Faculty Research and Scholarship

3-2011

On the predictability of the U.S. elections through search volume activity

Catherine Lui

P. Takis Metaxas
pmetaxas@wellesley.edu

Eni Mustafaraj
emustafa@wellesley.edu

Follow this and additional works at: http://repository.wellesley.edu/scholarship

Version: Post-print

ON THE PREDICTABILITY OF THE U.S. ELECTIONS
THROUGH SEARCH VOLUME ACTIVITY

Catherine Lui, Panagiotis T. Metaxas and Eni Mustafaraj
Department of Computer Science, Wellesley College
Wellesley, MA 02481, USA

ABSTRACT

In recent years several researchers have reported that the volume of Google Trends and Twitter chat over time can be used to predict several kinds of social and consumer metrics. From the success of new movies before their release to the marketability of consumer goods to the prediction of voting results in the recent 2009 German elections, Google Trends and Twitter message volume have been treated as indispensable tools for not only recording current social trends, but even for predicting the future. This is surprising, given the significant differences in the demographics of voters and people who use social networks and Web tools. But is there some underline logic behind these predictions or are they simply a matter of luck? With this work we wanted to test their predictive power against the US elections. One could argue that, following the previous research literature, and given the high utilization that the Web and the social networks have in the US, Google Trends and Twitter volume may be able to predict the outcomes of the US Congressional elections. In this paper we report that Google Trends was, actually, not a good predictor of both the 2008 and 2010 elections, and we offer some explanation on why this may be the case. On a forthcoming paper we report on our analysis on Twitter.

1. INTRODUCTION

Recently, there has been increased interest in mining social network activity data in an effort to detect the present (Ginsberg et.al. (2009), Choi & Varian (2009), Metaxas & Mustafaraj (2010)) or predict the future (Asur & Huberman (2010), Tumasjan et. al. (2010), O'Connor et.al. (2010)). These efforts have focused on the assumption that the volume of keywords searched and the chatter in social networks are revealing the current thinking of a large and quickly growing section of the population (Khabrov & Cybenko (2010), Romero et.al., (2010)). In particular, the studies on influence and prediction have been focusing on (a) the volume of Google Trends (Google, 2010), representing the overall information interests of the general community of people using Google as their main search engine (Pew, 2008) and (b) the volume of Twitter chat, representing the interests of small but quickly growing community of people using short text messages to communicate with friends and followers.

The overall findings of the research literature so far are that, over time, both of these sources can be used to predict all kinds of social and consumer data: From the success of new movies before their release (Asur & Huberman, 2010) to the marketability of consumer goods (Choi & Varian, 2009) to the prediction of the recent German elections (Tumasjan et. al., 2010) and polls (O'Connor et.al., 2010), Google Trends and Twitter message volume have been treated as indispensable tools. This may be surprising given, for example, the significant differences in the demographics of voters and people who use social networks and Web tools.

With this work we examine whether Google Trends and Twitter chat volume could be used to predict the voting results of the US elections, as it has reportedly done for the recent German elections. We have analyzed the 2008 and the 2010 US Congressional elections and compare the voting results to two factors that traditionally are used by news media for predicting the US elections: Incumbency, i.e., the fact that

people who are already in office are more likely to be re-elected, and the polling analysis that the New York Times is conducting prior to elections.

In this paper, we report on some very interesting results: We have found that Google Trends was, in general, not a good predictor of both the 2008 and 2010 elections, as compared to incumbency, the NYT polls, and even chance. But this is not the whole story: In a specific, important and well-defined subset of data, that of highly contested 2008 races, we found Google Trends to be a better predictor. We discuss whether this was predictive or accidental.

On a forthcoming paper we report on our findings of Twitter data.

1.1   Google Trends

Google's dominance in the search engine field has had a substantial impact on how people navigate the Internet and design their websites. Because having one's website appear on the first page of search results for a relevant query means having immense exposure due to the high quantity of people who use Google every day (Pew, 2008), web pages are specifically engineered to be placed on that front page. Additionally, due to Google's success and reputation, the first few entries on the first page of results are considered the most reputable sources available online for the query, and users typically conclude their search after choosing one of those entries. Every other search engine now mimics the Google's algorithm that ranks the entries. So one can simply use Google as a reliable representative when studying the reliability and predictive power of search results.

