Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), pages 8–15
Marseille, 24 June 2022

Correlating Political Party Names in Tweets, Newspapers and Election Results

Eric Sanders, Antal van den Bosch
CLS/CLST Radboud University, KNAW Meertens Institute
e.sanders@let.ru.nl, antal.van.den.bosch@meertens.knaw.nl

Abstract

Twitter has been used as a textual resource to attempt to predict the outcome of elections for over a decade. A body of literature suggests that this is not consistently possible. In this paper we test the hypothesis that mentions of political parties in tweets are better correlated with the appearance of party names in newspapers than to the intention of the tweeter to vote for that party. Five Dutch national elections are used in this study. We find only a small positive, negligible difference in Pearson's correlation coefficient as well as in the absolute error of the relation between tweets and news, and between tweets and elections. However, we find a larger correlation and a smaller absolute error between party mentions in newspapers and the outcome of the elections in four of the five elections. This suggests that newspapers are a better starting point for predicting the election outcome than tweets.

Keywords: social media, Twitter, newspapers, elections

1. Introduction

For over a decade researchers have attempted to predict (political) election results on the basis of Twitter. Some results from the beginning of that period looked promising (Tumasjan et al., 2010; Sanders and Van den Bosch, 2013), but soon papers appeared that expressed doubts about the ability to correctly predict the election outcome based on tweets (Gayo-Avello, 2012). Although there are findings that support these doubts (e.g. (Sanders and van den Bosch, 2019)), still studies appear that report on attempts to forecast the elections with tweets, often including sentiment analysis (Nugroho, 2021; Batra et al., 2020; Rao et al., 2020) and others with mixing in additional information, such as economic indicators (Liu et al., 2020).

In his paper on the predictive power of tweets with regards to election results, Murthy concludes: "Twitter frequency and sentiment are hardly measures of 'victory'. They are better indicators of the social media 'buzz' around a candidate. Twitter also tends to act as a reactive rather than predictive media platform." (Murthy, 2015). Based on this finding we intended to investigate whether the mentioning of party names in tweets might be more influenced by what Twitter users hear in the media than their political preference. We did this by studying the correlation of the mentions of political party names in tweets and party mentions in the news, and compare these to the correlation of party mentions in tweets and election results. If Murthy's conclusion is right, we expect the former correlation to be larger than the latter.

For news we restricted ourselves to newspaper articles. The reason for this is that this is a relative limited textual resource that is relatively well accessible and searchable, in contrast with for example television or radio news broadcasts. We are supported in this choice by Druckman who writes in his paper "More important, I find that newspapers, and not television news, play a significant, although potentially limited, role in informing the electorate." (Druckman, 2005).

In earlier studies about the relation between tweets and newspapers, we find opposing findings. In 2015 Murthy concludes "Using the 2011–2012 U.S. Republican primary as a case study, this article evaluates whether the sentiment of traditional print media coverage of candidates is related to the frequency of their mentions on Twitter. We found that the two are generally not related." (Murthy and Petto, 2015), where Su finds in a study about climate change in the news in 2019 "The findings imply that Twitter is more likely to influence newspapers' agenda in terms of breaking news, whereas newspapers are more likely to lead Twitter's agenda in terms of ongoing discussions during non-breaking news periods." and "Overall, the agendas of Twitter and newspapers were significantly correlated." (Su and Borah, 2019).

To investigate our research question whether tweets are more influenced by news than by political preference we counted how often political party names occur in tweets, newspaper articles and how the parties score in the elections of five Dutch elections of national importance, and compare their percentages. The paper is organised as follows: in section 2 we present how we got our data, in section 3 we explain how we conducted our experiment, in section 4 we show our results and in sections 5 and 6 we discuss our findings and draw conclusions.

2. Data

2.1. TwiNL

The tweets we used in our study are taken from TwiNL (Tjong Kim Sang and Van den Bosch, 2013), a project in which Dutch tweets are collected since December 2010. The archive creators claim a coverage of about 60% to 80% of all Dutch tweets (based on the number of replies to a tweet that also appears in the collection). These are tweets that are either in the Dutch language or posted by a set of users known to post in Dutch. Language detection separates the Dutch from the non-Dutch tweets. In our experiments we only use the tweets that were detected as written in Dutch, which is not flawless, but sufficiently accurate for trustworthy numbers. Until February 2021, over 4.1 billion Dutch written tweets were collected.

