Machine Learning Based Prediction of Croatian 2017 Local Elections

A.Kišić* and B.Kliček*
* University of Zagreb,Faculty of Organization and Informatics, Varaždin, Croatia
{akisic, bklicek}@foi.hr

Abstract **- Internet development enabled intensive applications of new ways of communication. In today's digital age, social media became fundamental mean of communication. Political parties are increasingly using social media for the purpose of advertising and voters` mobilization. An extensive literature review has identified multiple benefits of social media usage, such as gaining publicity, spreading messages and mobilizing voters, but also the need to monitor the content that is published, and moreover to analyze the impact of that content on potential voters. This paper examines the influence of political candidates' activities on social media on elections outcome. By means of decision tree methodology, predictive model of election outcome is developed based on dataset consisting of candidates` characteristics, but also data of the candidates` activity on the social network Facebook. The predictive model is developed on dataset consisting of candidates in the local elections at large Croatian cities held in Croatia in 2017. The research identified the most important factors of political communication on the social network Facebook for election outcome and provides guidelines for the effective use of Facebook in political campaigns**

Keywords **- machine learning; predictive model; social network data; political elections**

I. INTRODUCTION

Political communication on social media is the focus of this research. According to a study of Pew Research, vast majority of Millennials (86%) say they use social media [1], thus becoming an important strategic tool for public relations in political campaigns. In previous surveys [2] a need for monitoring the content which is published on social media has been identified, as well as for analyzing the impact of such content on the target audience. Therefore, the purpose of this research is to determine the following: (i) whether or not social media are an efficient tool of political communication, (ii) how to measure the effects of social media on the election outcome? Today, in the age of a large amount of data, it is possible to achieve that. The approach to data has enabled a shift from the mere assessments made through data based decisions and procedures. What data and for what purpose have been used in the previous research conducted? The literature review has shown that there are surveys examining the use of specific data from social media for the purpose of predicting the election outcome. DiGrazia and associates claim that the election results can be predicted based on the activity on social media [3]. In the vast majority of such papers, data collected from Twitter were used, but in some of the papers other data sources were used, as well, such as Facebook [4]. Data obtained from Twitter are mostly used, and several studies have been conducted for the purpose of examining the role of Twitter in the election process Portugal [5], United Kingdom [6]) and USA [7]. In the previous surveys, data obtained from Twitter were used predominantly for convenience as they are easily accessible. Moreover, most of the previous approaches relied on very simple measurement data, such as the number of tweets, as well as the number of followers the candidate has gained. While, on one hand, the results achieved by [8], show that tweets can predict the election results by examining the frequency of tweets during the campaign, the results obtained by [9], on the other hand, show that tweets are not a good predictor. [10] write about the two predictive models. One of them, described by [8], was based on counting the number of tweets mentioning each of the candidates. The other group of predictive models was elaborated by [11], who describe the calculation methodology of a sense of the topic addressed on Twitter by implementing the sentiment analysis. The predictive and sentiment analysis based on the mere number of entities participating in tweeting can be improved significantly, thus becoming almost as good and even a better prediction than the traditional opinion polling [12]. However, [13] suggest that there is a need for a more advanced analyses and data collection methodology, as well as the filtering of raw data obtained from social media. This is the same direction [14] is taking, stating that simple sentiment analyses are not the element sufficient for prediction. Moreover, Metaxas and associates have concluded that the number of Facebook friends or Twitter followers does not necessarily lead to success in the elections [10], but the number of posts shared and the constant content posting is an important tool used to increase the level of involvement [15]

Using the data collected from Facebook pages of the candidates in local elections in Croatia held in 2017, this research paper elaborates on the predictive model of the election outcome. The predictive model is based on approximately ten activity index components of the candidates on Facebook.