Google Trends (Google, 2010) analyzes a portion of Google web searches to compute how many searches have been done for the terms you enter, relative to the total number of searches done on Google over time. For brevity, we will refer to it also as G-trends. A study by Choi and Varian (2009) investigated the possibility of utilizing G-trends to assay interest in buying cars and homes. It is based on the assumption that the search terms people were querying Google with, were an indication of what they were thinking about. While traditional surveillance methods for analyzing the current spending tendencies of a population lag by a considerable amount of time, Google data is more easily accessible and can provide important information in short time and without cost. For example, G-trends has been used to positively predict influenza epidemics (Mohebbi et. al., 2009). By observing how frequently Google is queried with search terms related to sickness symptoms and the flu, and what regions of the world are searching from, researchers have been able to determine which parts of the world will be stricken with the flu, more efficiently than traditional clinical surveillance methods.

But besides consumer goods and epidemics outbreaks, is G-trends really able to predict elections? A Web search with terms "predicting election results with Google Trends" returns many sources in which journalists answer "yes" to this question. These are clearly non-scientific analyses because they pick and choose which results to analyze and how. Nevertheless, if there is a widespread belief among the journalists that G-trends have such a predictive power, it may not be long before it becomes a self-fulfilling prophecy, influencing voters' decisions: reassuring and exciting some, while discouraging others from voting in pursuit of a lost battle.

2.   METHODOLOGY

Our primary tool for obtaining Google search data was Google Trends provided by Google Analytics (http://www.google.com/trends). Because raw search volume data is sensitive information that Google does not publish openly, G-trends does not provide concrete numbers. However, they do provide the Search Volume Index of a keyword, which is a processed version of the data: it has been scaled based on the average search traffic of a keyword and normalized to accommodate for densely populated regions in the world. The foremost use of G-trends is not to get a sense of the absolute quantity of searches, but to see how a key term's search frequently has changed over time based on the average number of times the key term has been searched for some time period. G-trends is also optimized to compare the search frequencies of multiple key terms, which was suitable for our purpose of comparing the relative search popularity of two candidates vying for a political position.

Google Trends make Search Volume Index (SVI) data available through a downloadable file in .csv (comma-separated values) format that is customizable through the Trends interfaces. We utilized the pyGTrends Google Trends API, developed by Juice Analytics, to expedite the process of downloading all of the .csv files of the Senate and House races. For the 2008 elections we focused on the data from the week before the elections, since it was the one with the largest collection of data and because traditionally that is the week that most of the undecided voters make up their minds on who to vote for.

We also downloaded the names of the candidates and the results of elections using the New York Times (http://www.nytimes.com/politics/), one of the more authoritative newspapers in the US, which maintains detailed data of prior and current elections, including information of incumbent success. We used page scrapes since there is no available API that we could use, but we made sure to obey the established crawling protocol so that we would not flood their server with requests.

For the 2010 elections, we use the same technique, but we retrieve the data twice a day. We also recorder the predictions of the newspaper, based on polls that were coming in. At the time of our research, no predictions of the 2008 elections based on polls were available, however.

3. PREDICTING US ELECTIONS RESULTS WITH GOOGLE TRENDS

First, we list the comparative data of the two elections and then we discuss them.

Table 1: The 2008 and 2010 US Congressional Elections.

	2008	2010
Total Races	413	441
House Races	381	408
Senate Races	32	33
Highly contested	61	125
Democrats	237	200
Republicans	177	241
"landslide win"	Democrats	Republicans

During the 2008 Congressional Elections there were 413 races in total but we only follow 412 here, since one was uncontested (only one candidate). Of those, 380 races were for the House or Representatives and 32 for the Senate. By comparison, in the 2010 elections, there were 441 races, 408 for the House and 33 for the Senate. The increased number of House seats reflects increase in population and redistricting. Since our analysis is identical for these two categories, we will be reporting our results together.

As it usually happens in the US elections, not all races were highly contested, according to the political media. Sixty-one (61) of them were considered highly contested by the media during the last month before the 2008 election, and 125 of the 2010 elections. In the remaining races, one of the candidates was so far ahead in the polls that it was unlikely to lose the race. In our results we study both of these categories.

