user to the county in which the most frequent location is found among the user's tweets between 8:00 pm and 8:00 am the next day. This is based on the reasoning that most people work during the daytime and stay at home during the night. If no single most frequent county can be found during night time, we use the most frequent county at all times as one's home county. In the end, this process converts the database from the tweet level (**57,912 tweets**) to the user level (**8346 users**). Each record in the database contains user id, the user's residence county, and the user's average sentiment towards each candidate. Figure 11 shows the number of Twitter users per county. 122 out of 159 counties in Georgia have at least one Twitter user and cities with a high population like Atlanta, Augusta, Columbus, Macon, Savannah are represented with the high number of Twitter users.

It is noteworthy that a bot check process was conducted at the beginning of the process. Due to the open structure of the Twitter platform, not only real humans can send tweets, some automated programs, known as bots, also send posts on Twitter for news sharing, advertising, or other reasons. The existence of these bots and the spam spread by them cause problems to the extraction of valid Twitter users or meaningful information in

**Figure 10.** The graph of fuzzy membership function.

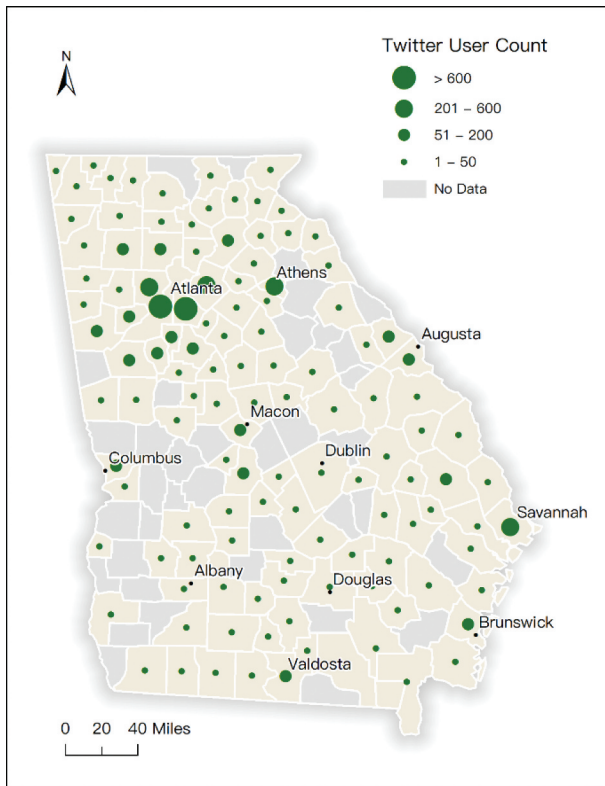


Figure 11. The number of Twitter users per county.

tweets (Chu et al. 2012; Guo and Chen 2014). Bessi and Ferrara (2016) found that the presence of social media bots can potentially distort public opinion and have a negative effect on the discussion about the presidential election on Twitter. In this research, only human accounts on Twitter can represent people in the real

world who can vote in the presidential election. Figure 12 shows the relationship between the number of tweets and the number of users, with both variables log-transformed. It shows that a small number of IDs are responsible for extremely large numbers of posts. However, after manually check, we find that IDs who sent a large number of tweets are actually real humans. Therefore, they are kept in the study.

### Twitter-based support rate calculation

In this step, the sentiments at the Twitter user level are aggregated to the county level. The supporting rate for the candidate of the incumbent party (Clinton in this election) in a county is calculated as the number of users whose sentiment is positive to Clinton divided by the total number of users in this county. Note that fuzzy logic is used so that partial memberships are also included in the calculation process. Because no Twitter users are identified in some counties, we do not include these counties in the final model. After this process, **8346 users** are aggregated to **122 counties**. Figure 13 shows the distribution of Clinton's Twitter-based support rate. We can see that there are more counties against Clinton than supporting her. Figure 14 illustrates the spatial distribution of the Twitter-based support rate for Clinton.

