Seminar Paper

Advanced Topics in Social Network Analysis

Learning from 2016: predicting elections through Twitter

Submitted by:

Daniel Stroh

Sebastian Reiserer

Table of contents

1. Abstract	Page 3
2. Introduction	Page 4
3. Literature Review	Page 6
4. Research Question & Hypotheses	Page 10
5. Methodology	Page 12
6. Results	Page 14
7. Discussion	Page 17
8. Conclusion	Page 20
9. References	Page 21

1. Abstract

What happened during the news coverage for the 2016 election cycle(s) is a case study that will be studied for years to come. The failure to accurately predict Brexit or Donald Trump's success in the US presidential election left everybody perplexed as neither, especially the latter, was ever predicted to occur. This research investigates the failures of the media to predict Trump's election and investigates whether Twitter can be used in the place of existing procedures to correctly predict elections. Data obtained through this research shows that online discussion of Trump by Twitter users was higher than discussion of Hillary Clinton during four critical events leading up to election day. This paper argues that this data can be used to make successful predictions for election outcomes. The findings and conclusion help underscore the importance of this type of research, not only in terms of predicting elections via Twitter, but also in terms of how such insights might benefit the media and its audiences.

2. Introduction

In the year 2016 two major election upsets occurred that went contrary to the popular belief held by experts, traditional mainstream media and the general populace alike. Despite the upset that was 'Brexit', no shock was bigger than Donald Trump's victory in the 2016 US presidential election. After continuous media coverage suggesting that he had no chance of winning, we ask ourselves how the media could have gotten it so wrong.

Learning how and why the media, particularly television and online media, was unsuccessful in accurately predicting the elections is essential to preventing similar blunders in the future. Being able to predict elections creates stability in economic, political and policy aspects while also ensuring that people have realistic views and expectations formed by correct media coverage. The collective failure of the media to correctly predict recent elections also means that the techniques they use to make predictions must be insufficient. We therefore want to see if there is another way to accurately predict elections.

With the emergence and rapid rise of social media in our everyday lives we want to know if it is possible to research these users to help determine voting patterns to help predict elections. We are not the only ones seeing the potential of social media to predict elections; other researchers are already conducting research into this phenomenon. Through the use of prior papers such as Tumasjan's 2009 paper on the German federal elections, or the research of Colleoni et al. into echo chambers, we will form our bedrock of understanding for this research paper.

In section three we will investigate and learn from previous research looking into the prediction capacity of both social media and news media. In the following section we will present our arguments, leading us to the research question. The method(s) that we will employ to answer the research question will be clearly defined in section five, titled 'Methodology'. The next part (section six) will be used to present our research data results, which we will thoroughly analyze afterwards in section seven. At this point we will also examine how our findings relate to and help answer our research question. In the final part, section eight, we will provide a brief overview of the steps taken to answer our research question.

3. Literature Review

Research on social media

Rapid developments in society and technology are causing an ever-increasing presence of online activity. During the 2008 US presidential election campaign, Barack Obama's strategies focused on incorporating and utilizing social networking platforms such as MySpace, Twitter, and Facebook. The 2008 campaign represents a watershed moment for social networks, as it was their first large-scale application in US presidential primaries. As Tumasjan et al. point out, "Twitter has become a legitimate communication channel in the political arena as a result of the 2008 campaign" (Tumasjan, Sprenger, Sander, & Welpe, 2010). Every subsequent election has heavily focused on developing strategies for social media, making it an indispensable part of any modern election. Of course, social media can be used for a variety of purposes besides being used as a political platform (e.g., Hayat & HersHKovitz, 2018; Hayat, HersHKovitz & Samuel-Azran, 2019). As important as social media is in getting elected, can it also become a way to predict elections? To find out we will examine research addressing this inquiry.

