Issues in Information Systems
Volume 22, Issue 4, pp. 321-331, 2021

DOI: https://doi.org/10.48009/1_iis_2021_321-331

Influence of data-driven methods in predicting U.S. presidential elections for a specific age ranging using social media

Mel Tomeo, *Miami Dade College, mtomeo@mdc.edu*
David Hertz, *Pittsburgh Technical College, hertz.david@ptcollege.edu*
John J. Scarpino, *Pittsburgh Technical College, scarpino.john@ptcollege.edu*
Richard Bryant, *Pittsburgh Technical College, bryant.richard@ptcollege.edu*
Michael Hodder, *Pittsburgh Technical College, hodder.michael@ptcollege.edu*

Abstract

The purpose of this case study is to provide a better understanding on how data analytics affects presidential elections and how election campaigns use data-driven methods to connect with a specific age range of voters. Big data analytics can be used to predict future voter behavior patterns from previous events. Data-driven methods can be used ranging from recommending products to customers on e-commerce websites to electing individuals to a political office. Big data analytics can now help election campaigns communicate effectively and target voters. A review of the literature regarding the general background of big data analytics and related technologies is provided. Hypotheses were tested using a research model. Data was collected through a survey that was conducted on 91 participants who used LinkedIn and Facebook as social media tools. The findings indicated that individuals of a similar age changed their opinion on who they voted for in the 2020 United States presidential election based on television ads and news media outlets.

Keywords: Data Analytics, Data-driven Methods, Presidential Elections, Elections, Big Data, Behavior Patterns, Election Campaigns, Election Influences, Votes, Media, Social Media

Introduction

Big data has become a common term that can have several definitions with many different perspectives. Ward and Barker (2013) described big data as the storage and analysis of large and complex data sets using a series of techniques such as NoSQL, MapReduce, and machine learning. Watson (2014) explained that big data is a term that can be used to describe data that has high volume, high velocity, and high variety that requires technology to capture, store, and analyze the data. Watson (2014) further explained that big data can be used to enhance decision making and provide insight to optimize results. Labrinidis and Jagadish (2012) defined big data as data volume, acquisition velocity, or data representation that limit the ability to perform effective analysis using traditional relational approaches for efficient processing. Big data analytics can be described as a complex process of examining big data to find information, such as hidden patterns, correlations, market trends, and individual preferences (Kambatla et al., 2014). This process can help businesses make educated decisions to produce the best results. Individuals can use big data analytics to impact the world, ranging from recommending products to customers online to influencing political elections.

Data analytics has become a critical aspect of every election campaign. Election campaigns have used data analytics to target specific voters on social media websites (Ma-Kellams et al., 2018). Social websites such as Facebook and Twitter can adjust their newsfeeds to target a specific audience to promote voting. In the 2016 Presidential elections Google trends was used to predict each state's votes (Ma-Kellams et al., 2018). Chauhan et al. (2021) found that the 2020 United States Presidential election indicated a significant impact in how elections are managed from the usage of big data analytics and data-driven methods through social media. Prescriptive analytics can be described as an analytical technique to help businesses make better decisions through the understanding of raw data (Rajaraman, 2016). This technique can be used to factor specific information into the data, such as possible situations or scenarios, past performance, and current performance (Bertsimas & Kallus, 2014).

The following will present a review of the literature to give a better understanding of the importance of this study.
Previous studies focused on how using data analytics on social media can influence election results will be discussed. The methodology of this study will be described, followed by the results. Finally, the limitations, conclusions, and future recommendations on how to extend this study will be offered.

Literature Review

The proliferation of social media over the past several years has given rise to a new era of data collection and data analysis. Through analysis of social media posts, trends can be more easily identified, predicted, and at times, even manipulated (Jain & Kumar, 2017). This has become increasingly apparent when it comes to elections. Ceron et al. (2016) investigated the reliability and accuracy of election polls. Ceron et al.'s (2016) results indicated that traditional methods of polling and surveying voters is quickly becoming outdated due to issues of low response rates. Ceron et al. (2016) concluded that analyzing online data has several advantages over traditional media and polling methods. One advantage they found was the ability to project and detect trends and major events, as well as the ability to analyze large amounts of data and interactions. The following literature will present previous studies that focused specifically on how data analytics have been used in elections in the United States and foreign countries.