### Modelling results

A series of regression and classification models are constructed for election forecasting. For regression, the dependent variable is the percentage of votes for the candidate of the incumbent party (Clinton). For classification, the dependent variable is the binary voting results

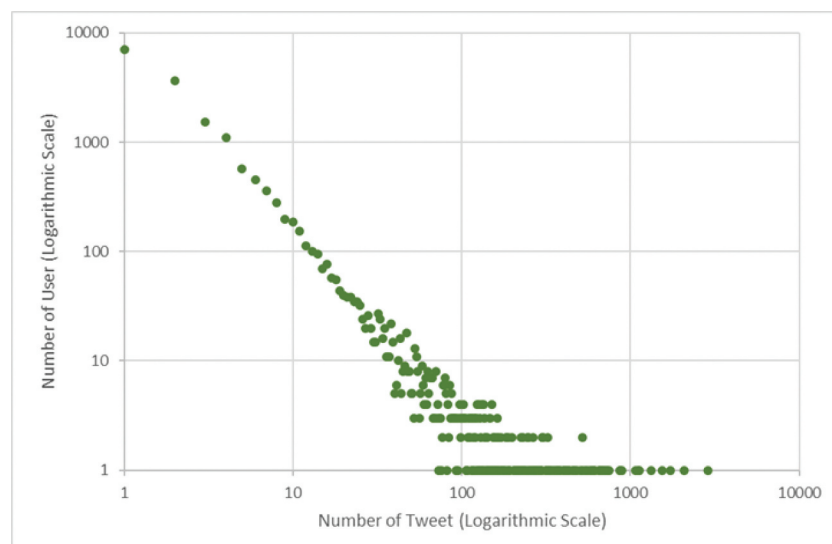


Figure 12. Number of users vs number of tweets.

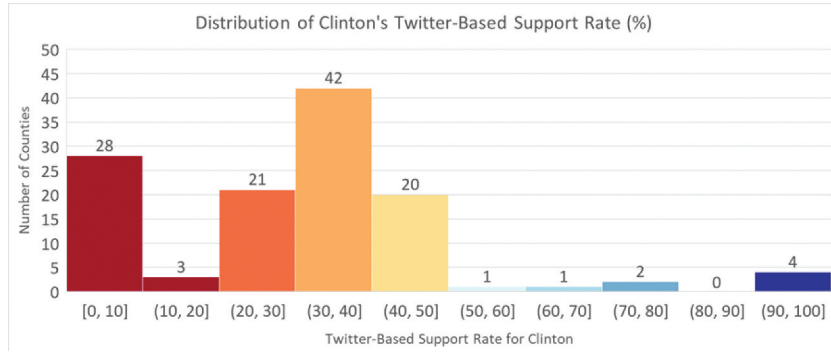


Figure 13. Distribution of Clinton’s Twitter-based support rate (%).

for Clinton. The independent variables for the models are the same, namely the Twitter-based support rate, GDP growth rate, per capita personal income growth rate, unemployment rate growth rate. We use the 10-fold cross-validation to train and validate the models.

For model performance evaluation, root mean squared error (RMSE) is used for the regression models, while accuracy, precision, recall, and F1-score are used for the classification models. The performance of these models is listed in Tables 6 and 7. K-Nearest Neighbours (K-NN) is the best model for regression while Decision Tree (DT) is the best for classification because it has the

highest precision, recall, and F1 score. K-NN regression has the RMSE value of 0.1526, which suggests a 15% difference between the predicted vote and actual vote.

Same as the definition of precision, recall, F1 score in section 4.1, DT classification has a precision of 70.43% which means the ratio of the number of counties classified correctly to the total predicted counties; recall of 67.50% which means the ratio of the number of counties correctly classified to the number of known counties; F1 score of 67.28% is the harmonic mean between both precision and recall; the accuracy of 81.28% means overall 81% of 122 counties are classified correctly.

Noted that the precision, recall, and F1 score here in Table 7 are the macro value which is computed as the average of precision, recall, or F1 score of the two classes in the classification models. For example, macro precision is computed as:  $P_{Macro} = \frac{P(VoteforTrump) + P(VoteforClinton)}{2}$ .

