IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

A SYSTEMATIC REVIEW OF PREDICTING ELECTIONS BASED ON SOCIAL MEDIA DATA

[1]K SUPARNA, [2]R. VENKATA NARESH

[1](Assistant Professor), MCA, DANTULURI NARAYANA RAJU COLLEGE(A) PG COURSES, BHIMAVARAM ANDHRA PRADESH

[2]MCA, scholar, DANTULURI NARAYANA RAJU COLLEGE(A) PG COURSES, BHIMAVARAM ANDHRA PRADESH

ABSTRACT

The way politicians communicate with the electorate and run electoral campaigns was reshaped by the emergence and popularization of contemporary social media (SM), such as Facebook, Twitter, and Instagram social networks (SNs). Due to the inherent capabilities of SM, such as the large amount of available data accessed in real time, a new research subject has emerged, focusing on using the SM data to predict election outcomes. Despite many studies conducted in the last decade, results are very controversial and many times challenged. In this context, this article aims to investigate and summarize how research on predicting elections based on the SM data has evolved since its beginning, to outline the state of both the art and the practice, and to identify research opportunities within this field. In terms of method, we performed a systematic literature review analysing the quantity and quality of publications, the electoral context of studies, the main approaches to and characteristics of the successful studies, as well as their main strengths and challenges and compared our results with previous reviews. We identified and analysed 83 relevant studies, and the challenges were identified in many areas such as process, sampling, modelling, performance evaluation, and scientific rigor. Main findings include the low success of the most-used approach, namely volume and sentiment analysis on Twitter, and the better results with new approaches, such as regression methods trained with traditional polls. Finally, a vision of future research on

integrating advances in process definitions, modelling, and evaluation is also discussed, pointing out, among others, the need for better investigating the application of state-of-the-art machine learning approaches.

1.INTRODUCTION

Social media (SM) has played a central role in politics and elections throughout this decade. We have entered a new era mediated by SM in which politicians conduct permanent campaigns without geographic or time constraints, and additional information about them can be obtained not only by the press but also directly from their profiles on social networks (SNs) and through other people sharing and amplifying their voices on SM. In this new scenario, SM is used extensively in electoral campaigns [1], and an online campaign's success can even decide elections. In practice recent examples of SM engagement and electoral success include the 2016 U.S. residential election, when Donald Trump focused his campaign on free-media marketing [2], and the 2018 Brazilian presidential election, when the candidate with more SM engagement but little exposition on traditional media was elected [3].

Moreover, in some way, it is possible to measure how a politician's message is spreading over SM and try to estimate how much attention a candidate is receiving or how many people are talking about a candidate. Thus, considering a large amount of data available in real time and the low cost of their acquisition, combined with the advances of techniques for processing them, a new research subject has emerged, focusing on using the SM data to predict election outcomes.

Only 2 years after Twitter and Facebook's launch for the general public, studies to predict elections based on the SM data started to be published: Tilton [4] can be considered a preliminary study focused on student elections, published in 2008. In addition, two studies published in 2010 at the same forum, Tumasjan *et al*. [5] and O'Connor *et al*. [6], are considered seminal studies regarding predicting political elections based on SM. The former presented an approach based on the volume counting of posts on Twitter (tweets), and the latter was based on the sentiment extracted from those tweets.

One decade after Tumasjan's and O'Connor's seminal studies had claimed promising results, several initiatives focused on predicting elections worldwide, such as in Europe [7], [8], Asia [9], [10], Latin America [11], [12], Africa [13], [14], and the USA [15]–[17], just to cite some. These studies presented a variety of methods, were applied in many different electoral scenarios, used different SNs as information source, and had different outcomes. Many studies claimed very positive results, others challenged the predictive power of SM, and even the same study may achieve positive results in one context and negative results in another context [18].

Thus, there is not yet a common perspective on the literature or well-established methods, processes, and tools for predicting election results based on the SM data. Moreover, even the SM context has changed over the years. For example, Face book surpassed the number of active users of Twitter, and new SN has emerged, such as Instagram.

In this context, this article aims to give a thorough review and investigation of the state of both the art and practice of predicting election outcomes based on the SM data and identify key research challenges and opportunities in this field. We systematically reviewed 83 studies from 2008 to 2019, identify the context of studies, main models, strengths, and challenges of this new area, as well as the main characteristic present on successful studies, and deeply discuss future directions

The remainder of this article is organized as follows: Section II presents the background and previous studies related to this work as well as an analysis of the main points of similar comparative studies. In Section III, we present the review method and procedure employed in this study, followed by Section IV, which provides an overall summary of the selected studies and assesses their quality. In Section V, we discuss the answers to three of the predefined research questions regarding the electoral context of studies, main approaches, and main characteristics of successful studies. Section VI answers the last research question regarding the main strengths and challenges, summarizing the results, and ends with a discussion about future directions. Section VII presents a comparison with previous works and reviews the limitations of this study, followed by Section VIII, which concludes and summarizes the outcomes.