Elections in the United States

González (2017) investigated how presidential candidate Donald Trump used data analytics in 2016 by hiring the company Cambridge Analytica several months before the presidential election. González found that Cambridge Analytica utilized Facebook user profiles to narrowly target voters. González discovered that during the same election, the data analytics system used by the Hillary Clinton campaign failed to predict the loss of Wisconsin and Michigan. The system that Clinton used was biased and did not count the impact of rural voters at the same level that it counted the impact of urban and suburban voters. González's results indicated that Clinton's system created incorrect simulations that predicted a victory for the Clinton campaign that did not materialize. Xie et al. (2018) extended González's (2017) study by using data analytics to examine the data that was collected from the 2016 presidential election campaigns on three popular social media sites: Facebook, Twitter, and Google. Xie et al.'s (2018) results indicated that using this technique correctly predicted the winner for the presidential election.

Cambridge Analytica created FB "personality quizzes" encouraging users to take online quizzes and in the process grant Cambridge Analytica access to their FB profiles (Kozlowska, 2018). Cambridge Analytica's marketing materials claimed to have 5,000 data points on over 22 million Americans. Cadwalladr and Graham-Harrison (2018) investigated Cambridge Analytica's claims of swaying the 2016 election. They

found that the claims were exaggerated. However, their results indicated that the ability of data scientists working with psychologists and other social scientists can influence human behavior.

Guo et al. (2016) examined the 2012 United States presidential election using data analytics to find the most frequent tweets. Guo et al. used a data analytical tool to analyze 77 million tweets. They found that the two most frequent tweets were Barack Obama's related foreign policy and Mitt Romney's related taxation plan. Jain and Kumar (2017) conducted a similar study regarding how data analytics was successfully used on social media and other data sources in recent elections. Jain and Kumar (2017) found that big data analytics was heavily used during the 2008 and 2012 Obama presidential election campaigns to target a specific audience. Lindoo (2020) extended Jain and Kumar's (2017) study by investigating how data analytics could be used to micro target groups of voters to sway their decisions on voting choices. Lindoo (2020) found that the presidential election in 2012 and in 2016 used data analytics to target a specific age of voters.

Elections in Foreign Countries

VV (2019) researched how the use of big data analysis became increasingly common in Indian elections from 2014 through 2019. VV found that during the 2014 Indian elections, politicians used big data to understand voters' socio-economic status and political leanings. Their results indicated that politicians used this data to target their messages to specific voters. By the 2019 elections, the use of these practices had spread to regional political parties as well as national ones. Several 2019 elections demonstrated the incorporation of big data in predicting outcomes. Awais and Ahmed (2019) attempted to predict the outcome of each district of the Pakistani 2018 elections. They created a formula to gather the results of the previous four elections, public polling records from the previous two years, and tweets captured three weeks before the election. Their model proved to be 83% effective in predicting the outcomes of each district election.

Zhou and Makse (2019) investigated the 2019 Argentinian presidential election. They investigated how Artificial Intelligence (AI) and big data analysis could be used to predict and influence elections, as well as impact important referendums. They focused on why pollsters failed to correctly predict results. Their results indicated that big data could be used to detect significant shifts in opinions and that traditional polling could be very inaccurate. Grimaldi et al. (2020) extended Zhou and Makse's study by using data analytics to successfully predict not only the outcome of the 2019 Spanish presidential race, but to correctly rank the individuals in the order of votes. They concluded that polling will eventually be replaced by AI tools for predicting the outcomes of elections.

Research

Research Question and Hypotheses

The main research question that this study addressed was: Did data analytics influence and assist in the United States 2020 presidential election? The specific research question that this study addressed was:

Did data-driven analytic methods influence a specific age range by using social media to affect the results of the 2020 presidential election in the United States?

The following hypotheses were tested:

> H_1. Individuals of a similar age allowed ads from social media to impact their decisions on who they voted for in the 2020 United States presidential election.
>
> H_2. Individuals of a similar age learned about the people who were running for president of the United States from social media.
>
> H_3. Individuals of a similar age determined their voting choice in the 2020 United States presidential election based on social media ads.
>
> H_4. Individuals of a similar age determined their voting choice in the 2020 United States presidential election based on television ads and news media outlets.

