Tweeting opinions
- How does Twitter data stack up against the polls and betting odds?
Abstract

With the rise of social media, people have gained a platform to express opinions and discuss current subjects with others. This thesis investigates whether a simple sentiment analysis — determining how positive a tweet about a given party is — can be used to predict the results of the Swedish general election and compares the results to betting odds and opinion polls. The results show that while the idea is an interesting one, and sometimes the data can point in the right direction, it is by far a reliable source to predict election outcomes.

Keywords: sentiment analysis, twitter, social media, election results, betting odds, swedish general election 2018
Contents

1 Introduction 4
   1.1 Background 5
   1.2 Related work 6
   1.3 Problem formulation 7
   1.4 Motivation 7
   1.5 Objectives 7
   1.6 Scope/Limitation 8
   1.7 Target group 9
   1.8 Outline 9

2 Method 10
   2.1 Sentiment Analysis Library 10
   2.2 Data Gathering 10
   2.3 Reliability and Validity 10
   2.4 Ethical Considerations 12

3 Implementation 13
   3.1 Implemented Scripts 13
      3.1.1 TwitterSearch 13
      3.1.2 CleanData 14
      3.1.3 SentimentAnalysis 15
      3.1.4 PollCrawler 16
      3.1.5 ScoreCalculator 17

4 Results 18
   4.1 Data points 18
      4.1.1 Total number of tweets 18
      4.1.2 Sentiment analysis average score 18
      4.1.3 Comparative sentiment average 18
      4.1.4 Percentage positive 18
      4.1.5 Betsson and Unibet odds 18
      4.1.6 Polls 18
   4.2 Result Presentation 19
      4.2.1 Total number of tweets 19
      4.2.2 Sentiment analysis average score per tweet 19
      4.2.3 Comparative average score 20
4.2.4 Percent of all tweets that were positive 20
4.2.5 Betsson odds 21
4.2.6 Unibet odds 21
4.2.7 Opinion polls 22

5 Analysis 23

6 Discussion 25

7 Conclusion 26
   7.1 Future work 26

References 27
1 Introduction

Social media can be a powerful platform to express opinions and discuss current subjects with other users. This could lead to the conclusion that the data generated by thousands of users can be used to gauge the public sentiment about current events, and even predict the outcome of elections.

1.1 Background

Sentiment analysis belongs to the field of Natural Language Processing (NLP), which allows a machine to understand the natural language used by humans to communicate with each other. This technique has received increasing interest over recent years, a big reason is an increased activity on social media — where users share their opinions publicly and engage in debates and discussions. Since the creation of Twitter in March 2006 — when the founder, Jack Dorsey, posted the now famous first tweet: “just setting up my twtr” [1] — the public have had an open channel in which to express their sentiment about a vast number of subjects. Due to the limited number of characters that are allowed in a single tweet Finn Årup Nielsen developed a new sentiment lexicon: AFINN. His sentiment lexicon is updated to better score the valence — or emotional value — of words used on microblogging platforms, such as Twitter [2].

Advancements in machine learning have enabled social networks to use algorithms to classify and determine the sentiment of their users, and thereby recommend content they are likely to be interested in. The use of these algorithms has also put social networks under fire. Critics argue that the algorithms place users in so-called “filter bubbles” in which users are shown more of what they are expected to like, which in turn is based on what they already like. People are trapped in a feedback loop, or an "echo chamber", where it seems like everyone else is of the same opinion as they are [3].

It might, therefore, seem certain that social media is not up to the task of predicting public sentiment. However, that might only be true from a single user’s perspective. As a social media network is a larger structure of many of these filter bubbles, all populated by their millions of users, the data generated could be able to give a much clearer picture of what goes on. At least if the unfiltered data can be analysed.

Many social media platforms provide APIs to developers and researchers for this purpose: to enable unfiltered access to the mountains of
data stored on their servers. However, social media platforms may attract a certain demographic and not be representative of the broader society.

While sentiment analysis has only, relatively, recently begun to be applied to predicting elections – betting odds have a much longer history as a resource to predict the result of an election, with studies analysing data that goes back to the 1880’s [4]. However, the accuracy isn’t always high, as pointed out by Wall et al [5], and the question of demographics remain: even though anyone over 18 is able to place a bet, do all demographics of the constituency have an even ratio of gamblers? Some researchers have still suggested it as a metric worth noting [6].

1.2 Related work

The topic of sentiment analysis and election prediction with Twitter data is not new. In fact, several attempts have been made prior to this one: on the American presidential election [7, 8]; the German federal election [9]; the French presidential election [10] and the UK general election [11] to name a few. The methods of these prior studies have been both the use of the simple volume metric — i.e. the candidate with most mentions is seen as most likely to win — as well as sentiment analysis.