Table 8 lists the feature importance of DT classification models. Variable *Sen\_SupportR\_Clinton* is the Twitter-based support rate which ranks 2<sup>nd</sup> important feature in DT classification. This means the support rate calculated from Twitter sentiments has made a major contribution to the models and has the potential to predict elections in the future. Same as the explanation in section 3.2, variables *GDP\_GR*, *PC\_PI\_GR*, and *UnemployR* are growth rates of GDP, per capita personal

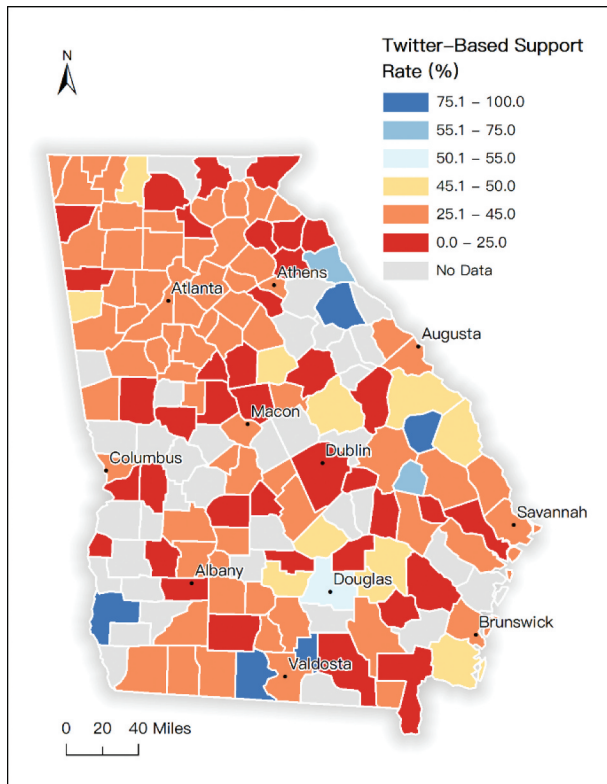


Figure 14. Twitter-based support rate for Clinton by county.

Table 6. Performance of the regression models.

Regression Model	RMSE
K-Nearest Neighbours	0.1526
Gradient Boosting Trees	0.1531
Lasso Regression	0.1539
Generalized Linear Model	0.1588
Linear Regression	0.1599
Ridge Regression	0.1599
Multi-layer Perceptron	0.1637
Random Forest	0.1641
Linear SVR	0.1948
Decision Tree	0.2059

**Table 7.** Performance of the classification models.

Classification Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)
Gradient Boosting Trees	82.88	62.12	58.50	58.23
Decision Tree	81.28	70.43	67.50	67.28
Logistic Regression	81.22	40.61	50.00	44.80
Kernel SVC	81.22	40.61	50.00	44.80
K-Nearest Neighbours	81.22	40.61	50.00	44.80
Random Forest	79.68	59.71	63.28	61.03
Multi-layer Perceptron	78.78	56.06	56.83	55.25
Gaussian Naive Bayes	78.72	53.73	53.67	52.18
Linear SVC	58.53	47.08	50.89	39.27

income, unemployment rate from 2015 to 2016 by counties in Georgia.

In order to apply the trained K-NN or DT models for forecasting a presidential election in the future, the Twitter-based support rate and the economic growth variables will need to be updated with the real data close to the election. This study applies it to the 2016 election just for a demonstration purpose. It should be noted that this is not a real prediction because the data have already been used to train the models, thus the results do not necessarily reflect the real prediction accuracy.