2.2. LexisNexis

For newspaper articles we used a huge online collection of Dutch newspapers provided by LexisNexis. It has a special service for academia, called LexisUni (Knapp, 2018). It contains an archive of forty years of weekly and daily newspapers. All major Dutch national newspapers (*Telegraaf, Volkskrant, Algemeen Dagblad, NRC, Parool, Trouw, Financieel Dagblad*) and many regional newspapers are present[1]. In contrast to the tweets we do not have the texts of the newspaper articles. We use the search engine of LexisNexis that returns the number of newspapers in which a search term was found within an indicated date range. This number was used in our experiments.

2.3. Elections and Parties

We studied five Dutch elections of national importance. In 2012 and 2017 elections were held for the *Tweede Kamer* (comparable to the House of Representatives in the USA) and in 2011, 2015 and 2019 elections were held for the *Eerste Kamer* (comparable to the Senate in the USA). Eleven political parties participated in all five elections. These are the parties that were taken into account in our experiments. Table 1 shows the eleven parties. See (Sanders and van den Bosch, 2019) for a more detailed description of the Dutch electoral system and the various political parties.

3. Experiments

3.1. Counting Political Party Names

To find the (”political”) correlation between tweets and newspapers, we count how often political party names appear in them. We use case insensitive pattern matching of different manners of writing of the party names. For tweets we use more elaborate regular expression to find the party names, for an extensive description, see (Sanders and van den Bosch, 2019). For the newspapers, we use a simpler set, because newspaper are much more unambiguous in their way of spelling party names and misspelling will be so infrequent that they can safely be ignored. Table 1 shows the political parties that gained at least one seat in all five elections under study. In 2017 and 2019, two other parties also gained seats in the elections: FvD and DENK. We decided to not include these in our experiments for two reasons: 1) By having the same set of parties over all elections makes it much easier to compare between the different elections. 2) ’Denk’ is also a conjugation of the Dutch verb ’Denken’ (to think), which is very common in the Dutch language. For tweets, we can disambiguate between the party name and the verb by means of automatic classification, but for newspaper articles this is not possible, because we do not have the texts of the articles.

We did some sample searches with all party names that might be used in the newspapers (also with their full names) and for most parties only their common abbreviation was sufficient to catch almost all news paper articles in which this party was mentioned. For a few parties we needed both the abbreviation as well as the full name. Note that the party *50Plus* is sometimes also written as *50+*, but this is not a possible search term in LexisNexis, because the plus-sign is ignored. Our estimation is that we did not miss many newspaper articles because of this.

Table 1: Search terms in LexisNexis of the political parties.

Party Name	Search Terms LexisNexis (case insensitive)
VVD	VVD
PvdA	PVDA
CDA	CDA
PVV	PVV
SP	SP
D66	D66
GroenLinks	GroenLinks ”Groen Links” GL
ChristenUnie	ChristenUnie ”Christen Unie” CU
50Plus	50Plus ”50 Plus”
SGP	SGP
PvdD	PvdD ”Partij voor de Dieren”

3.2. Correlation and Absolute Error

To determine the correlation between the number of party names mentioned in tweets on the one hand and in newspaper articles on the other hand, we counted mentions of the names in a period of ten days before election day. This period is long enough to smooth out fluctuations in reporting about specific parties, effects of one source influencing the other and the fact that in the Netherlands newspapers do not appear on Sundays. It is also short enough to make sure that the mentioning of parties is likely related to the elections.

We decided to take only singular copies of tweets and newspaper articles into account. Thus, we leave out

[1] https://www.lexisnexis.nl/over-lexisnexis/dutch-news-content

all retweets and replies to a tweet in which a party is mentioned out of our counts; also, we count identical articles, in which a party is mentioned, that appear in several newspapers as one. It is to be expected that including duplications will normalise over all parties and an earlier study showed that there is no substantial difference in including or excluding retweets with respect to the relation between party mentions in tweets and the outcome of elections (Sanders and van den Bosch, 2019).

Figure 1 shows the number of tweets per day in the ten days before election day. Retweets and replies to tweets are excluded from this set. For the elections in 2011 and 2012 there are considerably more tweets in the set than for the later elections, although we will see later that the number of tweets with party names in them are more comparable over the years. For every day and for every election there are at least 450,000 tweets in the collection.

Our research question as posed in the introduction is whether the correlation between political parties mentioned in tweets and in newspaper articles is bigger than the correlation between parties mentioned in tweets and the outcome of elections. To complete the triangular relation between these measurements, we also computed the correlation between parties mentioned in newspaper articles and the outcome of elections and compared these to the other two correlations.

We use two measurements to investigate the relationship between tweets and newspaper articles: Pearson correlation and Absolute error. Pearson correlation (Benesty et al., 2009) is a well known way to indicate the strength of the relation between two series of numbers. In our case we relate the percentages of the number of times the parties are mentioned in two different sources, tweets and newspaper articles.