Questionnaire Development and Testing

The questionnaire contained 7 questions. A demographic question was asked to gather the age range of the participants. A pretest was conducted on a group of participants who completed the questionnaire by themselves, without intervention or support from the researcher. The pretest was given to participants from specific targeted audiences. The purpose of the pretest was to validate the questions on the questionnaire. This test was conducted through a survey questionnaire created on www.surveymonkey.com. The researchers used SurveyMonkey due to its reputation of stability and for the simple appearance of the interface that it provides. SurveyMonkey uses traditional web widgets such as checkboxes and radio buttons. This interface helped reduce the amount of instructions on how to reply to the questions. SurveyMonkey was chosen by the researchers due to the built-in functions to analyze the results of the data collection. These tools have been tested and validated by previous studies. The tools that were provided by SurveyMonkey were at no cost to the participants or researchers.

Data Collection Methodology

All constructs were measured with previously validated instruments. In this study, one survey was conducted on a variety of participants. The instruments used in the study were a series of survey questions that were measured on a 5-point Likert-type scale in which 1 denoted "Strongly Agree (SA)," 2 denoted "Agree (A)," 3 denoted "Neither Agree Nor Disagree (NAND)," 4 denoted "Disagree (D)," and 5 denoted "Strongly Disagree (SD)." The participants for the survey were given the link through a social media posting on LinkedIn and Facebook. The participants were able to access the questionnaire between March 22, 2021, and April 5, 2021. The targeted participants were individuals who used social media. Participants were given an introduction and the purpose of the survey before being asked to take it. Participants were expected to fully understand the purpose of the survey and agree to the terms and conditions before proceeding to complete the survey. The purpose of the survey was to collect data and analyze the results to add to the body of literature regarding how data analytics can influence elections in the United States. Surveys and questionnaires are widely used in research to target a specific population with questions designed to measure and collect data pertaining to a specific topic (Alvarado et al., 2016). This technique provides precise calculations of the variables that are being used in the study.

The following statements in the questionnaire were given to the participants of this research study:

RQ_1: Please select your age range.
RQ_2: Ads that I have seen on social media websites impacted my decision on who I voted for in the 2020 United States presidential election.
RQ_3: I learned about the people who were running for president of the United States from social media websites.
RQ_4: I believe that social media ads can change how an individual feels towards someone running for a political position.
RQ_5: I believe that offline marketing like billboards and television ads were deployed to target a particular audience to gain their vote.
RQ_6: I feel that social media can play a role in who is elected as president of the United States.
RQ_7: I believe that general news media outlets play a role in who is elected as president of the United States.

Questionnaire Deployment

The approach to invite participants to the survey was done online through two social media postings on LinkedIn and Facebook. This was the only contact with the participants, and it explained the purpose of the research, who the researchers were, and the average time that would be spent to complete the questionnaire. SurveyMonkey.com provided a header for the survey questionnaire to include additional information for the participants. This helped show the participants that the research was focused on a specific topic.

Data Analysis

Civelek (2018) published a book on structural equation modeling (SEM) and explained that using this technique in a research study could reveal the relationships among the variables that are not directly measured. Civelek demonstrated how the SEM technique can be used to reveal direct and indirect relationships between variables. The SEM technique was used to analyze the relationships between age in this study. This technique was chosen to measure how the variation of the latent variables (how ads from social media play a role in elections, learning election knowledge from social media, ads targeting a specific audience, and general news media role in elections) impacted the measured variable (age). The goal was to understand and indicate a relationship between these latent variables and the measured variable. Replication of this study using this framework and research design is possible. Figure 1 shows the theoretical framework for this study. This theoretical framework displays the latent variables and the measured variable on how data-driven prescriptive analytic methods influence elections through social media.

