Machine Learning Techniques for Analysis of Political Campaign

Ge Xu Suchada Hapikul

***Abstract*—In the latest U.S. election of 2020, the voting results confirmed Biden as the 46th president of the United States. After the U.S. presidential election results have been announced, we have seen in some news reports, social media, and other online channels some voting data and public reactions. U.S politics has lately been in the center of the world's attention with the defeat of controversial 45th US president, Donald Trump and his claims of election fraud in the latest US presidential election. During his government, Trump made extensive usage of social media platforms to share and promote his thoughts, actions and opinions claiming that major media channels failed to portray the truth about his government. As a result, popularizing the term "fake news" as a reference to those channels. We find big data analytic inextricably linked to U.S. elections.**

***Index Terms*—Machine Learning, Fake news, Misinformation, Data**

I. Introduction

This has the potential of finding the factors that make the current solutions effective or ineffective in curving the number of gun shooting incidents. Lastly, identifying the most at risk group of people can be used to ensure that the correct community health or social welfare programs are available so that that potential victims do not end up on the New York City's report of shooting incidents [1] [2], [3] [4], [5], and [3].

In recent years, social media has taken up a large part of people's daily lives, keeping up with real-time news, expressing their opinions, etc [6] [5], [7]–[13]. Due to its unprecedented popularity, these have prompted politicians to use this channel to spread their ideas and political views and reach out more directly to potential voters. It is not uncommon for election candidates to post their daily activities and political statements on social media and even debate social media before and during the campaign (R, Hillegersberg J, Huibers, 2011) (PT E, 2012) (Graham T, M, K, 2012) (Enli GS E, 2015).

These actions attract large numbers of Internet users to discuss them online, and it is an easier way to gather broad public opinion about candidates than traditional polls. Social media mining with ML models reveals useful data [14]–[31] [?]. Several studies have shown the predictability of election results based on social media messages in different countries and regions, including the United States (C G) (J, K, J, F, 2013) (MC, 2015) , the United Kingdom (P, R, L, R, M, 2015), Germany (A, TO, PG, IM, 2010), the Netherlands (ETK J, 2012), and South Korea (M, MC, YK, 2014). The behavior of Internet users and articles has analyzed social media to infer election results.

The best way to predict the future is to study the past. This is one of the basic ideas behind big data analytics. As an example, the 2008 Obama campaign was one of the first to utilize a data- driven approach in its campaign for the elected presidency. The Obama campaign had a data analytics team of 100 people. It shows how deeply data analytics has impacted the world, from recommending products to customers on e-commerce sites (i.e., using predictive analytics) to electing the most powerful officials in the free world. Big data analytics is truly everywhere (Big Data Analytics and Predicting Election Results, 2019).

From the beginning of the U.S. election, statistics of various aspects of each candidate have been counted on multiple websites. Big data analytics can analyze each candidate's "tags," the slogans and strategies of different candidates, and the public's attention to these keywords and analyze the public trends. Social media can use big data statistics and judgment to make people's decisions deviate and guide the whole public opinion and even change many people's original intention.

The two companies that specialize in providing data analysis and services for the Democratic and Republican parties are TargetSmart and DeepRoot Analytics. The former specializes in providing big data analysis and services to Democrats and state Democrats and their allies, while the latter provides data analysis to the Republican Party and its affiliated teams.

Both TargetSmart and DeepRoot use Alteryx's software to illustrate their ability to accommodate, cleanse, blend, and analyze large-scale information from diverse sources. This approach focuses on exploring the age structure of voters, segmenting and scoring them according to age groups, and then using this information to optimize their media spending, especially on the all-important TV commercials, to amplify the impact of the campaign and make things work twice as well.

II. literature Review

The main areas that will be covered by the literature research are presidential election voter participation by state and voter participation in Georgia in the last 3 US presidential elections. In addition, we will perform an analysis of a data set containing posts and/or comments about the most recent US presidential election from one social media platform. The group believes that looking at this information will help us identify patterns and connections in the data to formulate our solutions. We propose to look at data from MIT Science Lab, FiveThirtyEight, The University of California Irvine, and Kaggle.com.

In addition, we will also look at articles and journals containing big data analysis of US politics, we will later list these resources in our final paper.

Obviously, data analytic is a powerful tool which is widely used to increase benefits by predication consumer's needs. It is also used politically in elections to forecast impossibility and seeking opportunity to provide information in order to convince voters. Singh A. (2019) reported that since the Obama campaign, data analytics has grown into the brain of every election campaign.