Each of the two Congressional elections proved to be a "landslide win" for one of the parties, since the US electoral system is a "winner-takes-all" system: Even a small advantage of one party can win a much larger majority of seats. In the 2008 elections, the Democrats were the winners, while in the 2010 elections, the Republicans won, as Figure 1 below indicates.
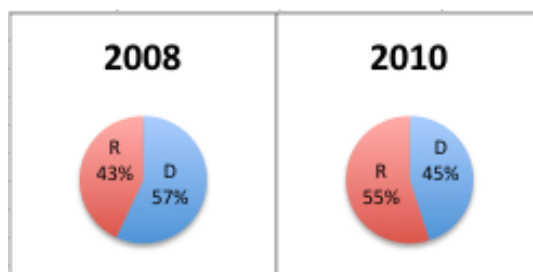

FIGURE 1: Elections in 2008 and 2010 won "in a landslide" but by a different party.

Traditionally, incumbents (candidates who were holding the seat in the previous Congress) have fared very well, and these elections were no exception: the overall incumbency re-election rate in 2008 was **91.6%,** while in 2010, when the anti-incumbent sentiment was running high, the rate dropped to **84.5%.** We decided to use the incumbent re-election rate as a lower bound of predicting the election, since one can do a very well in predicting the outcome of the elections just by betting on incumbents: anyone would like to bet on a game that gives a chance of winning over 80% in the worst case. The question is, can other methods of predicting, such as the G-trends, do better?

We retrieved the G-trends of both 2008 and 2010 multiple times. We observed that, over time, the reported numbers changed a little, as expected by the sampling method that Google employs in collecting them, but the results of our analysis did not change significantly.

To help the visualization of the results and the trends, we also developed the visual interface shown in Figure 2 (at http://cs.wellesley.edu/~webtrust/insights/)
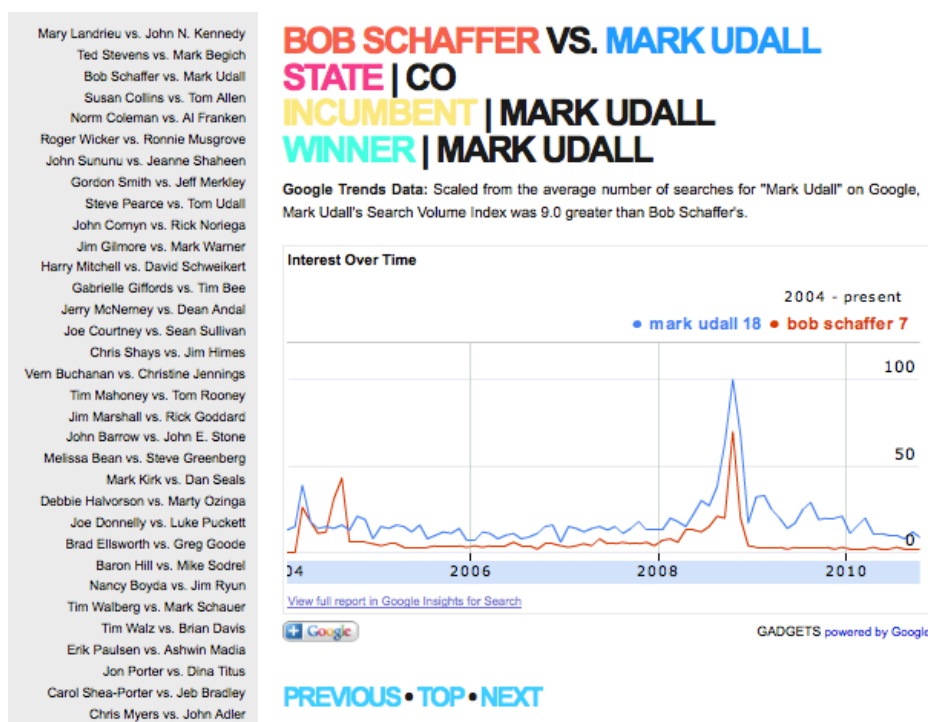

FIGURE 2: A visual interface for exploring the G-trends between pairs of candidates in the 2008 Congressional elections.