The paper '*Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment*' (Tumasjan et al., 2010) researched the 2009 parliamentary elections in Germany. This paper is useful to help us understand the potential of Twitter for predicting elections. 104,003 tweets were gathered in this research to analyze the correlation between Twitter and voting. The research by Tumasjan et al. showed that nearly 70% of these tweets referred to six major parties, with the remaining tweets (approximately 30%) referring to politicians. The data was divided to find which party and their respective politician was tweeted about, with what sentiment, and how often. Tumasjan et al. concluded that the mere amount of tweets "mentioning a political party can be considered a plausible reflection of the vote share and its predictive power even comes close to traditional election polls." The paper concludes by stating that, according to its findings, mention of politicians and their parties on Twitter reflects their popularity as measured via polling by the

traditional media. Furthermore the research concluded that Twitter activity prior to the election seems to “validly reflect the election outcome.” This research is in support of our view about Twitter’s capacities to predict elections, but do more research papers come to the same conclusion? To test the validity of this research, more papers must be studied to cultivate an informed opinion.

Dhiraj Murthy investigates the role tweets play in elections in the paper *‘Twitter and elections: are tweets predictive, reactive, or a form of buzz?’* Based on the 2012 US Republican presidential primaries, the paper explores whether Twitter sentiment and a candidate’s Twitter use correlate with electoral success. Sentiment towards every candidate was separated into positive and negative categories, with the aggregated results indicating an almost neutral overall sentiment score for everyone. Candidates Ron Paul and John Huntsman employed buzz marketing to gain online popularity; they had the highest tweet frequency which resulted in them having the highest sentiment score. Despite heavily using social media, their social media presence was not significantly correlated with winning debates or primaries (Murthy, 2015). Romney, on the other hand, had among the lowest cumulative sentiment score during primaries and ultimately won the nomination. The differences between candidates’ sentiment scores and their actual performance during voting indicates that online ‘sentimental’ political capital does not in fact relate to how voters end up voting.

At this point we must also ask ourselves what discussion on social media sites such as Twitter is like. How do users communicate and interact with one another, and is the online Republican presence underestimated, as it was by traditional polls? Colleoni, Rozza, and Arvidsson discuss echo chambers and homophily (meaning the tendency of similar individuals to form ties with each other) on Twitter. Unlike previous studies, they used the entire population of Twitter users in their study. This gave them a much larger sample size, thereby preventing sampling bias of the population. Their goal was to see if Twitter encourages discussions between people with different views on politics or if Twitter users only interact with like-minded people,

indicating homophily.

Their findings indicated that, in general, Democrats have an overwhelming amount of ties with other Democrats and Republicans have a high amount of outbound ties with Democrats. Running more tests, they tested both Republicans and Democrats separately and saw that Democrats had an even higher homophily rate, and Republicans a lower homophily rate, than expected. These results were similar to those of previous studies like Pennacchiotti and Popescu (2011), which found that “Democrats tend to consistently have a large percentage of friends with the same affiliation. For Republicans, the political affiliation of the neighbors is more mixed”. But these results were also the opposite of what some other previous studies found, such as Boutyline and Willer, who found “that conservatives are more homophilous than liberals.” Colleoni, Rozza, and Arvidsson’s second experiment assumed that users’ political stance was reflected by which official account (Democrat or Republican) they were following. When they did this, their own prior results were reversed, with Republicans yielding a higher rate of homophily than Democrats. Though the studies cited above do not allow us to point to a conclusive difference in Democrats’ and Republicans’ rates of homophily, the findings do suggest that Twitter acts both as an open public sphere for participation and as an echo chamber with reciprocal ties (e.g., Hayat, Samuel-Azran & Galily, 2016).

Research on traditional media

To understand how the traditional media was inaccurate in its predictions regarding Brexit and Donald Trump’s election, we must first understand how it makes its predictions. In the paper ‘*The Evolution of Election Polling in the United States*’, Hillygus investigates how the media acquires its election predictions.

The traditional media uses professional survey companies to make initial predictions for elections. However, as these require long-term research and analysis Hillygus concludes that this form of prediction cannot account for

any of the unexpected events that occur during an election. Predictions that are made using past voter patterns can also suffer from the same problem, as these models “rarely engage the most up-to-date scholarly research on political participants” (Hillygus, 2011).

Hillygus says that in order to make updated predictions throughout the election cycle, the media as well as pollsters use data gained from surveys, interviews, third party sources, and on- and offline opinion polls. Hillygus finds many problems associated with this, as this data, and the methodologies used to obtain it, are often – unsurprisingly – unreliable.