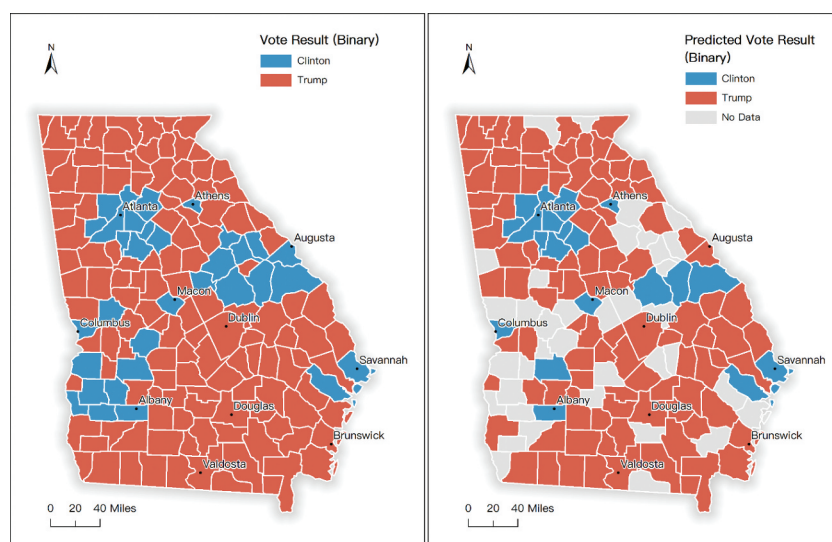
Figure 15 displays the estimated results versus the actual voting results. We can see the binary voting results by the DT classification model are quite consistent with the actual voting results. Tables 9 and 10 show the confusion matrix as well as accuracy, precision, recall,

and F1-score of the DT classification model. In the tables, the precision of voting for Trump is computed as the ratio of the number of counties classified correctly as voting for Trump to the total predicted counties in this class,  $P(Trump) = \frac{99}{99+4} = 0.9612$ ; recall of voting for Trump is computed as the ratio of the number of counties classified correctly as voting for Trump to the number of known counties in this class,  $R(Trump) = \frac{99}{99+0} = 1$ ; F1 score of voting for Trump is the harmonic mean between both precision and recall. The same calculations are applied for precision, recall, and F1 score of voting for Clinton. 96.72% accuracy means overall 97% of 122 counties are classified correctly.

The regression models have much lower accuracy, with an average of 15% deviation from the actual voting percentage. This is not surprising because an election prediction is more about telling the final results than telling the precise distribution of votes. It is very difficult, if not impossible, to be precise on the voting percentage. The more interesting question is whether these estimated percentages of votes can be aggregated to tell the correct final result. Indeed, when we aggregate the estimates

**Table 8.** Feature importance of DT classification models.

Variable	Feature Importance
Sen_SupportR_Clinton	0.40
GDP_GR	0.46
PC_PI_GR	0.12
UnemployR_GR	0.02

**Figure 15.** Actual votes versus estimated voting results in Georgia for the 2016 presidential election.

**Table 9.** Confusion matrix of DT classification model.

		Predicted	
		Vote for Trump	Vote for Clinton
Actual	Vote for Trump	99	0
	Vote for Clinton	4	19

**Table 10.** Accuracy, precision, recall and F1-score of DT classification model.

	Precision	Recall	F1
Vote for Trump	96.12	100.00	98.02
Vote for Clinton	100.00	82.61	90.48
Macro	98.06	91.30	94.25
Accuracy		96.72	

from the county to the state level, the final result correctly indicate that lower than 50% of the votes go to the Democratic candidate, which is consistent with the actual result.

Noted that the metrics of DT classification are different in [Tables 7 and 10](#). Because when choosing the best model, we use 10-fold cross-validation to train and evaluate the model performance. In this process the whole dataset is divided into 10 pieces, 9 pieces are used for training while 1 piece is used for validation each time and repeated for 10 times. All the metrics in [Table 7](#) are the average value of the 10 times. After finding Decision Tree as the best model, in order to present the estimated voting results, since there are no average parameters for DT classification from 10 fold cross-validation, we have to fit the model with all the data and then use the model to do the estimation. Therefore, we need to emphasize again this is not a real prediction but for a demonstration purpose.

