In: Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks, Avignon, France, 2012.
http://ifarm.nl/erikt/papers/sasn2012.pdf

Predicting the 2011 Dutch Senate Election Results with Twitter

Erik Tjong Kim Sang and **Johan Bos**
Alfa-informatica
University of Groningen
Groningen, The Netherlands
{e.f.tjong.kim.sang,johan.bos}@rug.nl

Abstract

To what extend can one use Twitter in opinion polls for political elections? Merely counting Twitter messages mentioning political party names is no guarantee for obtaining good election predictions. By improving the quality of the document collection and by performing sentiment analysis, predictions based on entity counts in tweets can be considerably improved, and become nearly as good as traditionally obtained opinion polls.

1 Introduction

Predicting the future is one of human's greatest desires. News companies are well aware of this, and try to predict tomorrow's weather and changes on the stock markets. Another case in point are the opinion polls, of which the news is abundant in the period before political elections. Such polls are traditionally based on asking a (representative) sample of voters what they would vote on the day of election.

The question we are interested in, is whether opinion polls could be conducted on the basis of the information collected by Twitter, a popular microblog website, used by millions to broadcast messages of no more than 140 characters, known as *tweets*. Over the last two years, we have collected a multi-billion-word corpus of Dutch tweets, with the general aim of developing natural language processing tools for automatically analyzing the content of the messages in this new social medium, which comes with its own challenges. When the Dutch Senate elections took place in 2011, we took this as an opportunity to verify the predictive power of tweets.

More concretely, we wanted to test whether by simply counting Twitter messages mentioning political party names we could accurately predict the election outcome. Secondly, we wanted to investigate factors that influence the predictions based on the Dutch tweets.

In this paper we present the results of our experiments. We first summarize related work in Section 2. Then we outline our data collection process (Section 3). The methods we used for predicting election results and the obtained results, are presented in Sections 4, 5 and 6. We discuss the results of the experiments in Section 7 and conclude in Section 8.

2 Related work

Tumasjan et al. (2010) investigate how Twitter is used in political discourse and check if political sentiment on Twitter reflects real-life sentiments about parties and politicians. As a part of their study, they compare party mentions on Twitter with the results of the 2009 German parliament election. They conclude that the relative number of tweets mentioning a party is a good predictor for the number of votes of that party in an election. A similar finding was earlier reported by Jean Véronis in a series of blogposts: the number

[1]The data and software used for the experiments described in this paper can be retrieved from http://ifarm.nl/ps2011/p2011.zip

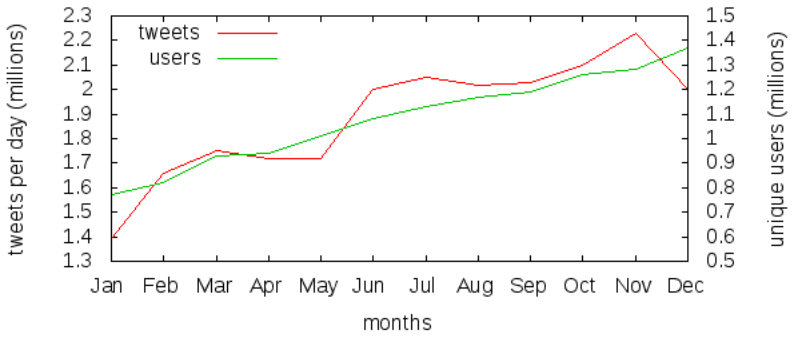

Figure 1: Overview of our collection of Dutch tweets of the year 2011. The data set contains almost 700 million tweets. Both the number of tweets (about two million per day) and the number of unique users (about one million) increase almost every month. The collection is estimated to contain about 37% of the total volume of Dutch tweets.

of times a French presidential candidate was mentioned in the press was a good prediction for his or her election results (Véronis, 2007). This prediction task involved only two candidates, so it was easier than predicting the outcome of a multiparty election.

Jungherr et al. (2011) criticize the work of Tumasjan et al. (2010). They argue that the choice of included parties in the evaluation was not well motivated and show that the inclusion of a seventh party, the Pirate Party, would have had a large negative effect on accuracy of the predictions. Furthermore, Jungherr et al. question the time period which was used by Tumasjan et al. for collecting the tweets and show that including the tweets of the week right before the election would also have had a significant negative effect on the prediction accuracy.