The paper found that despite the high demand for accurate predictions for elections, there is a "decline in the prominence of polls in election politics." This deterioration might be related to a current trend in traditional reporting which prioritizes sensationalism over the detached presentation of facts, as well as a general lowering of standards caused by financial pressures.

4. Research Question & Hypotheses

In the aftermath of the 2016 US presidential election, many were left wondering how this could have happened after constantly being told by the media that Hillary Clinton was going to win. Hillygus's research found how misguided the media was in dismissing Trump's chances, estimating at one point that "the presidential candidate has a 1% chance to win" (Diaz, 2015). This trend continued through the last few days leading up to the election, with some predictions giving Trump a mere 15% chance to win two days before election day (Katz, 2016). The media's inability to accurately predict election outcomes on numerous occasions throughout the year is of great concern to us. We therefore want to conduct our own research to find out why traditional media failed in its predictions and to discover if social media can be used as an alternative to accurately predict elections.

Several of the studies mentioned above support the notion that social media such as Twitter can be used as tools to predict elections, which is something we want to investigate ourselves. Intriguingly, research conducted by Murthy led him to the insight that social media sentiment for candidates does not in fact help researchers predict voters' actual intentions at the polls. This helps shape our research procedure, as sentiment will not be used in this investigation. Meanwhile, the research by Tumasjan et al. found that popularity on social media could be indicative of both candidates' and political parties' electoral success. The findings of these prior studies help shape our own inquiry into traditional and social media.

As the issues of inaccuracy in the media and the potential of social media to predict elections are interconnected, we decided to focus our research on both. Thus, our research question asks: **'Is the media's traditional way of predicting political election results obsolete, and can we use online social media (Twitter) in its place, to more accurately predict election results?'**

We have seen that the media was consistently inaccurate in predicting the 2016 US election, and we therefore hypothesize that at least some of their methods of predicting elections are outdated. Based on our preliminary understanding of social media research, gained through the literature review, we also hypothesize that social media does in fact possess the capacity to predict elections.

5. Methodology

Empirical analytical methods will be used in our paper in order to answer our research question. We will focus on testing the research question by analyzing Twitter to see if this particular social media network can support our theory, or provide alternative insights. After acquiring the relevant data, we will analyze, discuss, and explain our findings and relate them back to our research question.

To help us analyze Twitter we will use Buzzilla, an online tool that monitors, analyzes, and tracks social media conversations. This tool will provide us with the ability to obtain quantitative information based on tweets from the 2016 US election cycle. Taking a quantitative approach means that we will obtain numerical data on which candidate is more frequently discussed on Twitter, information which will then be compared to actual poll results.

The population used for the research encompasses all Twitter users of voting age (18 years old and over); this means that we will assume that the people tweeting about either candidate will participate in the voting process.

In order to address our research question we will conduct non-probability sampling of Twitter across four major election dates. These dates are:

Date	Event
September 26 th , 2016	First presidential debate
October 9 th , 2016	Second presidential debate
October 19 th , 2016	Third presidential debate
November 8 th , 2016	United States presidential election day

While we will be using non-probability sampling to gain data for these four dates, it does mean that we will unfortunately not be able to randomize the population and users of Twitter within our research. The downside of this sampling method is that researchers are unable to accurately represent the entire population of Twitter, meaning that results may not entirely reflect the actual voting population. However, as our research focuses on obtaining

unobtrusive insight into a sample of Twitter users, representative of the population as a whole, we believe bias won't impact our findings. Additionally, non-probability sampling is used to conduct our research as it offers benefits such as ease of availability of data (via Buzzilla), cost effectiveness, efficiency, and establishing so-called pilot data.

We will collect tweets referring to both the Republican and Democratic presidential candidates, i.e., Trump and Clinton, within our specified dates. On Buzzilla we will look at Twitter discussion of these two candidates by searching for keywords relating to each. This will then be used to see if the amount of discussion on Twitter translates into real life popularity.

Since our research paper hypothesizes that the media's traditional way of predicting political elections is obsolete, we will also be conducting our own investigation of said media's predictions about the 2016 US election. This will be accomplished by reviewing election forecasts and polling numbers relating to the four above-mentioned event dates. To get an accurate representation of predictions made by the media, we will use a variety of sources from different media networks and combine them into an average polling forecast to gain a general overview.