The absolute error is a measurement used to express the difference between a measured value and a real value. In earlier studies, we used this measurement to compute the distance between a prediction and the real outcome of elections (Sanders and van den Bosch, 2019). In these experiments we use the absolute error to measure the relation between mentions of political parties in tweets and newspapers. See equation 1 for the computation of the absolute error.

$$AE = \sum_{i=1}^{N} |Perc_1(i) - Perc_2(i)| \qquad (1)$$

Where AE is the Absolute Error, $Perc_1(i)$, the percentage of the mentions of party i in data stream 1, $Perc_2(i)$ the percentage of mentions of party i in data stream 2 and N is the total number of parties.

4. Results

The total number of tweets and newspaper articles in which one or more political parties were mentioned in the ten days before the elections are shown in Table 2. For newspaper articles, these numbers vary roughly between 7,000 and 16,000. For tweets these numbers are a factor 30 higher and vary roughly between 200,000 and 700,000. Figure 2 shows the number of tweets with party names per day in the ten days before the elections. From Table 2 and Figure 2 it can be observed that in 2012 and 2017 most tweets and newspaper articles with party names are found. This is to be expected, since these were the years that the elections for the *Tweede Kamer* took place, which are the most important elections in the Netherlands. In Figure 2 it can be seen that in the last one or two days before election day the number of tweets in which a party is mentioned increase substantially, which is also to be expected.

Table 2: Number of tweets and newspaper articles in which political parties were mentioned, for five elections.

year	#tweets	#newspaper articles
2011	304,933	11,175
2012	570,452	15,917
2015	413,967	10,928
2017	669,515	13,561
2019	206,159	7,261
total	2,165,026	58,842

For the five elections the percentages of party mentions in tweets and newspaper articles and the percentages of votes per party can be found in Figures 3, 4 and 5 respectively.

Comparing these figures it becomes apparent that they correlate to some extent. The largest parties (VVD, PvdA, PVV, CDA) have the largest percentages in all graphs, while the smaller parties (PvdD, SGP, 50Plus) are represented by small percentages in all graphs. Figure 6 confirms the visual correlation, showing Pearson's correlation coefficient for the three data pairs (news-tweets, tweets-elections, news-elections) for the five elections.

All Pearson's correlation coefficients lie between 0.67 and 0.95, which means that there is always at least a strong correlation. The correlation between newspaper articles and tweets is almost equal or higher than the correlation between tweets and election results in all cases. The hypothesis that tweets and news are more correlated than tweets and elections is not falsified by these results, but the differences are minimal.

At the same time we find that the correlation between newspaper articles and the election outcome is the highest in four of the five elections. This effect is clearer from Figure 7 in which the absolute errors for the three data pairs for the five elections are shown.

Figure 7 shows the same pattern as Figure 6: where the Pearson's correlation coefficient is higher, the absolute error is lower and vice versa. We observe that the absolute error of the news-elections relation is markedly lower in all elections except the one in 2012.