Chapter 2 gives overview of previous research. Chapter 3 describes the data used in the empirical part of the research and the methodology used in analyzing the said data. Chapter 3 provides the results of the research from the perspective of the interpretation of the obtained model and the assessment of such model. Finally, results are summarized in a conclusion

II. LITERATURE REVIEW

The spreading of social media has provided new methods of opinion polling measurement. Predicting the election outcome by means of large social media databases is a new form of political prediction and is used mostly for the purpose of producing election outcome predictions by collecting the relevant data from social media. Some authors suggest that analyzing data collected from social media during the election campaign might present a useful addition to the traditional polling methodology [12]. The main advantages of public opinion measuring through social media are accessibility and speed [16]. Moreover, in comparison to traditional public opinion surveys, social media can provide a continuous monitoring of the public opinion in real time [12]. Accordingly, access to data through social media could provide a solution to the limitations of traditional surveys. Politicians are increasingly using social media in their campaigns, thus producing such data. Barack Obama, the former President of the United States, is known as 'The first Internet President', because he is said to have used the Internet and information technology in his election campaign like no one before, thus changing the way of communication between the political candidates and their voters [17]. Since then, the research of the importance of social media for the elections has been receiving attention. The key topic associated with the election forecast by means of social media is the use of social media by politicians. The previous surveys conducted suggest that politicians have adopted the use of social media and that there are different factors affecting the patterns of their usage [18]. Recent studies were conducted with the purpose of determining the impact of social media on the political campaigns from several different perspectives, including the political participation, political knowledge and political efficiency [19]. What all of the aforementioned features have in common is persuasion, which is a crucial factor of each political campaign. Every political speech, phone call, every knocking at someone's door, every post on social media aims at promoting the political campaign in order to influence the voters. This leaves open the question of how much and to what extent does this affect the voters and whether such a propaganda is efficient or not.

III. RESEARCH DESCRIPTION

In this chapter, research objectives are defined and data collected for the purpose of achieving the set objectives are described.

There are several objectives of this research:

(i) systematize the results of previous surveys addressing the election result predictor and political communication on Facebook,

(ii) identify the impact of the candidates' Facebook activity on the proportion of votes received,

(iii) determine which of the activity index components has the greatest election outcome interpretation power?

A systematic literature review will be conducted for the purpose of achieving the first objective, thus providing a synthesis of the present knowledge about election outcome predictors and political communication on Facebook. The second objective aims to define and test the candidate's Facebook activity index. The activity index will be composed of several components, in line with the guidelines provided by [20] with some of them being the following:

- The number of photos posted (v1),
- The number of videos posted (v2),
- The number of links posted (v3),
- The number of statuses posted (v4),
- The number of created events (v5),
- Total number of reactions to posts (v6),
- Total number of posts shared (v7),
- The number of the page followers (v8).

In order to check the reliability of the created index, Cronbach alfa will be used, while the Principal Component Analysis (PCA) will be used for the purpose of the index substantive and constructive validity verification, in line with the guidelines proposed by [2].

In order to achieve the third research objective, the paper will propose the three predictive models of the election outcome. The dependent variable in all the three models is the outcome of the elections, constituting the actual results achieved by the candidate at the elections. The election outcome data will be downloaded from the following website: *http://www.izbori.hr/*. The election outcome will be a numeric variable indicating the percentage of the votes the candidate obtained.

Independent variables in the first model are the variables that were used in previous surveys for modeling the election outcome, including the following: sex, education degree, political affiliation, population living at the place where the elections are held, etc. In the second model, in addition to the said variables, the input variables measuring the frequency and type of communication of the candidates on Facebook are added, which are the following: the number of published images, videos, links, statuses, created events, the number of reactions and shared posts, the number of the candidate's page followers, according to [20]. Since most of the said variables correlate, and in order to reduce the number of input variables, in the third model the candidates' Facebook activity index is thus created. The creation of the three models allows us to determine whether or not the predictive power of the model has increased by adding the activity index to the model, and if it has increased indeed, to identify the index component contributing most to the increase. Such models can provide an answer as to whether or not the social media metrics can assist in future attempts to assess the election outcome. For this purpose, the decision tree method will be used. In previous surveys, linear regression was mostly used for the purpose of developing the predictive models of the election outcome. However, since the pilot study has shown that the obtained data does not meet the requirements of this parametric method, the non-parametric method will be used. The decision tree will be