Using Twitter data for predicting election results was popular in 2010 and 2011. Chung and Mustafaraj (2011) found that merely counting tweets is not enough to obtain good predictions and measure the effect of sentiment analysis and spam filtering. O'Connor et al. (2010) discovered that while volumes of mentions of *obama* on Twitter before the US presidential election of 2008 correlated with high poll ratings for Barack Obama, volumes of mentions of his rival *mccain* also correlated with high poll ratings of the election winner. Gayo-Avello et al. (2011) show that predictions based on Twitter only predicted half of the winners of US congressional elections with two candidates correctly, a performance which is not better than chance.

3 Data collection

We collect Dutch Twitter messages (tweets) with the filter stream provided by Twitter. We continuously search for messages that contain at least one of a list of about a hundred high-frequent Dutch words and a dozen frequent Dutch subject tags (hashtags). The results of this process also contain some false positives: tweets that contain apparent Dutch words but are actually written in another language. In order to get rid of these messages, we apply a language guesser developed by Thomas Mangin (Mangin, 2007). It ranks languages by comparing character n-grams of an input text to n-gram models of texts in known languages. We use a set of 74 language models developed by our students in 2007.

In order to estimate the coverage of our selection with respect to all tweets in Dutch, we collected all tweets of one month from 1,017 randomly selected users which predominantly post messages in Dutch. We compared the two data streams and found that the first contained 37% of the data found in the second. This suggests that we collect about 37% of all Dutch tweets. Our data collection process contains two filters: one is based on a word list and the other is the language guesser. The first filter lost 62% of the data while the second lost another 1%.

Party	Short name	Long name	Total	Seats Twitter	Seats PB	Seats MdH	Average polls
PVV	2226	1	2227	18	12	12	12
VVD	1562	0	1562	13	14	16	15
CDA	1504	0	1504	12	9	10	9.5
PvdA	1056	1	1057	9	13	13	13
SP	839	0	839	7	8	7	7.5
GL	243	505	748	6	5	3	4
D66	610	0	610	5	6	5	5.5
CU	159	79	238	2	3	3	3
PvdD	103	51	154	1	1	1	1
SGP	139	0	139	1	2	2	2
50+	6	43	49	0	1	2	1.5
OSF	-	-	-	1	1	1	1
			offset	21	4	4	-

Table 1: Frequencies of tweets mentioning one of 11 main political parties from one day, Wednesday 16 February 2011, converted to Senate seats (column Seats Twitter) and compared with the predictions of two polls from the same week: from Politieke Barometer of 17 February (Synovate.nl, 2011b) and from Maurice de Hond of 15 February (Peil.nl, 2011b). The offset value is the sum of the differences between the Twitter predictions and the average poll predictions. The OSF group is a cooperation of 11 local parties which were not tracked on Twitter.

4 Counting party names

The Dutch Senate elections are held once every four years. The elections are preceded by the Dutch Provincial Election in which the voters choose 566 representatives for the States-Provincial. Three months later the new representatives elect the new Senate. In the second election, each of the representatives has a weight which is proportional to the number of people he or she represents. The 2011 Dutch provincial elections were held on Wednesday 2 March 2011 and the corresponding Senate elections were held on Monday 23 May 2011. In the Senate elections 75 seats are contested.

Our work on predicting the results of this election was inspired by the work of Tumasjan et al. (2010), who report that basic counts of tweets mentioning a political party provided good predictions for the results of the 2009 German parliament election. We decided to replicate their work for the Dutch Senate Elections of 2011.

We started with examining the Dutch tweets of Wednesday 16 February 2011, two weeks prior to the Provincial elections. This data set consisted of 1.7 million tweets. From this data set we extracted the tweets containing names of political parties. This resulted in 7,000 tweets. This number was lower than we had expected. Originally we had planned to use the tweets for predicting local election results. However, further filtering of the tweets to require location information would have left us with a total of about 70 political tweets per day, far too few to make reliable predictions for twelve different provinces.