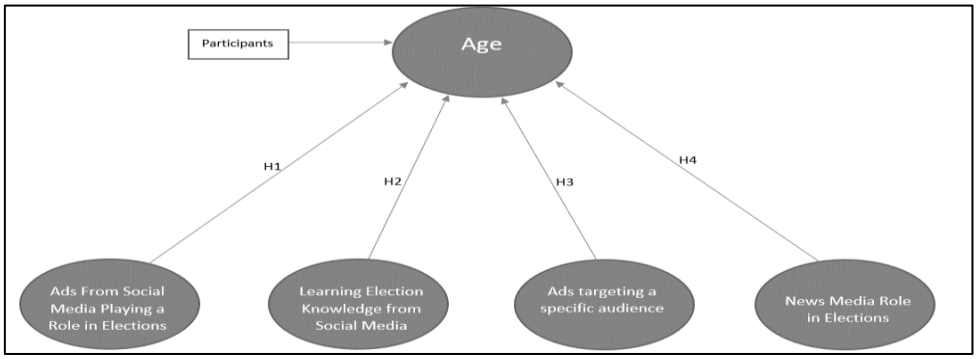

Figure 1. Theoretical Framework.

RESULTS

The questionnaire was taken by 91 participants. The data was collected from March 22, 2021, until April 5, 2021. The participants were over the age of 18 years. The possible age range choices were 18-24, 25-34, 35-44, 45-54, and over 55. Of the participants, 12% (11 participants) were in the 18-24 range, 20% (18 participants) were in the 25-34 range, 31% (28 participants) were in the 35-44 range, 16% (15 participants) were in the 45-54 range, and 21% (19 participants) were over the age of 55. The results from the survey showed 67.03% of the individuals fell in the age range of 25-54. Question 1 asked the participants to select their age range. The demographic factor (age) of the participants who completed this survey can be found in Figure 2.

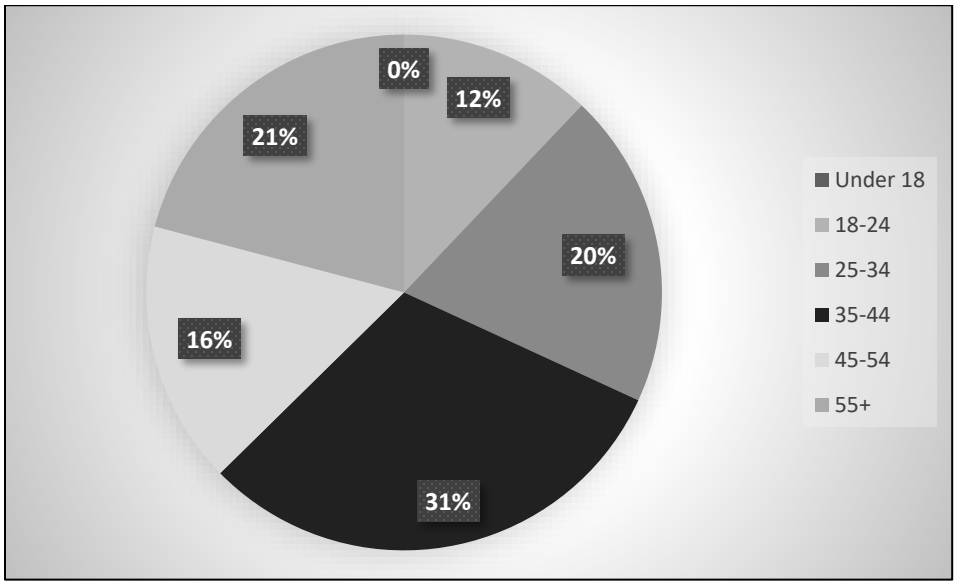

Figure 2. Demographics of the participants

Findings

In research question 2, 75% of the participants did not feel that social media influenced their decision of who they voted for in the 2020 presidential election. In research question 3, 56% of the participants did not learn about who was running for president from social media websites. In research question 4, 73% of the participants believed that social media ads could change how an individual feels toward a candidate for president. In research question 5, 74% of the participants believed that offline marketing like billboards and television ads was deployed to target a particular audience to gain their vote. In research question 6, 73% of the participants felt that social media could play a role in who is elected as president of the United States. In research question 7, 78% of the participants believed that general news media outlets played a role in who was elected as president of the United States. It is important to observe that over 70% of the participants selected "Strongly Agree" or "Agree" for research questions 4, 5, and 6. Another important observation is that 75% of the participants selected "Disagree" or "Strongly Disagree" for research question 2. Table 1 shows the breakdown of the survey results.