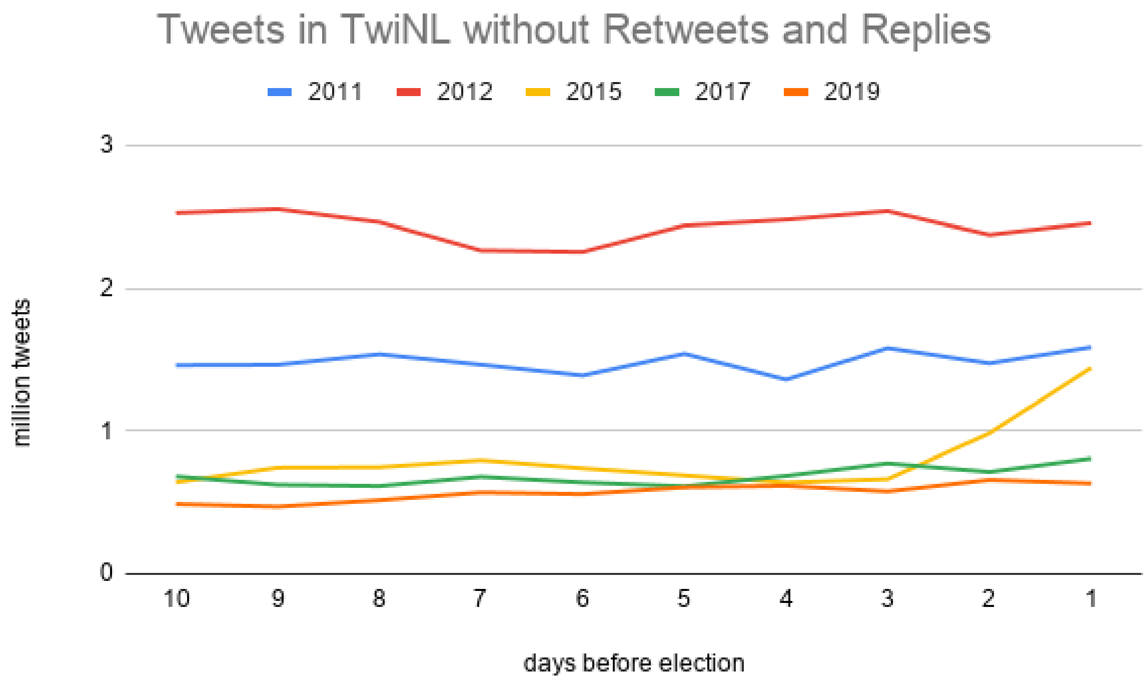

Figure 1: Number of million Dutch tweets per day in TwiNL in the 10 days before the election, excluding retweets and replies

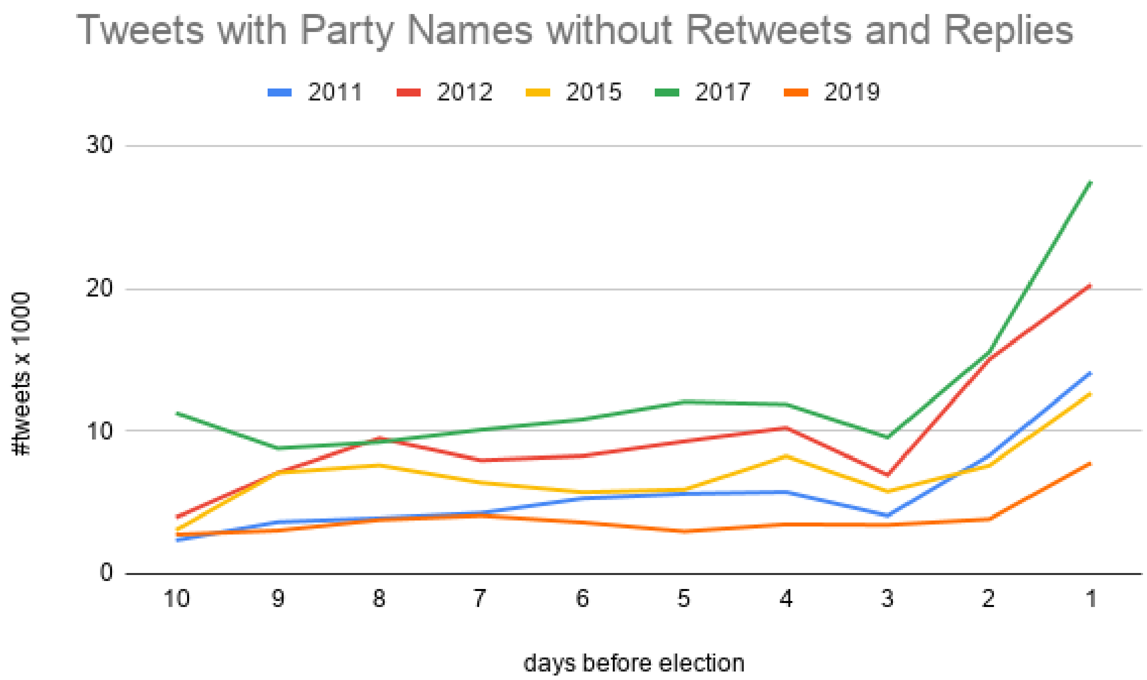

Figure 2: Number of tweets (excluding retweets and replies) with one or more party names in the ten days before the elections, for the five elections.

5. Discussion

Both a larger Pearson correlation (in four of the five elections) and a smaller absolute error (in all elections)