used to determine the impact level of the specific input parameters (such as the frequency and the way of using social media) to the output, that is, the election outcome. [21] suggests that the accuracy and reliability of the model is considered when measuring the quality level of the predictive model of the election outcome. Therefore, the said parameters will be taken into consideration when evaluating the model quality. The difference in accuracy and reliability of the model will explain the predictive power of additional variables and will enable the identification of the impact of the Facebook activity on the election outcome. The diagram describing the three models is given below in Figure 1.

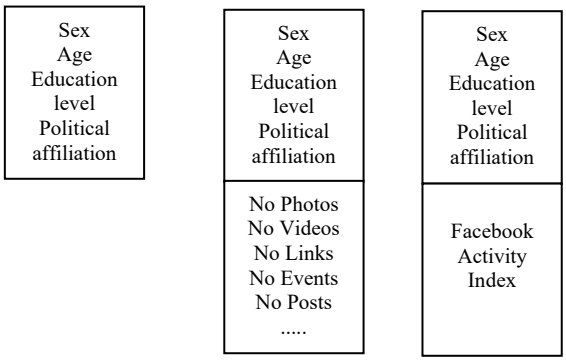

FIGURE 1. RESEARCH MODEL

Using the model sensitivity analysis, the most significant factors of the political candidates' communication on Facebook will be identified and the guidelines for the efficient digital campaign will be provided.

A. Data description

The previous research on the campaigns was mostly based on two different methods, such as the experiment and the questionnaire analysis [22]. Both methods have some disadvantages. Researchers have their doubts as to the validity of the experiment as they consider the impact of the messages conveyed at the laboratory may differ from the ones in real life and the answers obtained during the experiment may differ from the ones showing the actual behavior. The conducting of such surveys shows some other disadvantages [23]. The data obtained from the survey are prone to errors and rely heavily on the survey response rate. The researchers estimate that only 50% of the variance in the data collected through the survey is actual, while the rest is the result of sampling errors [23]. In this survey, the social media data are used for the purpose of designing the predictive models. In order to collect data from the political candidates' Facebook page, Sociograph which is a free tool will be used, which is available at the following website: https://sociograph.io/report.html. Data collection will be conducted on Facebook from the official websites of the candidates participating in the local elections in Croatia held in May 2017. The local elections were chosen based on the proposal of [24] who stated the need for linking the social media activity to individual candidates, not the political parties, in the guidelines provided for future research. Moreover, Facebook was chosen according to the suggestions of [25] claiming that political campaigns are increasingly using Facebook as a social network with the largest number of users. In most of the surveys conducted so far, the research was performed on Twitter, not on Facebook, because it is easier to download data from Twitter, and the tools suitable for collecting data from Facebook are lacking. However, since Facebook is more popular worldwide, particularly in Europe and Croatia, where the research is conducted, this research paper analyzes the data obtained from Facebook. In Croatia, Facebook is the most commonly used social network with 2.1 million users, while Twitter has 55,000 users in Croatia [25].

There are three groups of variables used in the research:
(i) Candidate's characteristics (Candidate's sex, Candidate's age, Education level, Political affiliation, Previous term of office),
(ii) Candidate's Facebook activity (Number of images, Number of videos, Number of links, Number of statuses, Number of events, Number of authors, Number of reactions to posts, Number of posts shared Number of comments to posts Number of reactions to posts, Number of users reacting, Top post category, Number of reactions to the top post, Number of top posts shared, Number of comments to the top post, Number of page likes,
(iii) Election outcome.