Table 1. Survey Results.

RQ	SA	A	NAND	D	SD
RQ_2	7	11	5	24	44
RQ_3	12	9	19	18	33
RQ_4	22	44	9	11	5
RQ_5	16	51	7	8	9
RQ_6	21	45	11	9	5
RQ_7	18	53	10	6	4

Data Synthesis for Research Question

The SEM technique was conducted to see if any relationships existed among the variables. Figure 3 displays the model and the path coefficients that were created in the PLS-SEM tool. The model represents the measured variable and the latent variables. The path coefficients were calculated through an algorithm in a sequence of regressions in terms of weight vectors. The weighting scheme consisted of 300 maximum iterations with a stop criterion (10^-X) of 7.

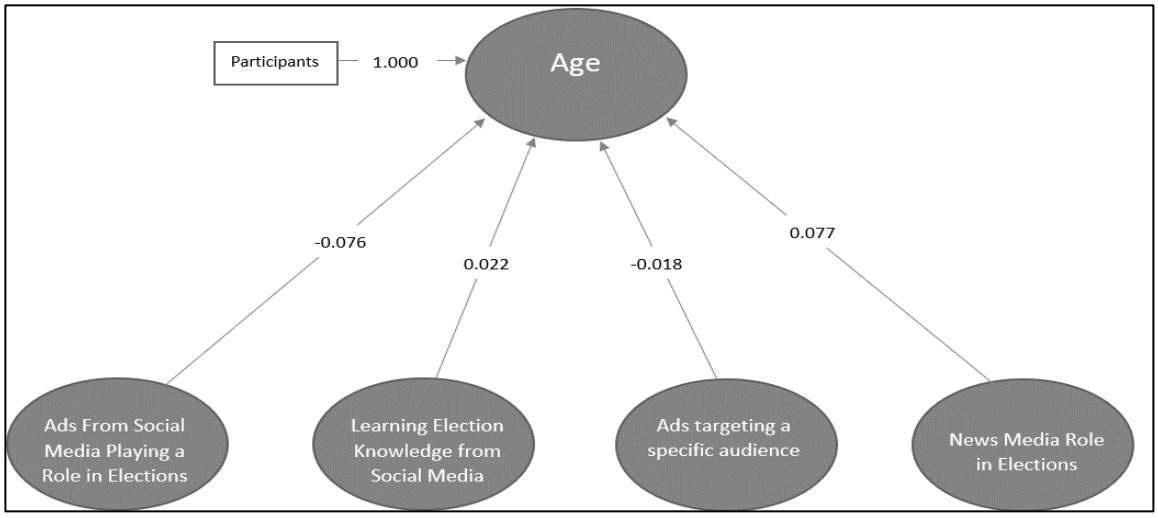

Figure 3. Coefficients of the SmartPLS Model.

Summary

An algorithm known as bootstrapping was used to test the statistical significance of the path coefficients of the SmartPLS model. A significance level of 1% was applied with a subsample of 500 and parallel processing to the tacit and explicit model. The breakdown of the *p*-values for H1 through H5 were: H1 resulted in generating a *p*-value of 0.490, H2 resulted in generating a *p*-value of 0.849, H3 resulted in generating a *p*-value of 0.874, and H4 resulted in generating a *p*-value of 0.010. Figure 4 displays the *p*-value results for the model. Based on the *p*-values derived from bootstrapping, only H4 was supported, while H1, H2, and H3 were not supported and must be rejected. Only H4 was supported and the rest of the hypotheses were not supported due to the significance level of 1% (*p*-value > 0.1).

Path Coefficients

Mean, STDEV, T-Values, P-Values | Confidence Intervals | Confidence Intervals Bias Corrected | Samples

	Original Sa...	Sample Me...	Standard D...	T Statistics (...	P Values
Ads From Social Media Playing a Role in Elections -> Age	-0.076	-0.082	0.110	0.690	**0.490**
Ads targeting a specific audience -> Age	-0.018	-0.015	0.114	0.159	**0.874**
General News Media Role in Elections -> Age	0.077	0.076	0.124	0.621	**0.010**
Learning Election Knowledge from Social Media -> Age	0.022	0.016	0.117	0.191	**0.849**

Figure 4. Bootstrap results of the SmartPLS model.