## Discussions and conclusion

This research provides a new perspective to think about forecasting the presidential election. As an innovative contribution, the study integrates a classical political science election prediction model and the emerging approach of using social media sentiments for prediction. By doing so, the proposed modelling strategy can extend the national level prediction by a classical model in political science to the county level, which is the finest geographical unit of voting results. Previous studies of the spatially fine-grained election models were essentially post hoc estimations and not prediction models, their key predictors were demographic variables and socio-economic variables. Because the direction and strength of relationships between these variables and the voting results towards the candidate of the incumbent party can change significantly from one election to

the next, the established models in those studies were not applicable for the next election. In comparison, the modelling strategy developed in this study subscribes to the theoretical basis in political science and adopts only the variables with relatively stable relationships towards votes. Specifically, these are key economic growth indicators and sentiments. As the relationships are generally applicable over time, the established models can be used for prediction in the future. Another contribution of the proposed modelling strategy is applying various machine learning methods to estimate model parameters in order to avoid oversimplified assumptions of linear relationships. To evaluate the performance of the developed models for presidential election prediction, it can be tested with the upcoming 2020 U.S. presidential election.

As the first study that draws on thoughts and methods in political science and in the big data research for election prediction, the current research is still preliminary with many limitations. Many research avenues can be explored for future improvements. Discussed here are some good examples that call for immediate attention. First, since the Twitter data collected by the API is only about 1% of all tweets, some of the counties in our study are lack of Twitter data. The problem is more severe in rural areas that have sparse twitter data in the first place. To improve the data coverage, future studies could use Twitter firehose data for better coverage. In addition, future research efforts can be made to interpolate sentiment data for areas that have no data. Second, the model performance can be sensitive to sentiment analysis results. Instead of using built machine learning models developed by a third party, future research can train their own election-related Twitter sentiment classification models which require more reliable labelling of the training data. Third, our study ignores the tweets which mention both candidates. In the future, it is necessary to employ more sophisticated methods to classify the tweets directly instead of relying on keywords searching. Fourth, there is room for improvement to infer the home location of a user. There is a strand of research about inferring home locations of the users (Cheng, Caverlee, and Lee 2010; Hecht et al. 2011; Huang, Cao, and Wang 2014). Simply considering the most frequent night location as a person's home location is intuitively reasonable but also crude. Last, we should not ignore the issue of bias of Twitter data or any other types of social media data. Not all people in a specific study area have Twitter accounts, neither are they all active Twitter users. Past studies have shown that the Twitter population is a highly non-uniform sample of the local population with regard to gender, race, and age (Mislove et al. 2011). The male Twitter users are

more likely to post political tweets (Barberá and Rivero 2015). Future research is in demand to evaluate the bias of Twitter data and to develop methods to account for such bias in election forecasting models.

In short, the study presents the first step towards a promising approach to forecasting elections in the United States. Sentiments gleaned from social media data will be used to surrogate for poll data. As the popularity of social media keeps growing, foreseeable is the increasing tendency that people express their opinions on Twitter or other social media platforms. Thus, the proposed election forecasting approach has great application potential in the future.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Ruowei Liu  <http://orcid.org/0000-0001-9495-366X>  
 Xiaobai Yao  <http://orcid.org/0000-0003-2719-2017>  
 Chenxiao Guo  <http://orcid.org/0000-0001-9200-6570>  
 Xuebin Wei  <http://orcid.org/0000-0003-2197-5184>