3.1   Candidate visibility on G-trends

Comparing the election results to the Trends during the last week before the elections, G-trends predicted correctly only a small fraction of them (121 races in 2008 and 99 in 2010).
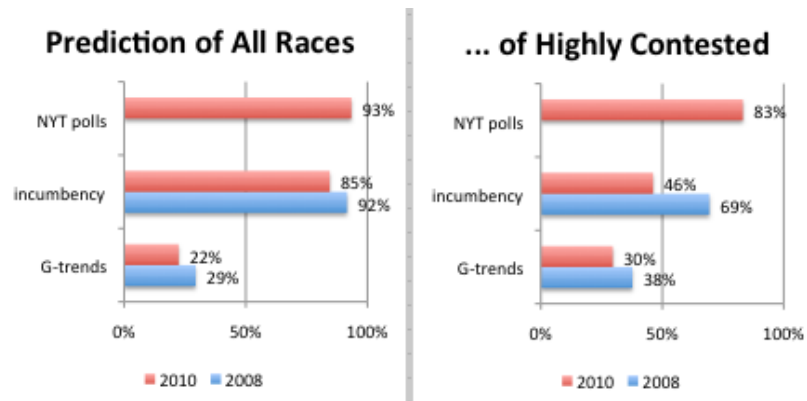
FIGURE 3: Including races that had candidates with no visibility on G-trends is not a fair way of assessing its predictive power, though it is fair for incumbency and NY Times polls. (For the latter, we did not have data for 2008 since they are removed after an election.)

However, this is not a fair assessment for G-trends, because not all of the races had candidates who were recorded in the trends (due to lack of visibility in Web searches).

To evaluate predictions in a fair way, we first removed from consideration group **T0**: the 203 races in 2008 and 223 in 2010 for which there were no G-trends recorded for either of the two candidates. It turns out that the incumbent re-election rate was 97.7% for this group. In the 2010 elections, however, anti-incumbency sentiments seemed to run high (Pew, 2010). Yet, even in this case 92.7% of incumbents got re-elected. The picture may be more complicated than this, as we will see.

Separating the T0 group is useful in defining the other two categories: **T1**, where only one of the candidates had trends recorded, and **T2** where both had trends recorded. One could expect that T1 should be a landslide for G-trends predictions, if higher visibility indicates positive attitude towards a candidate. By the same token, T2 should do well in terms of the trend predictions and be expected to correlate with the actual voting results.

3.2   T1: Single candidate visibility

We now consider the 115 races in 2008 and 102 races in 2010 for which only one of the two candidates had non-0 visibility in G-trends. The winners were not evenly divided between the two parties: Democrats did surprisingly well in both elections, winning 80 (69.6%) races in 2008 and 56 (54.9%) in 2010.
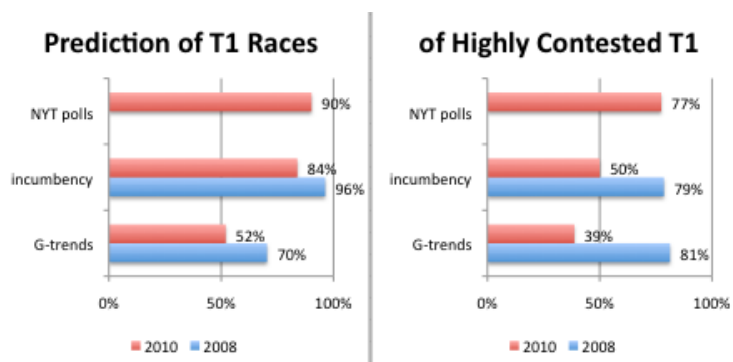

FIGURE 4: G-trends did not compare well with both incumbency and NYT predictions for the group T1 that included races with a single candidate having no visibility. It did quite well for the highly contested races in 2008, but quite poorly in 2010.

G-trends predicted correctly 81 (70%) of those races in 2008 and 53 (52%) in 2010, not a remarkable achievement considering the high rates of incumbency and polling. Its record is, however, surprisingly better when one considers the highly contested races in 2008: G-trends correctly predicted 13 of the 16 highly contested races (81%) beating narrowly even the incumbents' prediction (79%). Was it something in particular that we are observing or was just random luck?

The 2010 results indicate that it was just luck. In those elections G-trends correctly predicted only 39%, far less than chance, which was what incumbency predicted. Highly contested races in which one candidate has no visibility on Web searches may reveal simply a badly run campaign by the non-visible opponent.