B. Methodology description

The basic method used to perform the data analysis in this survey was the decision tree. The decision tree is a modern and popular technique used for solving classification and prediction problems. The simplicity of use of the decision tree lies in the fact that the data model can be read in the form of rules. Such rules can be directly interpreted or can be used in some of the programming languages in working with databases, so specific examples from the database can be singled out by using the rules generated by the decision tree.

There is a wide range of different algorithms used for designing a decision tree, the common and most widely used one being the C4.5 algorithm, i.e., its improved commercial version C5.0. At each tree node, the C4.5 selects an attribute that can most effectively split a set of data into subsets which are added to one or another class. The splitting criterion is the normalized information gain value obtained through the selection of a specific data splitting attribute. The attribute with the largest information gain is selected as a decision-making one.

IV. DISCUSSION

This chapter presents the research results from the perspective of the obtained decision tree model evaluation and the interpretation of the model using the model sensitivity analysis.

The model presented in this research paper was developed based on the sample of 112 candidates for Mayor at the local elections in Croatia in 2017. The survey includes the candidates in the county seats with the personal Facebook page. The model quality is presented in Table 1, showing the model reliability measured using RSquare and the model error measured using the RMSE.

TABLE I. MODELS EVALUATION

Evaluation	Model 1	Model 2	Model 3
RMSE	0.45	0.28	0.26
RSquare Adj.	0.52	0.68	0.71

The predictive models (models 2 and 3) including the candidates' Facebook activity has a minor error (RMSE) and a major reliability rate (RSquare). For the purpose of identifying if there is a statistically significant connection between the performance of specific models, a t-test was performed. The test results showed that the difference between the model 1 and model 2 is statistically significant at the $p<0.005$ level, while the difference between the model 1 and model 3 is statistically significant at the $p<0.001$ level. It has confirmed the presented hypothesis. The following section provides the interpretation of the model by conducting a sensitivity analysis in order to determine which social media activity components contribute the most to explaining the model.

A decision tree model sensitivity analysis was performed for the purpose of identifying the most significant variables determining the election outcome. The sensitivity analysis was performed for each of the models, and the obtained results are shown in Table 2.

TABLE II. SENSITIVITY ANALYSIS

Model 2 significant variables	Model 3 significant variables
Number of events	Activity index
Number of comments	Political affiliation
Political affiliation	Education level
...	...

The sensitivity analysis results show that the candidates' activity index is the most powerful election outcome predictor. The second model has singled out the number of created events and the number of comments to posts as the most significant index factors.

V. CONCLUSION

The connection between using Facebook by political candidates and the election outcome has been elaborated in the present survey, as well as the role of the interaction of the candidates on Facebook in explaining the election outcome. Research results provide the following scientific contribution:

(i) Systematization and synthesis of the previous knowledge of the election outcome predictors and the political communication on Facebook,

(ii) Identification of the most significant factors of political communication on Facebook for the election outcome,

In addition to the scientific contribution, this topic has provided some additional social contributions, which are the following:

(i) Contribution to the local professional literature in the field which has not been sufficiently studied,

(ii) Promotion of the significance of social media to the general public, particularly the interested groups,

(iii) The first research of this type in the Republic of Croatia.

Such research contributes to other studies seeking to develop the predictive models based on the data obtained from social media, as well as the studies examining the ways in which the citizens and politicians use social media. In the digital age, when social media form a significant part of one's life, it is important to examine their impact on the decisions made in different contexts. In addition to this, such research provides an insight into the understanding of the online political behavior of the citizens and politicians at the age of social media, through the prism of a large amount of data. In this research, data have been collected from several hundred candidates at the national level, unlike some previous surveys with limited data on several candidates or several political parties.

The present research has several limitations which must be taken into consideration when interpreting the results and generalizing the conclusions. The first limitation lies in the fact that Facebook users do not include a representative voter population. One of the limitations of using social media data is the fact that social media in some way enable data manipulation. Namely, fake accounts are easy to create and they can be used to reinforce the specific messages, thus reducing the quality of data and presenting a distorted state of affairs. Therefore, the research results must be interpreted with caution and placing the afore-mentioned research limitations within a context.