In the data, we searched for two variants of each party: the abbreviated version and the full name, allowing for minor punctuation and capitalization variation. For nearly all parties, the abbreviated name was used more often on Twitter than the full name. The two exceptions are GroenLinks/GL and 50Plus/50+ (Table 1). Party names could be identified with a precision close to 100% except for the party ChristenUnie: its abbreviation CU is also used as slang for *see you*. This was the case for 11% of the tweets containing the phrase *CU*. In this paper, the 11% of tweets have already been removed from the counts of this party.

Apart from the eleven regular parties shown in Table 1, there was a twelfth party with a chance of winning a Senate seat: the Independent Senate Group (OSF), a cooperation of 11 regional par-

ties. These parties occur infrequently in our Twitter data (less than five times per party per day), too infrequent to allow for a reliable base for predicting election results. Therefore we decided to use a baseline prediction for them. We assumed that the group would win exactly one Senate seat, just like in the two previous elections.

We converted the counts of the party names on Twitter to Senate seats by counting every tweet mentioning a party name as a vote for that party. The results can be found in the column *Seats Twitter* in Table 1. The predicted number of seats were compared with the results of two polls of the same week: one by the polling company Politieke Barometer of 17 February (Synovate.nl, 2011b) and another from the company Peil.nl, commonly referred to as Maurice de Hond, from 15 February (Peil.nl, 2011b). The predicted numbers of seats by Twitter were reasonably close to the numbers of the polling companies. However, there is room for improvement: for the party PVV, tweets predicted a total of 18 seats while the polling companies only predicted 12 and for the party 50+, Twitter predicted no seats while the average of the polling companies was 1.5 seats.

5 Normalizing party counts

The differences between the Twitter prediction and prediction of the polling companies could have been caused by noise. However, the differences could also have resulted from differences between the methods for computing the predictions. First, in the polls, like in an election, everyone has one vote. In the tweet data set this is not the case. One person may have send out multiple tweets or may have tweeted about different political parties. This problem of the data is easy to fix: we can keep only one political tweet per user in the data set and remove all others.

A second problem is that not every message containing a party name is necessarily positive about the party. For example:

> *Wel triest van de vvd om de zondagen nu te schrappen wat betreft het shoppen, jammer! Hierbij dus een #fail*
>
> *Sadly, the VVD will ban shopping on Sundays, too bad! So here is a #fail*

Party	One party per tweet	One tweet per user	Both constraints
PVV	22	17	19
VVD	12	13	13
CDA	12	12	12
PvdA	8	8	8
SP	6	8	7
GL	6	7	7
D66	5	5	5
CU	1	2	2
PvdD	1	1	1
SGP	1	1	0
50+	0	0	0
OSF	1	1	1
offset	29	22	25

Table 2: Senate seat predictions based on normalized tweets: keeping only tweets mentioning one party, keeping only the first tweet of each user and keeping of each user only the first tweet which mentioned a single party. The offset score is the seat difference between the predictions and the average poll prediction of Table 1.

While the tweet is mentioning a political party, the sender does not agree with the policy of the party and most likely will not vote for the party. These tweets need to be removed as well.

A third problem with the data is that the demographics of Dutch Twitter users are probably quite different from the demographics of Dutch voters. Inspection of Dutch tweets revealed that Twitter is very popular among Dutch teens but they are not eligible to vote. User studies for other countries have revealed that senior citizens are underrepresented on the Internet (Fox, 2010) but this group has a big turnout in elections (Epskamp and van Rhee, 2010). It would be nice if we could assign weights to tweets based on the representativeness of certain groups of users. Unfortunately we cannot determine the age and gender of individual Twitter users because users are not required to specify this information in their profile.

Based on the previous analysis, we tested two normalization steps for the tweet data. First, we removed all tweets that mentioned more than one party name. Next, we kept only the first tweet of each user. Finally we combined both steps: keep-

ing of each user only the first tweet which mentioned a single political party. We converted all the counts to party seats and compared them with the poll outcomes. The results can be found in Table 2. The seat predictions did not improve. In fact, the offsets of the three methods proved to be larger than the corresponding number of the baseline approach without normalization (29, 25 and 22 compared to 21). Still, we believe that normalization of the tweet counts is a good idea.

Next, we determined the sentiments of the tweets. Since we do not have reliable automatic sentiment analysis software for Dutch, we decided to build a corpus of political tweets with manual sentiment annotation. Each of the two authors of this paper manually annotated 1,678 political tweets, assigning one of two classes to each tweet: negative towards the party mentioned in the tweet or nonnegative. The annotators agreed on the sentiment of 1,333 tweets (kappa score: 0.59).