3.3 T2: Both visible candidates

Finally, we are taking a look at the races in which both candidates had visibility (non-0 in G-trends). There were 94 such races in 2008 and 116 in 2010.
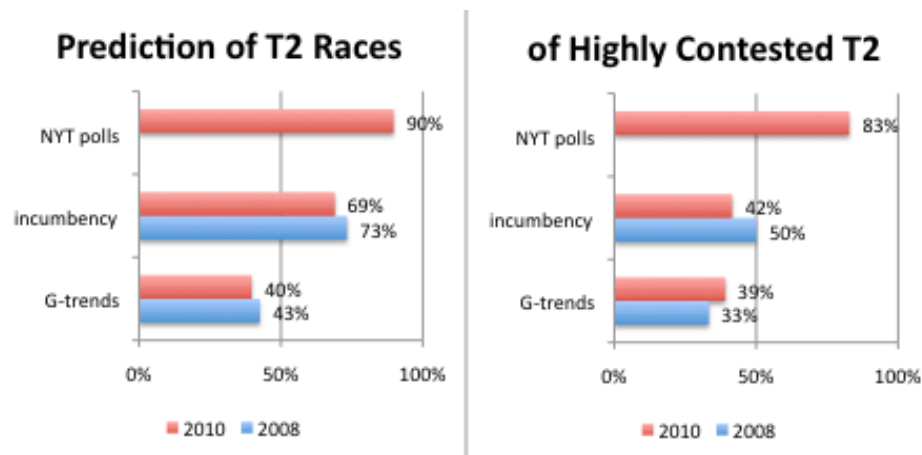

FIGURE 5: For the most competitive group T2 for which both candidates had visibility, G-trends did not compare well with incumbency. In the all-important high contested races, its performance was ever worse. The polls are proven to be the best way to make good predictions.

As one might expect, these were races harder to predict, but G-trends prediction of these races overall was worse than chance: 43% in 2008 and 40% in 2010. By comparison, incumbency was still able to predict 73% (2008) and 69% (2010).

The picture is even worse in the races that were considered highly contested which attract a lot of attention in the media: G-trends could predict correctly the outcome of only one in three (33.3%) in 2008 and 39% in 2010. Incumbency was able to do as well as chance (50%) in 2008 and worse than that (42%) in 2010. For those races, the polls was the best predictor by far, even though, for all the money and effort that goes into their collection, their prediction was only 83%.

We also looked at the correlation between the difference in votes and in the trend data. For the T1 group, which was the most promising for prediction, there was virtually no correlation for the non-contested races (0.02) while for the highly contested races it was moderate (0.68). See Figure 6 for the latter.
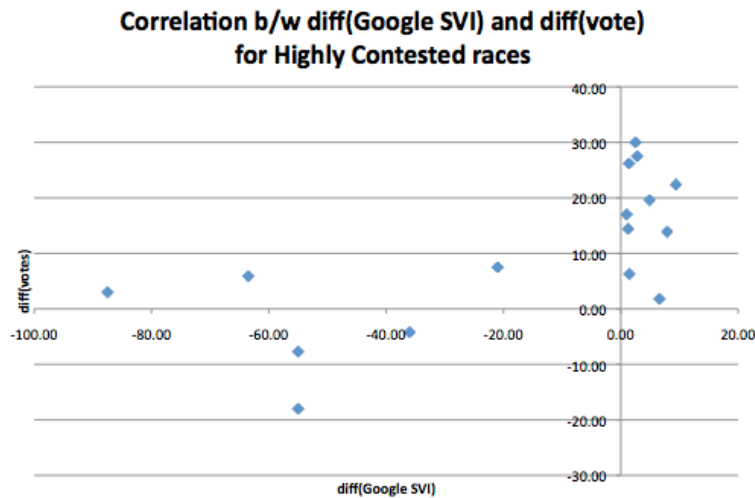
FIGURE 6: Correlation for the 2008 sixteen (16) highly contested races in the group T1 (where only one of the candidates had non-0 visibility in G-trends).

3.4   Discussion of Results

So, overall, G-trends was not a good predictor of both the 2008 and 2010 US congressional elections. Still, this performance begs some explanation.

One could argue that, for the group T1, when people search on Google for the name of a candidate but not for the name of the opponent, that is good news for the candidate: he or she has some high visibility while the opponent's candidacy is not raised even to the level of curiosity. In this case we may have indication of an unevenly run race and the Google search visibility reflect this fact.