6. Results

In our research we collected 555,993 tweets that discussed Trump (52%) or Hillary (48%) within our specified time period, mentioned in our methodology section. Depicted in figure 1, one can see the ongoing relationship between the number of tweets discussing Republican candidate Donald Trump (in red) and Democratic candidate Hillary Clinton (in blue).

Figure 1

By including data which ranges from our first date to our last date, figure 1 provides additional insight into the relationship being studied. A very noticeable takeaway from this data is that following any one of the four specified dates the amount of tweets increases significantly compared to any other day. The day after the first date, discussion of Trump increased by approximately 50.8%, with discussion of Clinton increasing slightly more, by 54.2%. This trend continued throughout the three remaining dates and their following day, with discussion of Trump increasing by 72.3%, 25.6%, and 234.9% after each respective date, and discussion of Hillary increasing by 84.4%, 8.3%, and 145.1% respectively.

To improve our understanding of the amount of traffic that occurred for each candidate on our targeted dates, we highlighted the amount of discussions, and their corresponding makeup of overall Twitter, in table 1.

Date	Candidate	% of Twitter	Discussion #	Discussion %
26.09.16	Donald Trump	1.94%	6,801	51.6%
26.09.16	Hillary Clinton	1.82%	6,371	48.4%
09.10.16	Donald Trump	2.20%	6,646	56.8%
09.10.16	Hillary Clinton	1.67%	5,054	43.2%
19.10.16	Donald Trump	2.41%	7,848	52.0%
19.10.16	Hillary Clinton	2.23%	7,247	48.0%
08.11.16	Donald Trump	4.71%	14,730	50.1%
08.11.16	Hillary Clinton	4.68%	14,642	49.9

Table 1

As shown in table 1, Trump consistently has a higher percentage of Twitter's population posting about him. The amount of discussion related to Donald Trump is higher on all four dates compared to his counterpart Hillary Clinton. Although Donald Trump does have higher discussion, on three out of four dates there is no significant difference between the two candidates as discussion traffic is within a 0.05% and 7% range. The one outlier date is the second debate in which Trump received 24% more discussion.

The tweets obtained in this research discussing both candidates are calculated into percentages in the last column of table 1. This simplification of the sampling data shows how much relative discussion the candidates received from the total sample population.

Trump had 51.6% of the population discussing him during the first debate compared to Hillary's 48.4%. During the second debate, Trump was even more popular at 56.8% while Clinton fell to 43.2%. During the third debate discussions of the candidates once more became more even, with discussion of Trump making up 52% of traffic compared to Hillary's 48%. On election day both candidates were nearly identical as Clinton's 49.9% narrowly lost out to Trump's 50.1% advantage.

In actuality, Donald Trump and Hillary Clinton were not the sole candidates participating in each event, meaning that polling numbers aren't split solely between them and their numbers do not by themselves make up the full 100% recorded. The averaged data collected from media forecasts for both candidates anticipated Hillary Clinton winning on every single date, both in the debates and the actual election.

For the first debate Hillary's average polling numbers were 43.0%, with Trump receiving 40.4% (First debate: Clinton & Trump, 2016). For the second debate, averages of media predictions saw Hillary winning by 44.8% compared to Trump's 40.8% (Second debate: Clinton & Trump, 2016). During the third debate Hillary increased her margin to 45.8%, ahead of Trump whose average polling rating decreased to 40.4% (Third debate: Clinton & Trump). With the help of the statistical analysis website FiveThirtyEight we obtained the odds given to each candidate to win on our fourth research date (election day). FiveThirtyEight's forecast gave Clinton a 71.4% chance of winning, with Trump's chance of winning being estimated at a mere 28.6%.

In the end Donald Trump won the 2016 US Presidential election after securing 306 Electoral College votes (270 are needed to win), with Hillary achieving 232. This essentially means that while Clinton was 16.3% away from reaching the minimum 270 Electoral College votes, Donald Trump's 306 votes made him a decisive winner with 15.5% over the required amount (an overall difference of 31.9% between the two candidates). The general election results saw Trump receive a lower percentage of the popular vote (45.9%) than Clinton, who received 48.0% of votes (Presidential election results, 2017).