REFERENCES

[1] Pew Research, available at: https://www.pewresearch.org/fact-tank/2019/09/09/us-generations-technology-use/

[2] Praude, V., & Skulme, R. (2015). Social Media Campaign Metrics in Latvia. Procedia-Social and Behavioral Sciences, 213, 628-634.

[3] DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. PloS one, 8(11), e79449.

[4] Gulati, G., & Williams, C. (2013). Social media and campaign 2012: Developments and trends for Facebook adoption. Social Science Computer Review, 31(5), 577–588.

[5] Fonseca, A. (2011). Modeling political opinion dynamics through social media and multi-agent simulation. In First doctoral workshop for complexity sciences.

[6] Tweetminster. (2011). Can word-of-mouth predict the general election Result? Atweetminster experiment in predictive modeling. Retrieved November 24, 2015, from http://www.scribd.com/doc/29154537/Tweetminster-Predicts.

[7] Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. Journal of Computer-Mediated Communication, 18(2), 40e60.

[8] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010a). Predicting elections withTwitter: What 140 characters reveal about political sentiment. In Paper presented at the 4th International AAAI Conference on Weblogs and Social Media, Washington, DC (pp. 23e26).

[9] Jungherr, A. (2013, October). Tweets and votes, a special relationship: The 2009 federal election in germany. In Proceedings of the 2nd workshop on Politics, elections and data (pp. 5e14). ACM.

[10] Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October). How (not) to predict elections. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 165-171). IEEE.

[11] O'Connor, B., Bamman, D., & Smith, N. A. (2011). Computational text analysis for social science: Model assumptions and complexity (pp. 1e4).

[12] Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 Dutch senate election results with twitter. Paper presented at the Proceedings of the Workshop on Semantic Analysis in Social Media.

[13] Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. Internet Research, 23(5), 544e559.

[14] Gayo-Avello, D. (2012). No, you cannot predict elections with twitter. Internet Computing, IEEE, 16(6), 91e94.

[15] Stefko, R., Dorcak, P., & Pollak, F. (2011). Virtual social networks and their utilization for promotion. Polish Journal of Management Studies, 4, 126-134.

[16] Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Research synthesis social media analyses for social measurement. Public Opinion Quarterly, nfv048.

[17] Greengard, S. The first Internet president. Communications of the ACM, 2009, vol. 52, no 2, p. 16-18.

[18] Strandberg, K. (2013). A social media revolution or just a case of history repeating itself? The use of social media in the 2011 Finnish parliamentary elections. New Media&Society, 1461444812470612.

[19] Rahmawati, I. (2014). Social media, politics, and young adults: the impact of social media use on young adults' political efficacy, political knowledge, and political participation towards 2014 Indonesia general election (Master's thesis, University of Twente).

[20] Straub, D., Boudreau, M. C., & Gefen, D. (2004). Validation guidelines for IS positivist research. Communications of the Association for Information systems, 13(1), 24.

[21] Lewis-Beck, M. S. (2005). Election forecasting: Principles and practice. The British Journal of Politics and International Relations, 7(2), 145-164.

[22] Ceron, A., & D'Adda, G. (2013). Enlightening the voters: The effectiveness of alternative electoral strategies in the 2013 Italian election monitored through (sentiment) analysis of Twitter posts. European Consortium for Political Research, 1-25.

[23] Wlezien, C., & Erikson, R. S. (2002). The timeline of presidential election campaigns. The Journal of Politics, 64(4), 969-993.

[24] Safiullah, M., Pathak, P., Singh, S., & Anshul, A. (2017). Social media as an upcoming tool for political marketing effectiveness. Asia Pacific Management Review, 22(1), 10-15.

[25] Chen, C. Y., & Chang, S. L. (2017). User-orientated perspective of social media used by campaigns. Telematics and Informatics, 34(3), 811-820.

[26] Statistics Portal for Market Data, available at: https://www.statista.com/