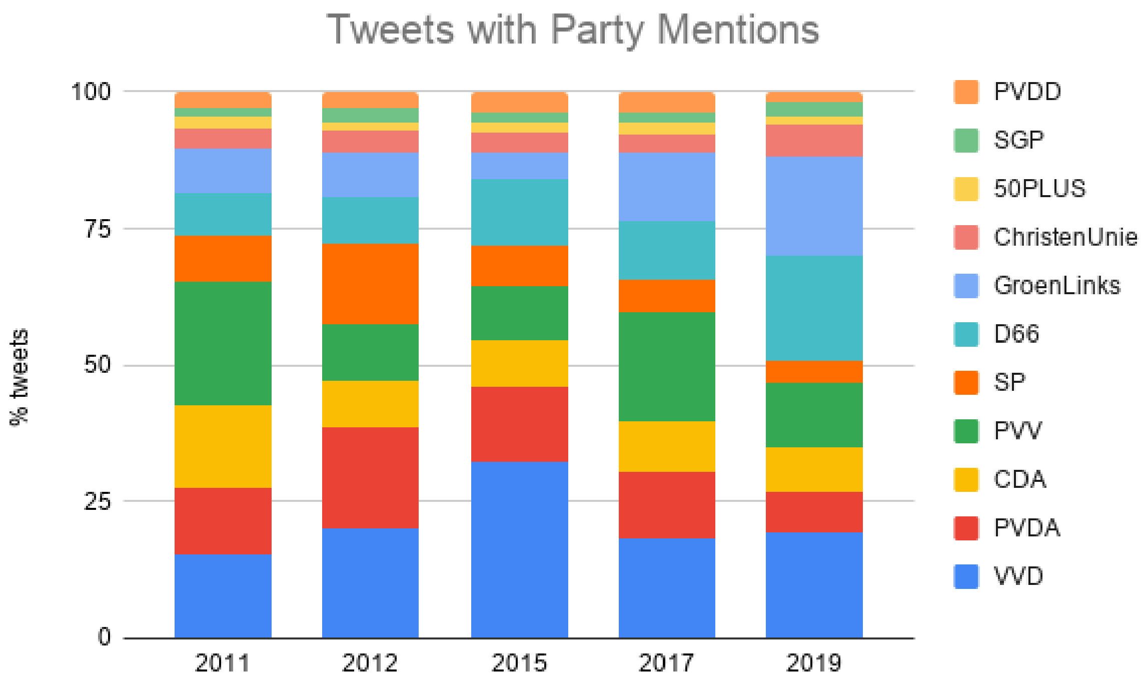

Figure 3: Percentages that indicate how often a political party was mentioned in tweets, for five elections.

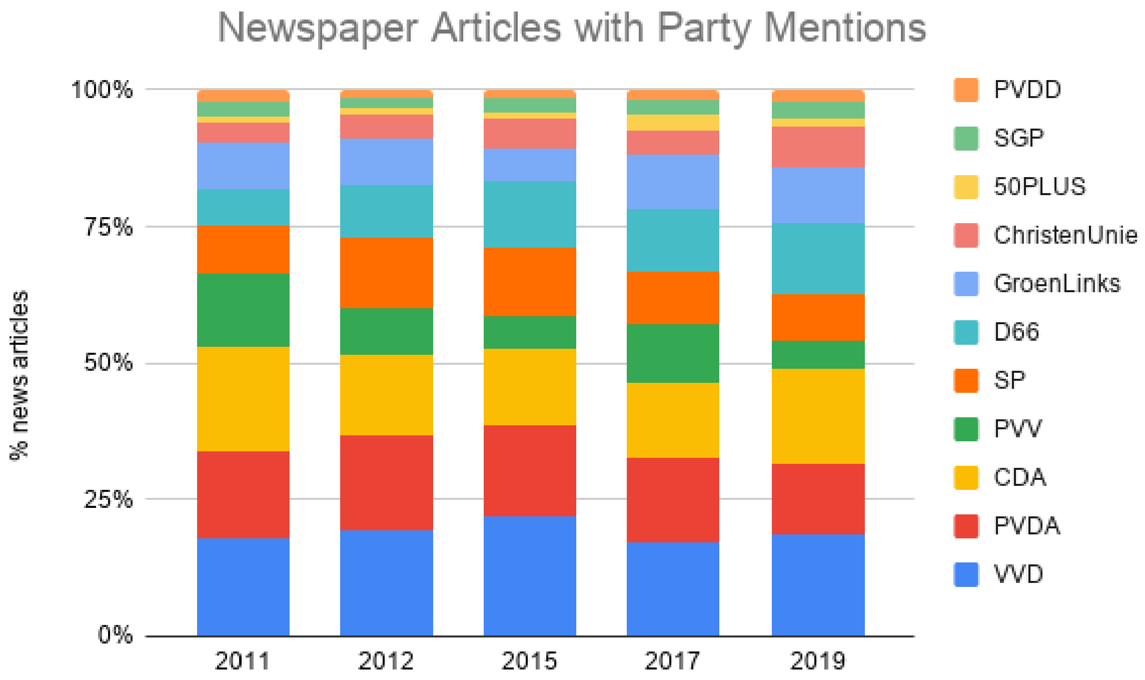

Figure 4: Percentages that indicate how often a political party was mentioned in newspaper articles, for five elections.

of the relation tweets—newspapers compared to the relation tweets—election results would confirm our assumption that party mentions in tweets are more influenced by the news than the political preferences of Twitter users, but the differences are overall very small.