2. EXISTING SYSTEM

In 2013, Kalampokis *et al*. [29] presented a systematic review aiming to understand the predictive power of SM, not only in the electoral context. By analysing 52 studies, 11 regarding election predictions, they identified that main approaches were based on volume, sentiment, and user profiling.

In addition, the use of predictive analysis using linear regression was identified, but not on the studies related to the political context. In addition, they verified that 40% of studies that had used sentiment-related variables challenged SM predictive power, i.e., was not successful, and this number increased to 65% in the case of lexicon-based approaches.

Finally, they emphasized the lack of predictive analytics evaluation and controversial results of electoral predicting studies. In the same year, Gayo-Avello [30] presented a study that we consider the first review specifically on predicting elections with SM, focused on Twitter. By analysing ten previous studies from 2010 to 2013, he concluded that “the presumed predictive power regarding electoral prediction has been somewhat exaggerated.” Moreover, as in [29], he identified volume and sentiment analysis as main approaches and the need to use more up-to-date methods for sentiment analysis. In addition, he expanded the list of challenges, such as the dependency of arbitrary decisions made by researchers regarding keywords, parties, candidates and selection of the data collection period, and problems related to Twitter, such as demographic and self-selection bias, and bias related to spam, misleading propaganda, and astroturfing. He ended the study pointing out that regression models may be a future direction.

In 2015, studies from Prada [31] and O’Leary [32] presented in general lines the main approaches for predicting using Twitter in many different domains, and briefly described a few studies related to election predictions (2 and 11 studies, respectively). In 2018, Kwak and Cho [33] presented the results of a survey including 69 papers that supported the argument that SM can be used in understanding political agenda, rather than in election forecast.

Ultimately, most recent studies [34], [35] presented limited non-systematic surveys, both analysing 13 papers, adding some arguments to the original review from Gayo-Avello [30]. Koli *et al*. [34] argued that prediction using Twitter can have better results in developed countries, due to a higher literacy rate and internet access, than in developing countries. In addition, Bilal *et al*. [35] considered the challenges of sentiment analysis in languages other than English. Despite these new arguments, recent studies fail to identify novel approaches as well as approaches using SM other than Twitter and Facebook.

DISADVANTAGES

1) Data uncover is the main weakness in the existing system.

2) The system doesn't have a technique to test and train for large scale data sets

3. PROPOSED SYSTEM

The proposed system aims at identifying the electoral contexts being studied, such as the year and country in which the election took place and the type of election. This question is intended to ascertain whether the studies are best suited or paying attention to any particular electoral context.

The objective of this proposed system is to identify the main approaches used, their main characteristics, how they are modelled and applied to predict elections, and what are the metrics used to assess their performance.

He objective of this proposed system is to identify the main characteristics of allegedly successful studies in order to identify in which specific contexts, which approaches, and which factors yield effective results.

After studying the context, approaches, and characteristics of successful studies, the answer to this question aims to summarize the main perceived strengths, weaknesses, challenges, and opportunities in this new research area to guide future research.

ADVANTAGES

Unique studies include approaches based on prediction market, cluster detection, centrality score, statistical physics of complex networks, and analysis of groups of supporters, solely or in combination with previously described approaches.

The system performed statistical tests on results to verify whether they were statistically significant.

4. OUTPUT SCREENS

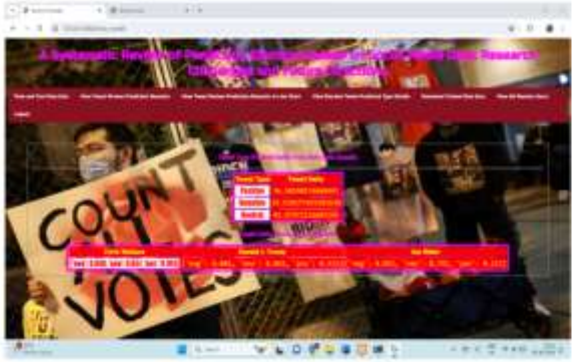
HOME PAGE

USER REGISTER PAGE

ELECTION PREDICTION

SERVICE PROVIDER LOGIN

ELECTION PREDICTION



5. CONCLUSION

This study collected more than 500 articles, 90 of which were focused on predicting elections based on SM data, investigating, and summarizing how this new research field has evolved since 2008. Among these studies, 83 are primary studies aiming at predicting elections and seven are surveys or reviews of past studies. The results show that the number of publications in this area is increasing and research is spread across 28 countries fro all continents. Nevertheless, there cannot yet be found any prominent researchers, research groups, or clusters performing sustainable research in the area. In addition, there was no identification of a common well-known forum for publication on this subject, and results are spread across many forums.