Data analytics aids the election campaign in better understanding voters; candidates can adapt those need into their preferences. Moreover, Nickerson W. D. Rogers T. (2013) indicated that campaign of data analyst is playing a bigger role in politics; developing predictive models can provide individual-level scores that forecast citizens' likelihoods of engaging in political behaviors and endorsing candidates.

There are some studies mentioned about the factor which impact the results of presidential election. Jackson, J. (1999)(1999) studied about the economic impact upon modern U.S. elections, the author reported that people vote in hopes of improving their personal finances while also helping the country's overall economic condition. Kurtbaş, İ. (2015) claimed that the philosophy of the candidate was cited by a number of voters as the most important factor affecting their vote in local elections. Lawless, J.L. and Fox, R.L. (2001) and Cohen, C.J. and Dawson, M.C. (2001)found a correlation between political participation and income levels (wealth, social well-being, money, and economic development); people with a high degree of income and wealth are more likely to engage in politics and government. Online information can be misleading in some cases [32], [33].

III. Method

We propose to parse through news articles, numerical voter support, and social media text data about the 3 last US presidential elections. We would like to produce a network or other types of visualizations formed by voter geographical and social information, which can be linked by relations of support and opposition to candidates. This network is studied by applying insights from several theories and techniques, and by combining existing, including graph partitioning, centrality, assorting, and hierarchy. The analysis will hopefully yield interesting and clarifying patterns. We would like to perform an Exploratory Data Analysis approach in which we maximize insight into the data sets, detect outliers and anomalies, and extract important variables.

Twitter is a popular microblogging website. Tweets are used to express a person's perspective on a subject, and every day 8TB of data is generated. We feel it is the most significant platform with data and can be analyzed, so Our primary evaluation method will be based on the Twitter data. The project requirements can be carried out by the Sentiment Analysis System that collects and processes machine-generated Big Data from Twitter to identify systematically, extract, and subjective information from the Twitter tweet's generated by a user and using text analysis methods and further process. The greatest task is to gather all the details and to summarize the tweets.

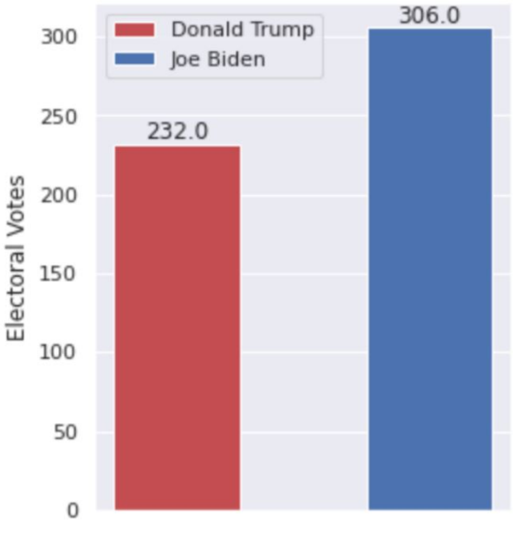

Fig. 1.

In this project, we have to collect all tweets related to 'user support for elections in different states in the United States' and performed sentimental analysis using python. For that, we have to process unstructured raw data by collecting from the social networking site "TWITTER" by using Twitter API tokens, which are generated after creating Twitter developer's account and now should process it to semi-structure data. The tweets are collected into JSON files and then imported into MongoDB.

A python script is to be written to retrieve the data from MongoDB or JSON. The evaluation of the data is done to get users' opinions about 'Presidential elections 2020 in each state' like percentage of user's tweets talking positively, negatively, and neutrally about 'Biden and Trump', which can be done by seeing polarity in each tweet or by using a lexicon. For all these analyses, massive data is required, so data is to be collected every day from Twitter using API keys and tokens. Data can be visualized in tableau, geolocation and leaflet.

IV. Results

From Figure 2, we can see that in this election, Biden won 26 of the 51 states or 51 percent. Biden received 76.19 million votes at the voter level, for the 71.5 million Trump received, or 51.6 percent. Trump only had an advantage in the number of counties won, with his total wins in 3,135 states, or 70 percent.