Reliability and Validity

Reliability was established by using three different subject matter experts (SMEs) to generate the survey questions. A SME is an individual who is a specialist in their field, with degrees and years of experience in a particular topic (Mattoon, 2005). The SMEs made the determination of which questions should be on the questionnaire based on their knowledge and experience. The candidates to be SMEs in this research were recruited through a list of college faculty members. The candidates were determined based on their experience working within their chosen field, collaboration techniques, and soft skills.

Discriminant validity was applied by using a statistical technique to determine the relationship between the variables. The Fornell-Larcker criterion was used to measure the discriminant validity. This approach compared the average variance extracted (shared variance within) of the constructs to the squared correlation between the constructs (shared variance between). Figure 5 displays the results of the Fornell-Larcker criterion. As all the coefficients in the diagonal are larger than the values in the table, discriminant validity is guaranteed.

Fornell-Larcker Criterion | Cross Loadings | Heterotrait-Monotrait Ratio (HTM

	Ads From S...	Ads targeti...	Age	General Ne...	Learning El...
Ads From S...	1.000				
Ads targetin...	0.110	1.000			
Age	-0.049	-0.009	1.000		
General Ne...	0.383	0.162	0.050	1.000	
Learning Ele...	-0.003	0.206	0.037	0.229	1.000

Figure 5. Results of the Fornell-Larcker Criterion.

Discussion

Previous studies regarding Cambridge Analytica showed the ability of data analytics to predict the outcome of a presidential election (González, 2017). However, those studies do not indicate the ability of social media to sway opinion on the matter of which candidate one chooses in an election. The rejections of H1 and H3 are novel findings in that the results of the questionnaire showed that the sample both believed that social media affected their choice of candidate, and that social media affected their opinion of their choice in a candidate. This could be due to several factors including the predominant age group, education level, or political ideals of the participants.

Additionally, the questions regarding the ability of social media to change one's mind regarding candidate choice may not have been read by the individual as pertaining to them but instead simply asking the individual's thoughts on society in general. The rejection of H1 and H3 effectively shows that their opinion that social media sways public opinion is not supported by the results of the study. Regarding H2, it is not surprising that the participants learned which candidates were running in the 2020 Presidential election due to the widespread media coverage and length of the election season.

The acceptance of H4 indicates that participants had a strong belief that traditional news media outlets and television advertisements did not have a strong effect upon public opinion with regards to their ability to effectively change one's decision regarding choice of candidate in an election. This may indicate that the participants held traditional communications methods as being more trustworthy as opposed to social media, which does not have the same level of professional standards as traditional communications channels.

Limitations

The implementation of this study was not without certain limitations. The study is limited by the fact that it only focused on the measurement of specific variables. A limitation existed regarding the political views of the participants. It is unknown whether certain answers to the questions were biased, based upon the underlying political views of the participants, therefore altering the acceptance of one or more hypotheses. Another limitation is the lack of awareness of the educational level of the participants and the size of the data sample. Further investigation is needed to establish if the same results could be duplicated through a larger data sample and applied across a broader context.

A future study could include the use of additional questions to determine the political affiliation of the participants and attempt to ensure an equal representation of both sides of the political spectrum. Additionally, an expansion of the sample size would be warranted to eliminate as much bias in the responses as possible. The need to balance the study by age group may also be warranted to remove any significant influence one large sub-sample may have upon the survey. Lastly, a second survey could be conducted asking the same questions to determine if opinions had changed since the participants' initial responses.

Conclusion

In conclusion, this study provided insightful information regarding if data analytics influenced and assisted in the United States 2020 presidential election. This study adds to the body of knowledge related to data-driven analytic methods influencing a specific age range by using social media to affect the results of the 2020 presidential election in the United States. Based on the findings, 75% of the participants believed that ads they had seen on social media websites did not impact their decision on who they voted for in the 2020 United States presidential election, 56% of the participants did not learn about who was running for president from social media websites, 73% of the participants believed that social media ads could change how an individual feels toward a candidate for president, 74% of the participants believed that offline marketing like billboards and television ads was deployed to target a particular audience to gain their vote, 73% of the participants felt that social media could play a role in who is elected as president of the United States, and 78% of the participants believed that general news media outlets played a role in who was elected as president of the United States.