## References

- Abramowitz, A. 2001. "The Time for Change Model and the 2000 Election." *American Politics Research* 29 (3): 279–282. doi:10.1177/1532673X01293004.
- Ahmed, S., K. Jaidka, and M. M. Skoric. 2016. "Tweets and Votes: A Four-country Comparison of Volumetric and Sentiment Analysis Approaches." In *Tenth International AAAI Conference on Web and Social Media*, Cologne, Germany.
- Anuta, D., J. Churchin, and J. Luo. 2017. "Election Bias: Comparing Polls and Twitter in the 2016 Us Election." *ArXiv Preprint ArXiv:1701.06232*.
- Barberá, P., and G. Rivero. 2015. "Understanding the Political Representativeness of Twitter Users." *Social Science Computer Review* 33 (6): 712–729. doi:10.1177/0894439314558836.
- Beauchamp, N. 2017. "Predicting and Interpolating State-level Polls Using Twitter Textual Data." *American Journal of Political Science* 61 (2): 490–503. doi:10.1111/ajps.12274.
- Birmingham, A., and A. Smeaton. 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)*, Chiang Mai, Thailand, 2–10.
- Bessi, A., and E. Ferrara. 2016. "Social Bots Distort the 2016 US Presidential Election Online Discussion." *First Monday* 21: 11–17.
- Bovet, A., F. Morone, and H. A. Makse. 2018. "Validation of Twitter Opinion Trends with National Polling Aggregates: Hillary Clinton Vs Donald Trump." *Scientific Reports* 8 (1): 8673. doi:10.1038/s41598-018-26951-y.
- Burnap, P., R. Gibson, L. Sloan, R. Southern, and M. Williams. 2016. "140 Characters to Victory?: Using Twitter to Predict the UK 2015 General Election." *Electoral Studies* 41: 230–233. doi:10.1016/j.electstud.2015.11.017.
- Campbell, J. E., L. L. Cherry, and K. A. Wink. 1992. "The Convention Bump." *American Politics Quarterly* 20 (3): 287–307. doi:10.1177/1532673X9202000302.
- Campbell, J. E., and K. A. Wink. 1990. "Trial-Heat Forecasts of the Presidential Vote." *American Politics Quarterly* 18 (3): 251–269. doi:10.1177/1532673X9001800301.
- Ceron, A., L. Curini, and S. M. Iacus. 2015. "Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence from the United States and Italy." *Social Science Computer Review* 33 (1): 3–20. doi:10.1177/0894439314521983.
- Ceron, A., L. Curini, S. M. Iacus, and G. Porro. 2014. "Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens' Political Preferences with an Application to Italy and France." *New Media & Society* 16 (2): 340–358. doi:10.1177/1461444813480466.
- Cheng, Z., J. Caverlee, and K. Lee. 2010. "You are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users." In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Toronto, ON, Canada, 759–768.
- Chu, Z., S. Gianvecchio, H. Wang, and S. Jajodia. 2012. "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?" *IEEE Transactions on Dependable and Secure Computing* 9 (6): 811–824. doi:10.1109/TDSC.2012.75.
- Chung, J. E., and E. Mustafaraj. 2011. "Can Collective Sentiment Expressed on Twitter Predict Political Elections?" *AAAI* 11: 1770–1771.
- Eisenberg, D., and J. Ketcham. 2004. "Economic Voting in US Presidential Elections: Who Blames Whom for What." *The B.E. Journal of Economic Analysis & Policy* 4: 1.
- Gayo-Avello, D., P. T. Metaxas, and E. Mustafaraj. 2011. "Limits of Electoral Predictions Using Twitter." In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain. Association for the Advancement of Artificial Intelligence.
- Gayo-Avello, D. 2011. "Don't Turn Social Media into another 'Literary Digest' poll." *Communications of the ACM* 54 (10): 121–128. doi:10.1145/2001269.2001297.
- Gayo-Avello, D. 2012a. "I Wanted to Predict Elections with Twitter and All I Got Was This Lousy Paper—A Balanced Survey on Election Prediction Using Twitter Data." *ArXiv Preprint ArXiv:1204.6441*.
- Gayo-Avello, D. 2012b. "No, You Cannot Predict Elections with Twitter." *IEEE Internet Computing* 16 (6): 91–94. doi:10.1109/MIC.2012.137.
- Gayo-Avello, D. 2013. "A Meta-analysis of State-of-the-art Electoral Prediction from Twitter Data." *Social Science Computer Review* 31 (6): 649–679. doi:10.1177/0894439313493979.
- Grover, P., A. K. Kar, Y. K. Dwivedi, and M. Janssen. 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. doi:10.1016/j.techfore.2018.09.009.
- Guo, D., and C. Chen. 2014. "Detecting Non-personal and Spam Users on Geo-tagged Twitter Network." *Transactions in GIS* 18 (3): 370–384. doi:10.1111/tgis.12101.