7. Discussion

To address and answer our research question we will take a look at both facets of the question. First, we will examine whether **‘the media’s traditional way of predicting political election results is obsolete’**. We also will discuss if Twitter is able to predict elections, and subsequently conclude if we can **‘use online social media (Twitter) in the place of traditional media, to more accurately predict election results.’**

Through our research we unintentionally found a pattern in which discussion increased significantly the day after our research dates. This increase in the volume of traffic represents the highest fluctuation in discussion within our data range. A possible reason for this is that the buzz surrounding the event dates caused an inflation of the traffic the following day. Despite this, the data collected on these days does not indicate a deviation from Twitter users’ normal discussion patterns, merely an increase in volume.

The research dates show that Donald Trump consistently received more discussion, which according to our hypothesis would predict Trump’s victory. This was partially true, because although he received 2.1% fewer votes in the popular vote, the decisive electoral vote was 31.9% in his favor. This data shows that Trump’s popularity was much higher than was predicted by the traditional media. The dates that we researched also suggest that Trump’s popularity superseded Hillary’s, with Twitter discussions on election day indicating a 0.2% advantage for Trump (50.1% to Hillary’s 49.9%). This evidence both supports and undermines the claim that social media can be used to more accurately predict election results as it incorrectly predicted the popular vote, but correctly predicted the president. A possible interpretation of this outcome is that because the election was extremely close, more dates should have been sampled to improve the observable popularity trends. Another factor that influences the findings is that the US election system is not based on overall popularity, but rather on the Electoral College system. The prediction capacity of Twitter is supported by the findings of Tumasjan et al. wherein they also discovered that mentions of politicians (and their parties) are reflective of the voting outcome. Our research did successfully predict the

overall popularity of Trump among voters, and that this would lead to his election.

Speaking in terms of traditional media, the results clearly indicate that throughout the election the media failed to accurately predict the real support Trump had at every stage. The media's own opinion polls conducted after each debate showed a trend of Hillary's support growing over Trump's, supporting their coverage that predicted her election victory. This trend is not supported by our data, which shows it was Trump who gained popularity after each debate leading up to the election. This only helps underscore how the traditional media's reluctance to use social media to predict elections results in inaccurate reporting.

Regardless of how the media collected its predictive data regarding the 2016 US presidential election, the undeniable fact is that they failed to give an accurate picture of Trump's popularity. This failure should not however mean that the media's traditional way of predicting political elections has become entirely obsolete. Rather, the outcome of the 2016 election should present the traditional media with a watershed moment, one that should motivate them to incorporate social media as an additional way of predicting elections to more accurately predict the popularity of candidates in future elections.

Twitter has been shown to have some capacity to predict elections, which combined with the constant increase in social media usage overall, makes it both an attractive and essential avenue for further research. While such research does not impact Twitter users themselves, it does influence other researchers and, possibly, the media and other related institutions.

Every researcher (us included) tries to be as accurate and objective as possible throughout their research process. We should however also be able to reflect on our research and acknowledge any limitations, which is why we will offer a few ways the research could be improved. If this research were to be repeated, we would suggest that a larger sample size population be used. Another way to improve on our research would be to increase the sampling scale by collecting data from more than four dates. An in-depth study of the historical accuracy of the media's ability to predict elections should be performed in order to contextualize the media's current predictions. Furthermore, the results of this research also raise the question of whether,

and how, other types of social media may help predict elections. Another question that arises from our research is unrelated to elections, namely: what other practical uses might Twitter's prediction capacities have?

In accordance with our hypothesis we have established and clearly demonstrated the media's failure to predict the 2016 US presidential elections, although we did conclude that this does not entirely invalidate their prediction methods. Our research into social media and its prediction capacity did yield results in support of our hypothesis that it possesses the qualities to accurately predict elections, which should be used as an additional tool employed by the traditional media.

8. Conclusion

In our research we have shown that the current way the media predicts elections is outdated, inaccurate, and needs to adopt new approaches, approaches that social media such as Twitter are capable of fulfilling. The findings show distinctive advantages that researching and analyzing Twitter has with regard to predicting elections. Our research findings provide benefits to a variety of fields, such as social and news media, politics (pollsters and campaigns), fellow researchers, and more. Our research informs them about the extent of Twitter's prediction capacity and in doing so will make them better equipped to predict candidates' success in future elections.