On the other hand, for the group T2, especially for the highly contested races, Google searches may be a sign of trouble. It has been documented (Mustafaraj & Metaxas, 2009) that there is significant political spam during the elections. (In fact, as of the writing of this, a series of Google bombs have been openly announced, as reported in http://www.politico.com/news/stories/1010/43767.html). One possible explanation is that in the T2 categories, people who search for the candidates' names may be finding negative information and may be a sign of the trouble that is coming at the polls.

4. CONCLUSIONS

Comparing the set of candidates who won the 2008 congressional elections to the set of candidates who were queried more frequently than their opponent on Google did not reveal evidence of a strong correlation between search popularity and likelihood of winning. This was repeated during the 2010 Congressional elections, in which a different party won. Therefore, we have strong evidence of the limitations of the predictive power of G-trends when it comes to elections. The only exception observed was in races when one of the candidates has not utilized the Web and social media in gaining exposure in cyberspace. However, this category of candidates will not last for long as more people use the Web for being informed and deliberating on voting and more candidates use the Web and social media in their election strategy.

A variable that may have affected G-trends effectiveness as a tool for predicting political elections is the sentiment of a user's query. It is difficult, though not impossible, to determine the circumstances behind a user's search of the profile of a certain candidate to make a guess about that candidate's public image and why a user might be interested in the candidate. This is part of future research that we plan for the next stage of our work.

One question that remains largely open is, how can one categorize the types of events that can actually be measured correctly or even predicted by G-trends?

ACKNOWLEDGEMENT

REFERENCES

Asur, S. & Huberman, B. A. (2010). Predicting the Future With Social Media.. In http://arxiv.org/abs/1003.5699.

boyd. d., Golder, S., Lotan, G., 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In HICSS-43. Kauai, HI. IEEE.

Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K. P., 2010. Measuring User Influence in Twitter: The Millon Follower Fallacy. In 4th International AAAI Conference on Weblogs and Social Media (ICWSM).

Choi, H., Varian, H., 2009. Predicting the Present with Google Trends. In Google Research Blog, Thursday April 2, 2009. http://googleresearch.blogspot.com/ 2009/04/predicting-present-with-google-trends.html.

Ginsberg, J, Mohebbi, M., Patel, R, Brammer, L., Smolinski, M., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. In Nature, Vol. 457, 19 Feb. 2009; doi:10.1038/nature07634.

Google, 2010. How does Google Trends Work, Official Site, http://www.google.com/intl/en/trends/about.html

Huberman, B. A., Romero, D. M., & Wu F., 2009. Social networks that matter: Twitter under the microscope. In First Monday, 14(1), 2009.

Java, A., Song, X., Finin, T. & Tseng, B., 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In 9th WebKDD and 1st SNA-KDD workshop on web mining and social network analysis.

Khrabrov, A., Cybenko, G., 2010. Discovering Influence in Communication Networks using Dynamic Graph Analysis. In SocialComm 2010.

Lampos, V. & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the Social Web. In IAPR 2nd Workshop on Cognitive Information Processing.

Metaxas, P., Mustafaraj, E., 2010. From Obscurity to Prominence: Political Speech and Real Time Search. In Web Science 2010. Raleigh, NC.

Mustafaraj, E., Metaxas, P., 2009. The Battle for the 2008 US Congressional elections on the Web. In WebScience 2009, Athens, Greece.

O'Connor, B., Balasubramanyan, R., Routledge, B., Smith, A., 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In 1st International AAAI Conference on Weblogs and Social Media. Washington, DC, May.

Pew, 2010. GOP Likely to Recapture Control of House. In Pew Research Center for the People & the Press. October 31,2010, http://people-press.org/report/671/ The Pew Foundation.

Romero, D., Galuba, W., Asur, S. Huberman, B., 2010. Influence and passivity in social media. In http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1653135.

Smith, A Rainie, L., 2008. The internet and the 2008 election. In Pew Internet and American Life Project. June 15, 2008, The Pew Foundation.

Tumasjan, A., Sprenger, T. O., Sandner P. G. & Welpe I. M., 2010. Predicting Elections with Twitter: Whan 140 Characters Reveal about Political Sentiment. In 4th International AAAI Conference on Weblogs and Social Media (ICWSM).

Zhang, X., Fuehres, H., Gloor, P., 2010. Predicting Stock Market Indicators Through Twitter. In Procedia – Social and Behavioral Sciences. Elsevier.