Although it was not the focus of our research, we found that the correlation between party mentions in newspaper articles and the election results is the largest in four

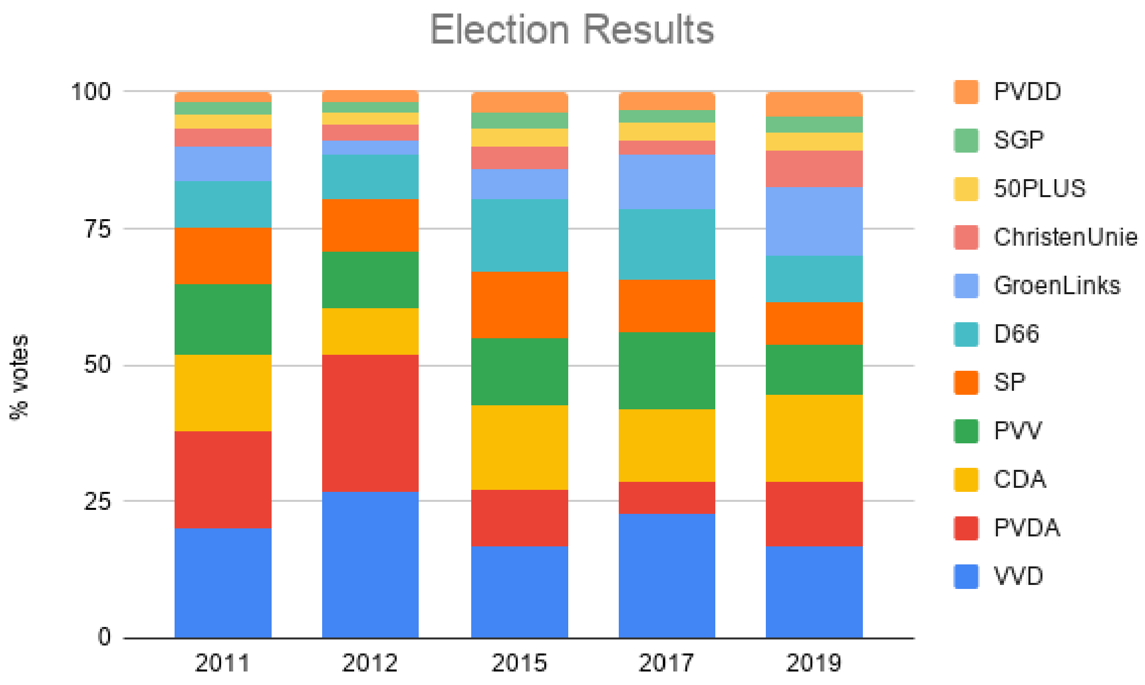

Figure 5: Percentages that indicate how many votes a political party got, for five elections.