Figure 3 shows the number of electoral votes in each state, the candidates they support, and their approval rates. The dark

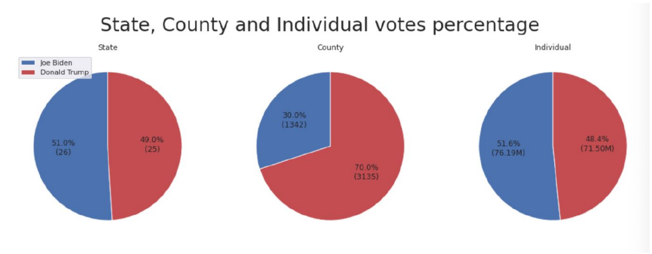

Figure 2

Fig. 2.

red and dark blue columns indicate the states that supported Trump and Biden and their votes; the light red and light blue columns indicate the ”swing states” that supported Trump and Biden and their votes; the dotted line indicates the support rate of each state for the corresponding candidate. Through the above data analysis and visual display, the following conclusions can be found:

1. The ratio of votes received by Biden and Trump is 306:232, with a large difference between them. The distribution of votes is consistent with our traditional view that the Democratic Party mainly relies on supporting the economically developed regions on the east and west coasts. In contrast, the Republican Party primarily depends on the support of the central agricultural provinces. However, it should be noted that there are more ”swing states” in this election, and Texas, the traditional Republican vote, is not a powerhouse for Trump.

2. At the state, county, and individual voter levels, Biden narrowly won the state and voter levels, while Trump won a relatively large number of counties. In essence, Biden not only gained a relatively large number of electoral votes by winning 26 states, but he also gained more than 50% support among the public, which can be considered an overall victory. The main reason Trump was able to win at the county level is that his votes mainly came from the Midwestern agricultural states, vast and sparsely populated. Although many of them are by county units, their populations are small, and their votes are small, and winning these counties did not help him win the election.

3. From the distribution of electoral votes in each state and the degree of support for the two parties in each state, most of the votes for the Republican Party are small. Its primary sources of votes (Iowa, Ohio, Texas, Florida, North Carolina, and other states) are ”swing states”, although they choose to support the Republican Party, but not overwhelming support. On the contrary, the Democratic Party has not only gained support from some states in the ”swing states”, but its own vote sources are also more stable than the Republican Party, such as New York, Illinois.

References

[1] M. Heidari, J. H. J. Jones, and O. Uzuner, “Offensive behaviour detection on social media platforms by using natural language processing models,” 2021.

[2] S. H. Bae, H. Shin, H.-Y. Koo, S. W. Lee, J. M. Yang, and D. K. Yon, “Asymptomatic transmission of SARS-CoV-2 on evacuation flight,” *Emerging Infectious Diseases*, vol. 26, pp. 2705–2708, Nov. 2020.

[3] E. M. Choi, D. K. Chu, P. K. Cheng, D. N. Tsang, M. Peiris, D. G. Bausch, L. L. Poon, and D. Watson-Jones, “In-flight transmission of SARS-CoV-2,” *Emerging Infectious Diseases*, vol. 26, pp. 2713–2716, Nov. 2020.

[4] N. C. Khanh, P. Q. Thai, H.-L. Quach, N.-A. H. Thi, P. C. Dinh, T. N. Duong, L. T. Q. Mai, N. D. Nghia, T. A. Tu, L. N. Quang, T. D. Quang, T.-T. Nguyen, F. Vogt, and D. D. Anh, “Transmission of SARS-CoV 2 during long-haul flight,” *Emerging Infectious Diseases*, vol. 26, pp. 2617–2624, Nov. 2020.

[5] M. Bielecki, D. Patel, J. Hinkelbein, M. Komorowski, J. Kester, S. Ebrahim, A. J. Rodriguez-Morales, Z. A. Memish, and P. Schlagenhauf, “Air travel and COVID-19 prevention in the pandemic and peri-pandemic period: A narrative review,” *Travel Medicine and Infectious Disease*, vol. 39, p. 101915, Jan. 2021.

[6] T. W. Russell, J. T. Wu, S. Clifford, W. J. Edmunds, A. J. Kucharski, and M. Jit, “Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study,” *The Lancet Public Health*, vol. 6, pp. e12–e20, Jan. 2021.

[7] S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, “Emotion detection of textual data: An interdisciplinary survey,” in *IEEE 2021 World AI IoT Congress,AIIoT2021*, 2021.

[8] A. Adekunle, M. Meehan, D. Rojas-Alvarez, J. Trauer, and E. McBryde, “Delaying the COVID-19 epidemic in australia: evaluating the effectiveness of international travel bans,” *Australian and New Zealand Journal of Public Health*, vol. 44, pp. 257–259, July 2020.

[9] M. Heidari, S. Zad, B. Berlin, and S. Rafatirad, “Ontology creation model based on attention mechanism for a specific business domain,” in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[10] M. Heidari and J. H. Jones, “Using bert to extract topic-independent sentiment features for social media bot detection,” in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pp. 0542–0547, 2020.