Previous literature review has shown that data-driven analytical methods using social media and the ability to analyze large amounts of data quickly have proven to be reliable for predicting elections. Data-driven

analytical methods have also demonstrated to be effective in predicting important referendums. Data-driven analytical methods can detect trends and events that will greatly impact the future of elections globally. Big data is a powerful tool for people seeking political office. How that tool will be used in the future to not only predict, but to influence, elections will be of great significance. The data sources identified in most of the literature reviewed came from social media sites. These sites can generate an enormous amount of data very quickly, which can be used with analytic techniques for finding specific information. As massive amounts of data accumulate, researchers will need a method by which they can analyze the data. By using data driven methods to analyze data, an individual or organization can recommend a course of action or strategy to predict the outcome of a particular situation.

References

Awais, M., Hassan, S. U., & Ahmed, A. (2019). Leveraging big data for politics: Predicting general election of Pakistan using a novel rigged model. *Journal of Ambient Intelligence and Humanized Computing*, 1-9.

Bertsimas, D., & Kallus, N. (2014). From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481*.

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, *17*, 22.

Ceron, A., Curini, L., & Iacus, S. M. (2016). *Politics and big data: Nowcasting and forecasting elections with social media*. Taylor & Francis.

Chauhan, P., Sharma, N., & Sikka, G. (2021). The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, *12*(2), 2601-2627.

González, R. J. (2017). Hacking the citizenry? Personality profiling, 'big data' and the election of Donald Trump. *Anthropology Today*, *33*(3), 9-12.

Grimaldi, D., Cely, J. D., & Arboleda, H. (2020). Inferring the votes in a new political landscape: The case of the 2019 Spanish presidential elections. *Journal of Big Data*, *7*(1), 1-19.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332-359.

Holbrook, T. M., & McClurg, S. D. (2005). The mobilization of core supporters: Campaigns, turnout, and electoral composition in United States presidential elections. *American Journal of Political Science*, *49*(4), 689-703.

Jain, V. K., & Kumar, S. (2017). Towards prediction of election outcomes using social media. *International Journal of Intelligent Systems and Applications*, *9*(12), 20.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561-2573.

Kozlowska, I. (2018). Facebook and data privacy in the age of Cambridge Analytica. *Seattle, WA: The University of Washington. Retrieved August 1*, 2019.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, *5*(12), 2032-2033.

Lindoo, E. (2020). The making of a president using data analytics and social media. *Journal of Marketing Development and Competitiveness*, *14*(1), 75-82.

Ma-Kellams, C., Bishop, B., Zhang, M. F., & Villagrana, B. (2018). Using "Big Data" versus alternative measures of aggregate data to predict the US 2016 presidential election. *Psychological Reports*, *121*(4), 726-735.

Mattoon, J. S. (2005). Designing and developing technical curriculum: Finding the right subject matter expert. *Journal of STEM Teacher Education*, *42*(2), 5.

Mavragani, A., & Tsagarakis, K. P. (2019). Predicting referendum results in the Big Data Era. *Journal of Big Data*, *6*(1), 1-20.

Rajaraman, V. (2016). Big data analytics. *Resonance*, *21*(8), 695-716.

VV, A. V. (2019, December). Big data analytics and transformation of election campaign in India. In *Proceedings of the 2nd International Conference on Information Systems & Management Science (ISMS)*.

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, *34*(1), 65.

Ward, J. S., & Barker, A. (2013). Undefined by data: A survey of big data definitions. *arXiv preprint arXiv:1309.5821*.

Xie, Z., Liu, G., Wu, J., & Tan, Y. (2018). Big data would not lie: Prediction of the 2016 Taiwan election via online heterogeneous information. *EPJ Data Science*, *7*, 1-16.

Zhou, Z., & Makse, H. A. (2019). Artificial intelligence for elections: the case of 2019 Argentina primary and presidential election. *arXiv preprint arXiv:1910.11227*.