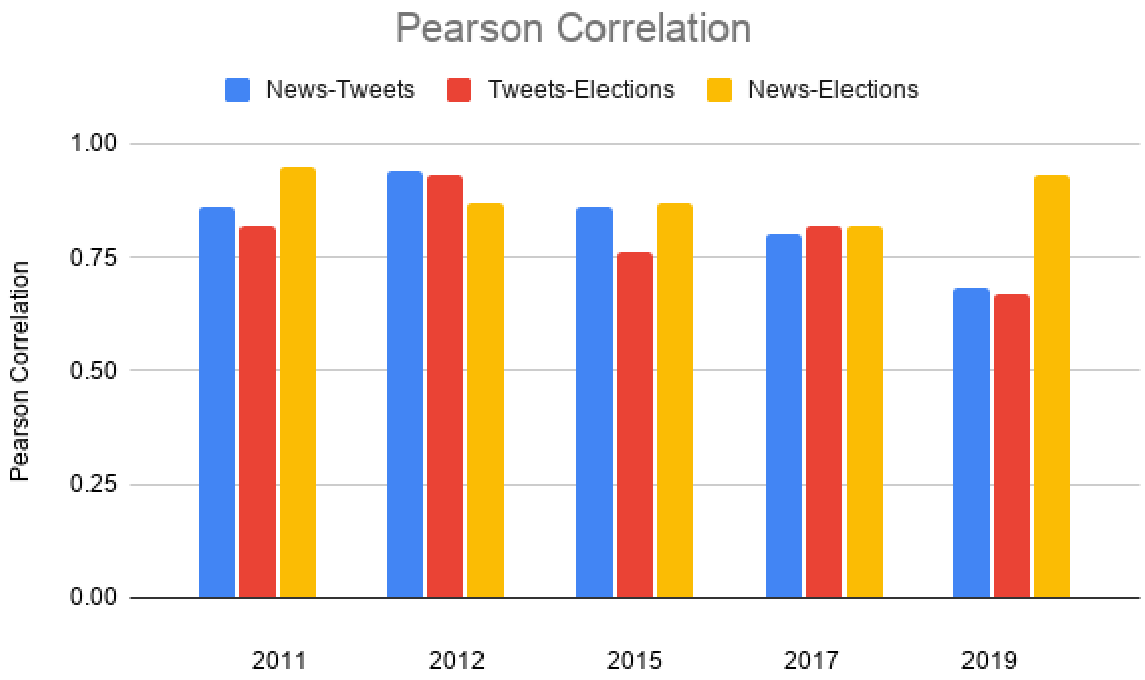

Figure 6: Pearson's correlation coefficients for the relation pairs between newspaper articles, tweets and election results for five elections.

of the five elections. Especially the absolute error is significantly lower in these four cases. The exception is in 2012 when two parties were in a duel to become the largest. It seems that newspapers reflect the political preferences in society better than tweets do. That would make them a better basic predictor of the election results. This is in accordance to what Barclay et al. conclude in their paper about the political bias of In-

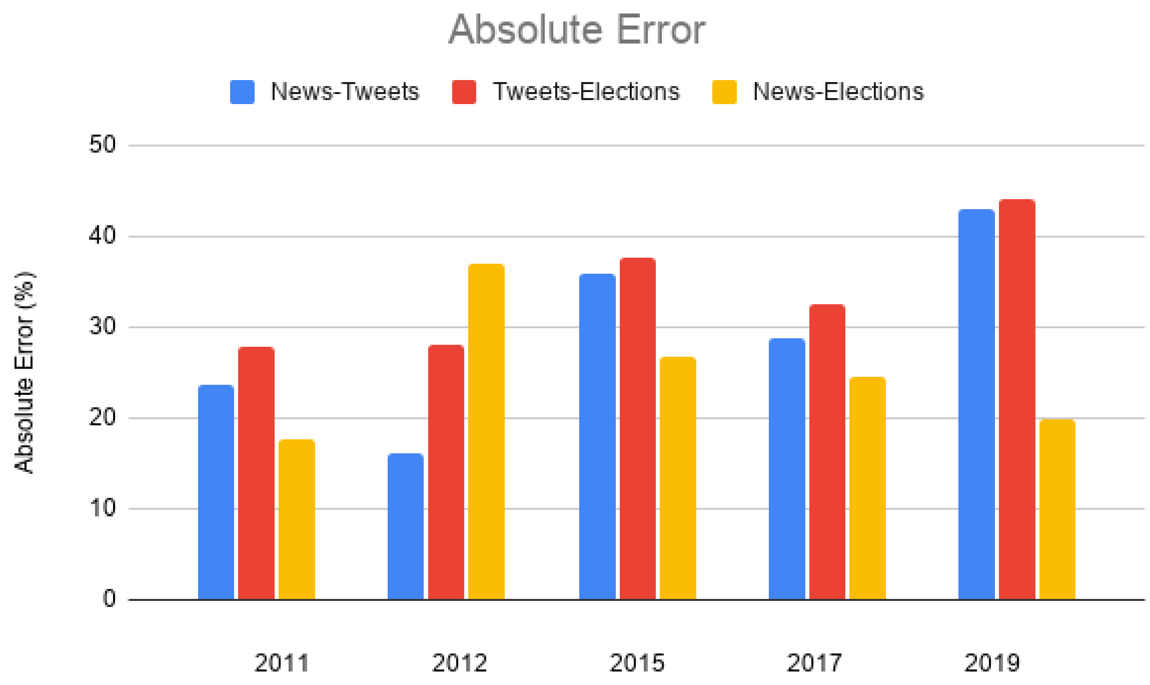

Figure 7: Absolute error for the relation pairs between newspaper articles, tweets and election results for five elections.

dian English newspapers in the 2014 elections in India: ”This overall Press bias was observed to have a strong and positive correlation with the vote count, supporting the strong effects paradigm.” (Barclay et al., 2015).

We were investigating whether news has a bigger impact on which political party people tweet about than their intention to vote for that party. The first step to find out was to look at the correlations and absolute errors. We realise that these measurements in itself do not tell anything about the influence of one data stream on the other. To be able to have more insight in the direction of influence we would need to take a closer look to the chronological order in which the parties are mentioned in different media and in what context they are mentioned.