[11] M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. P. y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini, and A. Vespignani, “The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak,” *Science*, vol. 368, pp. 395–400, Mar. 2020.

[12] S. Zad, M. Heidari, J. H. J. Jones, and O. Uzuner, “A survey on concept-level sentiment analysis techniques of textual data,” in *IEEE 2021 World AI IoT Congress,AIIoT2021*, 2021.

[13] M. Heidari and S. Rafatirad, “Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews,” in *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*, pp. 1–6, 2020.

[14] S. Chen, S. Owusu, and L. Zhou, “Social network based recommendation systems: A short survey,” in *2013 International Conference on Social Computing*, pp. 882–885, 2013.

[15] S. Lin, C. Liu, and Z.-K. Zhang, “Multi-tasking link prediction on coupled networks via the factor graph model,” in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, pp. 5570–5574, 2017.

[16] M. Heidari, J. H. J. Jones, and O. Uzuner, “Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter,” in *IEEE 2020 International Conference on Data Mining Workshops (ICDMW), ICDMW 2020*, 2020.

[17] Y. Chu, F. Huang, H. Wang, G. Li, and X. Song, “Short-term recommendation with recurrent neural networks,” in *2017 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 927–932, 2017.

[18] C. Yang, X. Chen, T. Song, B. Jiang, and Q. Liu, “A hybrid recommendation algorithm based on heuristic similarity and trust measure,” in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pp. 1413–1418, 2018.

[19] S. Ji and J. Liu, “Interpersonal ties and the social link recommendation problem,” in *2019 6th International Conference on Systems and Informatics (ICSAI)*, pp. 456–462, 2019.

[20] M. Heidari and S. Rafatirad, “Bidirectional transformer based on online text-based information to implement convolutional neural network model for secure business investment,” in *IEEE 2020 International Symposium on Technology and Society (ISTAS20), ISTAS20 2020*, 2020.

[21] J. Wang, H. Song, and X. Zhou, “A collaborative filtering recommendation algorithm based on biclustering,” in *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 803–807, 2015.

[22] S. Chen, S. Owusu, and L. Zhou, “Social network based recommendation systems: A short survey,” in *2013 International Conference on Social Computing*, pp. 882–885, 2013.

[23] M. Heidari and S. Rafatirad, “Semantic convolutional neural network model for safe business investment by using bert,” in *IEEE 2020 Seventh International Conference on Social Networks Analysis, Management and Security, SNAMS 2020*, 2020.

[24] A. Gatzioura, J. Vinagre, A. M. Jorge, and M. Sànchez-Marrè, “A hybrid recommender system for improving automatic playlist continuation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 5, pp. 1819–1830, 2021.

[25] Z. Liao, Y. Song, Y. Huang, L.-w. He, and Q. He, “Task trail: An effective segmentation of user search behavior,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 3090–3102, 2014.

[26] M. Heidari, J. H. J. Jones, and O. Uzuner, “An empirical study of machine learning algorithms for social media bot detection,” in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[27] C.-Y. Chi, Y.-S. Wu, W.-r. Chu, D. C. Wu, J. Y.-j. Hsu, and R. T.-H. Tsai, “The power of words: Enhancing music mood estimation with textual input of lyrics,” in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–6, 2009.

[28] A. Gatzioura, J. Vinagre, A. M. Jorge, and M. Sànchez-Marrè, “A hybrid recommender system for improving automatic playlist continuation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 5, pp. 1819–1830, 2021.

[29] H. Yang, C. He, H. Zhu, and W. Song, “Prediction of slant path rain attenuation based on artificial neural network,” in *2000 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 1, pp. 152–155 vol.1, 2000.

[30] M. Heidari, S. Zad, and S. Rafatirad, “Ensemble of supervised and unsupervised learning models to predict a profitable business decision,” in *IEEE 2021 International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021*, 2021.

[31] M. Thompson, G. Alshabana, T. Tran, and A. Chitimalla, “Predict covid-19 cases using opensky data.” http://mason.gmu.edu/ ttran81/, 2021.

[32] M. Heidari, J. H. J. Jones, and O. Uzuner, “Fraud detection to increase customer trust in online shopping experience,” 2021.

[33] R. Gao, J. Li, B. Du, X. Li, J. Chang, C. Song, and D. Liu, “Exploiting geo-social correlations to improve pairwise ranking for point-of-interest recommendation,” *China Communications*, vol. 15, no. 7, pp. 180–201, 2018.