We used these 1,333 tweets with unanimous class assignment for computing sentiment scores per party. We removed the tweets that mentioned more than one party and removed duplicate tweets of users that contributed more than one tweet. 534 nonnegative tweets and 227 negative tweets were left. Then we computed weights per party by dividing the number of nonnegative tweets per party by the associated total number of tweets. For example, there were 42 negative tweets for the VVD party and 89 nonnegative, resulting in a weight of 89/(42+89) = 0.68. The resulting party weights can be found in Table 3.

We multiplied the weights with the tweet counts obtained after the two normalization steps and converted these to Senate seats. As a result the difference with the poll prediction dropped from 25 to 23 (see Table 3). Incorporating sentiment analysis improved the results of the prediction.

After sentiment analysis, the tweets still did not predict the same number of seats as the polls for any party. For nine parties, the difference was two and a half seats or lower but the difference was larger for two parties: GL (5) and PvdA (6). A possible cause for these differences is a mismatch between the demographics of Twitter users and the Dutch population. We have no data describing this discrepancy. We wanted to build a model for this difference so we chose to model the difference by additional correction weights based on the seats differences between the two predictions. We based the expected number of seats on the two poll results of the same time period as the tweets (Synovate.nl, 2011b; Peil.nl, 2011b). For example, after normalization, there were 811 tweets mentioning the PVV party. The party has a sentiment weight of 0.49 so the adjusted number of tweets is 0.49*811 = 397. The polls predicted 12 of 74 seats for this party. The associated population weight is equal to the average number of poll seats divided by the total number of seats divided by the adjusted number of tweets divided by the total number of adjusted tweets (2,285): (12/74)/(397/2285) is 0.93.

The population weights can be found in Table 4. They corrected most predicted seat numbers of Twitter to the ones predicted by the polls. A drawback of this approach is that we have tuned the prediction system to the results of polls rather than to the results of elections. It would have been

Party	Tweet count	Sentiment weight	Seats Twitter
PVV	811	0.49	13
VVD	552	0.68	13
CDA	521	0.70	12
PvdA	330	0.69	7
SP	314	0.90	9
GL	322	0.81	9
D66	207	0.94	6
CU	104	0.67	2
PvdD	63	1.00	2
SGP	39	0.86	1
50+	17	0.93	0
OSF	-	-	1
		offset	23

Table 3: Sentiment weights per party resulting from a manual sentiment analysis, indicating what fraction of tweets mentioning the party is nonnegative and the resulting normalized seat predictions after multiplying tweet counts with these weights. The second column contains the number of tweets per party after the normalization steps of Table 2.

Party	Population weight	Seats Twitter	Average polls
PVV	0.93	12	12
VVD	1.23	15	15
CDA	0.80	10	9.5
PvdA	1.76	13	13
SP	0.82	8	7.5
GL	0.47	4	4
D66	0.87	5	5.5
CU	1.33	3	3
PvdD	0.49	1	1
SGP	1.84	2	2
50+	2.93	1	1.5
OSF	-	1	1
	offset	2	-

Table 4: Population weights per party resulting from dividing the percentage of the predicted poll seats (Synovate.nl, 2011b; Peil.nl, 2011b) by the percentage of nonnegative tweets (Table 3), and the associated seat predictions from Twitter, which are now closer to the poll predictions. Offsets are measured by comparing with the average number of poll seats from Table 1.

Party	Result	Seats PB	Seats MdH	Seats Twitter
VVD	16	14	16	14
PvdA	14	12	11	16
CDA	11	9	9	8
PVV	10	11	12	10
SP	8	9	9	6
D66	5	7	5	8
GL	5	4	4	3
CU	2	3	3	3
50+	1	2	2	2
SGP	1	2	2	2
PvdD	1	1	2	2
OSF	1	1	0	1
offset	-	14	14	18

Table 5: Twitter seat prediction for the 2 March 2011 Dutch Senate elections compared with the actual results (Kiesraad.nl, 2012a) and the predictions of two polling companies of 1 March 2011: PB: Politieke Barometer (Synovate.nl, 2011a) and MdH: Maurice de Hond (Peil.nl, 2011a).

better to tune the system to the results of past elections but we do not have associated Twitter data for these elections. Adjusting the results of the system to get them as close to the poll predictions as possible, is the best we can do at this moment.