When we take a closer look at the mentions of the individual parties in the newspapers and tweets in graphs 3, 4, 5, we see that PVV (an anti-islam party) is consistently underrepresented in the newspapers while CDA (Christian democrats) is over-represented. PVV has for a long time been a party that people typically will not say they will vote for. CDA on the other hand is a party in the middle of the political spectrum that has been the largest party for large periods in the previous century that is declining in support since a few decades, but still talked about in the newspapers. The same goes for PvdA (social democrats), but the overrepresentation in the newspapers is smaller than that of CDA.

We restricted ourselves to newspapers because of feasibility reasons. It would be interesting to study the difference with respect to party mentions in other media, such as radio, television and news websites. Unfortunately we do not have access to searchable resources that contain the transcriptions of radio and television broadcasts and news websites are often behind a paywall or very difficult if not impossible to search. We conjecture that they are likely to be correlated strongly to our newspaper measurements.

For our comparisons, we used raw counts of political party mentions in tweets and newspapers. This is very crude. Of course it would be best to normalise for all kinds of demographic variations of the tweeters, as that could help improving the results as the demographics of Twitter users are different (and changing over time) from the general voting populace; being able to correct for that would strengthen the assumption that when Twitter users mention a party, they often express their political preference for that party. However, as far as such a correction is technically possible, we have indications it does not offer an improvement (Sanders et al., 2016). Also taking context and sentiment into account does not appear to improve the preciseness of the counts, as we concluded in (Sanders and van den Bosch, 2020).

6. Conclusion

The goal of our research was to investigate the hypothesis that mentions of political party names in tweets are more influenced by what people read from the media (i.e. the news) than by what they (intend to) vote for. A first step in this investigation is to look at the correlation of party mentions in tweets and newspaper articles and the election results. Pearson correlation between party mentions in tweets and newspaper articles is larger than that of party mentions in tweets and the election results in four of the five elections and the absolute errors is smaller in all elections, which indicates that our hypothesis is confirmed. However, the differences are overall too small to be able to draw definitive conclusions.

It appeared that the correlation between the party mentions in the news and the election results are significantly higher and the absolute error significantly lower in four of the five cases. This leads us to conclude that newspaper articles might be a better predictor of the election outcome than tweets.

7. Bibliographical References

Barclay, F. P., Venkat, A., and Pichandy, C. (2015). Media effect: correlation between press trends and election results. *Media Asia*, 42(3-4):192–208.

Batra, P. K., Saxena, A., Goel, C., et al. (2020). Election Result Prediction Using Twitter Sentiments Analysis. In *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 182–185. IEEE.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 37–40. Springer.

Druckman, J. N. (2005). Media matter: How newspapers and television news cover campaigns and influence voters. *Political communication*, 22(4):463–481.

Gayo-Avello, D. (2012). No, you cannot predict elections with Twitter. *IEEE Internet Computing*, 16(6):91–94.

Knapp, J. A. (2018). Nexis Uni. *The Charleston Advisor*, 19(3):31–34.

Liu, R., Yao, X., Guo, C., and Wei, X. (2020). Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach. *Annals of GIS*, pages 1–14.

Murthy, D. and Petto, L. R. (2015). Comparing print coverage and tweets in elections: A case study of the 2011–2012 US Republican primaries. *Social science computer review*, 33(3):298–314.

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

Nugroho, D. K. (2021). US presidential election 2020 prediction based on Twitter data using lexicon-based sentiment analysis. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 136–141. IEEE.

Rao, D. D. R., Usha, S., Krishna, S., Ramya, M. S., Charan, G., and Jeevan, U. (2020). Result Prediction for Political Parties Using Twitter Sentiment Analysis. *International Journal of Computer Engineering and Technology*, 11(4).

Sanders, E. and Van den Bosch, A. (2013). Relating Political Party Mentions on Twitter with Polls and Election Results. In *Proceedings of DIR-2013*, pages 68–71.

Sanders, E. and van den Bosch, A. (2019). A Longitudinal Study on Twitter-Based Forecasting of Five Dutch National Elections. In *International Conference on Social Informatics*, pages 128–142. Springer.

Sanders, E. and van den Bosch, A. (2020). Optimising Twitter-based Political Election Prediction with Relevance and Sentiment Filters. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6158–6165.

Sanders, E., de Gier, M., and van den Bosch, A. (2016). Using Demographics in Predicting Election Results with Twitter. In *International Conference on Social Informatics*, pages 259–268. Springer.

Su, Y. and Borah, P. (2019). Who is the agenda setter? Examining the intermedia agenda-setting effect between Twitter and newspapers. *Journal of Information Technology & Politics*, 16(3):236–249.

Tjong Kim Sang, E. and Van den Bosch, A. (2013). Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal*, 3:121–134, 12/2013.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*, 10:178–185.