Regarding electoral contexts, most studies were performed in the context of a unique election, which may impact the results' validity. In addition, most were related to presidential elections at a national level with few candidates. Moreover, the most studied scenario was the U.S. presidential scenario, which can impact generalization due to its specificity.

Considering the main models used, we found that most studies used the approach of volume/sentiment analysis only on Twitter, in a variety of data collection approaches. We also found that regression and time series analysis is increasing, using multiple SNs, in addition to some supporting approaches, such as profile or post interactions and topic analysis.

By combining studies' characteristics and success we found that, despite being the most used approach, volume/sentiment does not presenting success rates, which is consistent with the conclusions of previous surveys. Thus, approaches such as regression or based on profile/posts interactions may be better to investigate and improve; even totally new approaches, such as one based on statistical physics of complex networks, may be tested. Finally, studies based on Twitter achieved significantly lower success rates than studies based on other SNs, such as Face book. Surprisingly, no studies based on Instagram were found.

Moreover, as main challenges, we identified issues in four areas. Regarding processes, we highlight the lack of well defined, replicable, and generalizable processes, and lack of prediction capabilities during the campaign. In sampling, issues are mainly related to the fact that SNs and Twitter data do not represent representative samples, and studies were performed with many arbitrary data collection choices. Regarding modeling, we found difficulties crossing data from multiple networks, the high susceptibility to volume manipulation, the lack of use of state-of-the-art ML techniques and technical modeling weaknesses. And considering performance evaluation and scientific rigor of studies, the lack of statistical analysis of results and of meaningful comparison with related works are also main issues.

Finally, the study presented the authors' point of view on the future directions of predicting elections using SM data in three axes: process definitions, model definitions and sampling, and study evaluation. As main directions, we highlight the need for repeatable processes based on well-known methodologies, for example, CRISP-DM or SEMMA; the use of state-of-the art methods for regression based on machine learning that can combine data from multiple SNs, such as ANN; and the use of statistical tests for results evaluation, such as Wilcoxon signed rank test and others.

The results from this review contribute to the research field by providing the academic community, as well as practitioners, with a better understanding of the research landscape and by identifying some of the gaps in the area that open up opportunities for future research. In addition to future directions presented, this literature review may also be extended in certain ways: a search extension may be performed to expand the search strategy and number of sources, thereby performing a broader study; a temporal update can be implemented without making modifications to the protocol, to expand the timeframe and co pare results over time; and finally, both approaches can be combined.

6. REFERENCE

1] A. Jungherr, "Twitter use in election campaigns: A systematic literature review," *J. Inf.Technol. Politics*, vol. 13, no. 1, pp. 72–91, Jan. 2016.

[2] P. L. Francia, "Free media and Twitter in the 2016 presidential election: The unconventional campaign of Donald Trump," *Social Sci. Comput. Rev.*, vol. 36, no. 4, pp. 440–455, Aug. 2018.

[3] K. Brito, N. Paula, M. Fernandes, and S. Meira, "Social media and presidential campaigns–preliminary results of the 2018 Brazilian presidential election," in *Proc. 20th Annu. Int. Conf. Digit. Government Res.*,Jun. 2019, pp. 332–341.

[4] S. Tilton, "Virtual polling data: A social network analysis on a student government election," *Webology*, vol. 5, no. 4, pp. 1–8, 2008.

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. '. M.Welpe, "Predicting Elections with Twitter: What 140 characters reveal about political sentiment," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, 2010,pp. 1–8.

[6] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith,"From tweets to polls: Linking text sentiment to public opinion timeseries," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, 2010,pp. 1–8.

[7] E. Sang and J. Bos, “Predicting the 2011 Dutch senate election results with Twitter,” in *Proc. Workshop Semantic Anal. Social Media, 2012*,pp. 53–60.

[8] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, “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 Soc.*, vol. 16, no. 2, pp. 340–358, Mar. 2014.

[9] K. Singhal, B. Agrawal, and N. Mittal, “Modeling Indian general elections: Sentiment analysis of political Twitter data,” in *Information Systems Design and Intelligent Applications* (Advances in Intelligent Systems and Computing). New Delhi, India: Springer, 2015.

[10] N. Dwi Prasetyo and C. Hauff, “Twitter-based election prediction in the developing world,” in *Proc. 26th ACM Conf. Hypertext social media (HT)*, 2015, pp. 149–158.