6 Predicting election outcomes

The techniques described above were applied to Dutch political tweets collected in the week before the election: 23 February 2011 – 1 March 2011: 64,395 tweets. We used a week of data rather than a day because we expected that using more data would lead to better predictions. We chose for a week of tweets rather than a month because we assumed that elections were not an important discussion topic on Twitter one month before they were held.

After the first two normalization steps, one party per tweet and one tweet per user, 28,704 tweets were left. The parties were extracted from the tweets, and counted, and the counts were multiplied with the sentiment and population weights and converted to Senate seats. The results are shown in Table 5 together with poll predictions (Synovate.nl, 2011a; Peil.nl, 2011a) and the results of the elections of 2 March 2011 (Kiesraad.nl, 2012a).

The seat numbers predicted by the tweets were close to the election results. Twitter predicted the correct number of seats for the party PVV while the polling companies predicted an incorrect number. However the companies predicted other seat numbers correctly and they had a smaller total error: 14 seats compared to 18 for our approach.

In Dutch elections, there is no strict linear relation between the number of votes for a party and the number seats awarded to a party. Seats that remain after truncating seat numbers are awarded to parties by a system which favors larger parties (Kiesraad.nl, 2012b). Furthermore, in 2011 there was a voting incident in the Senate elections which caused one party (D66) to loose one of its seats to another party (SP). In our evaluation we have compared seat numbers because that is the only type of data that we have available from the polling companies. The election results allow a comparison based on percentages of votes. This comparison is displayed in Table 6.

Party	Result	Twitter	offset
VVD	19.6%	17.3%	-2.3%
PvdA	17.3%	20.8%	+3.5%
CDA	14.1%	11.0%	-3.1%
PVV	12.4%	13.3%	+0.9%
SP	10.2%	8.5%	-1.7%
D66	8.4%	10.1%	+1.7%
GL	6.3%	4.8%	-1.5%
CU	3.6%	4.0%	+0.4%
50+	2.4%	3.1%	+0.7%
SGP	2.4%	3.1%	+0.7%
PvdD	1.9%	2.7%	+0.8%
OSF	1.4%	1.3%	-0.1%
offset	-	17.4%	

Table 6: Twitter vote prediction for the 2 March 2011 Dutch Provincial elections compared with the actual results in percentages[2].

Party	Result	Seats Twitter	Population weight
VVD	16	16	2.23
PvdA	14	13	1.93
CDA	11	10	1.41
PVV	10	12	1.78
SP	8	7	1.11
D66	5	5	0.82
GL	5	4	0.59
CU	2	3	0.45
50+	1	1	0.22
SGP	1	2	0.30
PvdD	1	1	0.15
OSF	1	1	-
offset	-	8	

Table 7: Seat prediction for the 2 March 2011 Dutch Senate elections based on an uniform distribution of tweets mentioning political parties.

With the exception of the three largest parties, all predicted percentages are within 1.7% of the numbers of the election. The percentages might prove to be more reliable than seat numbers as a base for a election prediction method. We hope to use percentage figures when the predicting the outcome of next parliament elections: one of the polling companies publishes such figures with their predictions of parliament elections.

7 Discussion

Although we are happy about the accuracy obtained by the Twitter predictions, we have some concerns about the chosen approach. In Table 4, we introduced poll-dependent weights to correct the demographic differences between the Twitter users and the Dutch electorate. This was necessary because we did not have information about the demographics of Twitter users, for example about their gender and age. As already mentioned, this choice led to tuning the system to predicting poll results rather than election results. But do the population weights not also minimize the effect that tweet counts have on the predictions? Does the system still use the tweet counts for the election prediction?

In order to answer the latter question, we designed an additional experiment. Suppose the tweets per party were uniformly distributed such that each party name appeared in the same number of tweets each day. This would make tweet counts uninteresting for predicting elections. However, how would our system deal with this situation? The results of this experiment are shown in Table 7.