- Hecht, B., L. Hong, B. Suh, and E. H. Chi. 2011. "Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, BC, Canada, 237–246.
- Holbrook, T. M., and J. A. DeSart. 1999. "Using State Polls to Forecast Presidential Election Outcomes in the American States." *International Journal of Forecasting* 15 (2): 137–142. doi:10.1016/S0169-2070(98)00060-0.
- Huang, Q., G. Cao, and C. Wang. 2014. "From Where Do Tweets Originate? A Gis Approach for User Location Inference." In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, Dallas, Texas, 1–8.
- Lacombe, D. J., and T. M. Shaughnessy. 2007. "Accounting for Spatial Error Correlation in the 2004 Presidential Popular Vote." *Public Finance Review* 35 (4): 480–499. doi:10.1177/1091142106295768.
- Le, H., G. R. Boynton, Y. Mejova, Z. Shafiq, and P. Srinivasan. 2017. "Bumps and Bruises: Mining Presidential Campaign Announcements on Twitter." In *Proceedings of the 28th ACM Conference on Hypertext and Social Media - HT '17*, 215–224. doi:10.1145/3078714.3078736.
- Lewis-Beck, M. S., and T. W. Rice. 1982. "Presidential Popularity and Presidential Vote." *Public Opinion Quarterly* 46 (4): 534–537. doi:10.1086/268750.
- Liu, B. 2012. "Sentiment Analysis and Opinion Mining." *Synthesis Lectures on Human Language Technologies* 5 (1): 1–167. doi:10.2200/S00416ED1V01Y201204HLT016.
- Metaxas, P. T., E. Mustafaraj, and D. Gayo-Avello (2011). "How (Not) to Predict Elections." In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference On Social Computing*, Boston, Massachusetts, USA, 165–171.
- Mislove, A., S. Lehmann, -Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. 2011. "Understanding the Demographics of Twitter Users." *Icwsn* 11 (5th): 25.
- Nasukawa, T., and J. Yi. 2003. "Sentiment Analysis: Capturing Favorability Using Natural Language Processing." In *Proceedings of the 2nd International Conference on Knowledge Capture*, Sanibel Island, FL, USA, 70–77.
- O'Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." *Icwsn* 11 (122–129): 1–2.
- Paul, D., F. Li, M. K. Teja, X. Yu, and R. Frost. 2017. "Compass: Spatio Temporal Sentiment Analysis of US Election What Twitter Says!" In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1585–1594. doi:10.1145/3097983.3098053.
- Poorthuis, A., and M. Zook. 2017. "Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media." *Journal of Urban Technology* 24 (4): 115–135. doi:10.1080/10630732.2017.1335153.
- Sang, E. T. K., and J. Bos. 2012. "Predicting the 2011 Dutch Senate Election Results with Twitter." In *Proceedings of the Workshop on Semantic Analysis in Social Media*, Avignon, France, 53–60.
- Socher, R., A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. 2013. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, 1631–1642.
- Swamy, S., A. Ritter, and M.-C. de Marneffe. 2017. "'I Have a Feeling Trump Will Win .... .. .': Forecasting Winners and Losers from User Predictions on Twitter." *ArXiv:1707.07212 [Cs]*. <http://arxiv.org/abs/1707.07212>
- Wang, H., D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. 2012. "A System for Real-time Twitter Sentiment Analysis of 2012 Us Presidential Election Cycle." In *Proceedings of the ACL 2012 System Demonstrations*, Jeju Island, Korea, 115–120.
- Wang, L., and J. Q. Gan. 2017. "Prediction of the 2017 French Election Based on Twitter Data Analysis." In *2017 9th Computer Science and Electronic Engineering (CEECE)*, 89–93. doi:10.1109/CEECE.2017.8101605.
- Yaqub, U., S. A. Chun, V. Atluri, and J. Vaidya. 2017. "Analysis of Political Discourse on Twitter in the Context of the 2016 US Presidential Elections." *Government Information Quarterly* 34 (4): 613–626. doi:10.1016/j.giq.2017.11.001.