From our original interest in how the media could be so mistaken in predicting the 2016 US presidential election and how this could be resolved, we researched and developed our own understanding that led us to our research question. With a strong understanding of the matter at hand we hypothesized what our data would later confirm (i.e., the media's inaccuracy and social media's successful prediction ability). To test our hypothesis we developed a strategy, using Buzzilla to determine candidates' popularity among Twitter users. This information highlighted the limitations of traditional prediction methods, while also demonstrating that Twitter could have been used to gain a more accurate picture of Trump's popularity, and consequently could have helped to successfully predict his electoral success.

9. References

- (2016, September 16). *CPD Invites Hillary Clinton and Donald J. Trump to Debate*. Retrieved from <http://debates.org>
- (2016, October 4). *CPD Invites Hillary Clinton and Donald J. Trump to Debate*. Retrieved from <http://debates.org>
- (2016, October 14). *CPD Invites Hillary Clinton and Donald J. Trump to Third Presidential Debate*. Retrieved from <http://debates.org>
- (2016, November 8). *Who will win the presidency?*. Retrieved from <https://projects.fivethirtyeight.com>
- (2016, November 8). *2016 General Election: Trump vs. Clinton*. Retrieved from elections.huffingtonpost.com
- (2017, August 9). *Presidential Election Results: Donald J. Trump Wins*. Retrieved from <https://www.nytimes.com>
- Bialik, C., & Enten, H. (2016, November 9). *The Polls Missed Trump. We Asked Pollsters Why*. Retrieved from <https://fivethirtyeight.com>
- Boutyline A., & Willer, R. (2013). *e social structure of political echo chambers: Ideology and political homophily in online communication networks*. Working paper, University of Berkeley.
- Colleoni, E., & Rozza, A., & Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*. 64(2), 317-332. DOI: 10.1111/jcom.12084
- Crespi, I. (1988). *Pre-election polling*. USA: Russell Sage Foundation.
- Diaz, Daniella. (2015, July, 9). *Think Trump has a chance to snag GOP nomination? Analysis gives him just 1%*. Retrieved from <http://edition.cnn.com>
- Hayat, T., Samuel-Azran, T., & Galily, Y. (2016). Al-Jazeera Sport's US Twitter followers: Sport-politics nexus?. *Online Information Review*, 40(6), 785-797.
- Hayat, T., HersHKovitz, A., & Samuel-Azran, T. (2019). The independent reinforcement effect: The role diverse social ties play in the credibility assessment process. *Public Understanding of Science*, 28(2), 201-217.
- Hayat, T., & HersHKovitz, A. (2018). The role social cues play in mediating the effect of eWOM over purchasing intentions. *Journal of Customer Behavior*. 17 (3), 173-187.

Hillygus, S. D. (2011). The Evolution of Election Polling in the United States. *Public Opinion Quarterly*, 75(5), 962–981.

Katz, J. (2016, November 8). *Who Will Be President?*. Retrieved from <https://www.nytimes.com>

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415–444.

Metaxas, P., & Gayo-Avello, D., & Mustafaraj, E. (2011). How (Not) to Predict Elections, IEEE Third International Conference on Social Computing, Boston, 2011, IEEE Publications.

Murthy, D. (2015). Twitter and elections: are tweets, predictive, reactive, or a form of buzz?. *Information, Communication & Society*, 18(7), 816-831.

(n, d). What Consumers Really Need Retrieved from <http://www.buzzilla.com/solutions/overview/>

Pennacchiotti, M., & Popescu, A. (2011). A machine learning approach to Twitter user classification, Proceedings of AAAI Conference on Weblogs and Social Media, 2011.

Pollster. (n.d.) In *Oxford Dictionaries*. Retrieved from <https://en.oxforddictionaries.com/definition/pollsters>

Tumasjan, A., & Sprenger, TM., & Sander, PG., & Welpe, IM. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, Fourth International AAAI Conference on Weblogs and Social Media, Washington, 2010.