Since we did not have data to base sentiment weights on, we assumed that all the sentiment weights had value 1.0. Since the tweet counts were different from those in the earlier experiments, we needed to compute new population weights (see Table 7). The seat numbers predicted by the system were equal to the average of the seat numbers of the two polls in Table 4 plus or minus a half in case the two numbers added up to an odd number. The VVD party gained one seat, as a consequence of the system of awarding remainder seats to larger parties. We assume that the tweet distribution will be uniform at all times and this means that the system will always predict the seat distribution. The offset of the new prediction was 3 seats for the test distribution of Table 4 and 8 seats for the election results (see Table 7), a

[2]CU and SGP were awarded an additional 0.3% and 0.2% for the 0.5% they won as an alliance.

smaller error than either of the polling companies (compare with Table 5).

This experiment has produced a system which generates the average of the predictions of the two polling companies from the week of 16/17 February as an election prediction. It does not require additional input. This is not a good method for predicting election outcome but by chance it generated a better prediction than our earlier approach and those of two polling companies. We are not sure what conclusions to draw from this. Is the method of using population weights flawed? Is our evaluation method incorrect? Are tweets bad predictors of political sentiment? Is the margin of chance error large? It would be good to test whether the measured differences are statistically significant but we do not know how to do that for this data.

8 Concluding remarks

We have collected a large number of Dutch Twitter messages (hundreds of millions) and showed how they can be used for predicting the results of the Dutch Senate elections of 2011. Counting the tweets that mention political parties is not sufficient to obtain good predictions. We tested the effects of improving the quality of the data collection by removing certain tweets: tweets mentioning more than one party name, multiple tweets from a single user and tweets with a negative sentiment. Despite having no gold standard training data, the total error of our final system was only 29% higher than that of two experienced polling companies (Table 5). We hope to improve these results in the future, building on the knowledge we have obtained in this study.

Acknowledgements

We would like to thank the two reviewers of this paper for valuable comments.

References

Jessica Chung and Eni Mustafaraj. 2011. Cam collective sentiment expressed on twitter predict political elections? In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. San Francisco, CA, USA.

Martijn Epskamp and Marn van Rhee. 2010. Analyse opkomst gemeenteraadsverkiezingen 2010.

Susannah Fox. 2010. Four in ten seniors go online. Pew Research Center, http://www.pewinternet.org /Commentary/2010/January/38-of-adults-age-65-go-online.aspx (Retrieved 8 March 2012).

Daniel Gayo-Avello, Panagiotis Metaxas, and Eni Mustafaraj. 2011. Limits of electoral predictions using social media data. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Barcelona, Spain.

Andreas Jugherr, Pascal Jürgens, and Harald Schoen. 2011. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, t. o., sander, p. g., & welpe, i. m. 'predicting elections with twitter: What 140 characters reveal about political sentiment'. *Social Science Computer Review*.

Kiesraad.nl. 2012a. Databank verkiezingsuitslagen. http://www.verkiezingsuitslagen.nl/Na1918/Verkiezingsuitslagen.aspx?VerkiezingsTypeId=2 (retrieved 27 February 2012).

Kiesraad.nl. 2012b. Toewijzing zetels. http://www.kiesraad.nl/nl/Onderwerpen/Uitslagen/Toewijzing_zetels.html (retrieved 27 February 2012).

Thomas Mangin. 2007. ngram: Textcat implementation in python. http://thomas.mangin.me.uk/.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Washington DC, USA.

Peil.nl. 2011a. Nieuw haags peil 1 maart 2011. http://www.peil.nl/?3182 (retrieved 5 March 2012).

Peil.nl. 2011b. Nieuw haags peil 15 februari 2011. http://www.peil.nl/?3167 (retrieved 1 March 2012).

Synovate.nl. 2011a. Nieuws 2011 - peiling eerste kamer - week 9. http://www.synovate.nl/content.asp? targetid=721 (retrieved 5 March 2012).

Synovate.nl. 2011b. Peiling eerste kamer - week 7. http://www.synovate.nl/content.asp?targetid=713 (retrieved 5 March 2012).

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth AAAI conference on Weblogs and Social Media*, pages 178–185.

Jean Véronis. 2007. 2007: La presse fait á nouveau mieux que les sondeurs. http://blog.veronis.fr /2007/05/2007-la-presse-fait-nouveau-mieux-que.html.