The results have varied from reported successes in predicting elections [9, 10, 12] to critically claiming such predictions were no better than chance [13] and proving that claim by failing to predict the US congressional election of 2010 using the same methods. Tunggawan and Soelistio also failed to predict the nominees of both the Democratic and Republican sides in the 2016 presidential election [7]. Their Naïve-Bayesian model erroneously predicted Ted Cruz as the Republican nominee and Bernie Sanders as the Democratic nominee.

There seems to be a lack of studies that compare the metrics of social media with betting odds data. However, there have been studies on how well betting markets can predict elections. For example, Rhode and Strumpf point out that betting markets were quite successful in predicting elections before the use of scientific polling [14]. This result is further backed up by Erikson and Wlezien who found that election markets were better predictors of the result before polling became available than after [4].
1.3 Problem formulation

Politicians and their constituents use opinion polls to gauge the political landscape of the electorate. The problem with opinion polls is that they are not necessarily indicative of the final election result. Further, it takes time to gather the data for the opinion polls, and interviewees may decline to answer the poll in the first place. By using sentiment lexicons and readily available data from social media and betting sites, another type of measurement can be used in a shorter time and complement the opinion polls.

1.4 Motivation

There is an increasing reliance on algorithms and as society and industry put an increasing amount of faith in these systems there is a risk of hubris arising. Researchers have been using Twitter and other social media platforms to predict earthquakes [15] and the spread of the flu [16]. However, while earthquakes and the flu affect people regardless of demographics, the same might not be said about political opinions and social media usage. This hasn’t prevented researchers from using social media to predict election results, to varying degrees of success [7, 8, 9].

A look at the Swedish election is interesting due to the country’s eight parliamentary parties, potentially distributing the demographics over the political spectrum. Another aspect of interest is the limited geographical influence of the election result. Swedish politics are rarely discussed outside its borders and the limited spread of the Swedish language makes for higher fidelity in data gathering as any posts in other languages can be discarded.

Researchers have also suggested betting sites as an accurate predictor of election polls [4], another set of data readily available for use in analysis. It is therefore interesting to investigate if this metric is more accurate than social media.

1.5 Objectives

<table>
<thead>
<tr>
<th>O1</th>
<th>Gather Twitter data</th>
</tr>
</thead>
<tbody>
<tr>
<td>O2</td>
<td>Gather betting odds data</td>
</tr>
<tr>
<td>O3</td>
<td>Categorize Twitter data according to sentiment (Positive or Negative)</td>
</tr>
<tr>
<td>O4</td>
<td>Analyse categorized data and use it to predict poll standings of the three largest parties in the Swedish parliament</td>
</tr>
<tr>
<td>O5</td>
<td>Compare the result of categorized to polls and betting odds</td>
</tr>
<tr>
<td>O6</td>
<td>Adjust algorithms if necessary</td>
</tr>
</tbody>
</table>
The result of the sentiment analysis is expected to give Sverigedemokraterna a high negative score due to the populist nature of the party. Further, many of the tweets mentioning Sverigedemokraterna are expected to be negative in nature due to disagreements with all other parties.

The betting odds, on the other hand, are correlated with the regular polls and current events which should give it a more balanced approach, especially due to the polls direct effect on the markets as suggested by Erikson and Wlezien [4].

1.6 Scope/Limitation

The research in this project is limited to the three currently largest parties in the Swedish parliament. To improve the accuracy there is a need for large amounts of data, which will be easier to satisfy for the larger parties. Further this project does not aim to predict a percentage but rather who will be the winner, runner-up and third position. Thus, the leading party of each block is selected: Moderaterna, Socialdemokraterna and Sverigedemokraterna. A problem with this approach is that it doesn't consider the other parties of each respective block. However, due to the 4% barrier to entry into the Swedish parliament the data needed to accurately predict such a result is beyond the scope of this project.

This project only gathers data from Twitter, this is due to the nature of Twitter posts which are limited to 280 characters, forcing users to be more concise. The downside is that the demographics are limited to that of Twitter’s Swedish user base. However, image or video-based social media platforms would increase the size of the scope far beyond what is possible in the allotted time frame for this project. Further, Twitter’s heavy use of hashtags and mentions makes it easy to query their API for a limited topic to procure a focused dataset.

Since it is beyond the scope of this project to make a new implementation of an NLP classifier, pre-existing libraries will be used for all sentiment analysis classification tasks.

Another limitation is due to the rate limitation of 250 requests per month on the Twitter Search API – where each request returns 100 tweets – which means this project can’t use the pure frequency metric of prior studies. It will instead focus on the sentiment analysis and betting odds data.
1.7 Target group

The target group of this study are primarily researchers that are interested in the comparison between social media data, betting odds and opinion polls to further build on this research. The secondary target group is politicians and professionals within the field of politics who may gain a new metric tool with which to predict the public opinion and react to it faster.

1.8 Outline

In the following chapters, the project is described in more detail. In chapter 2. Method, the approach to solving the problem of this thesis is outlined. The chapter details which sentiment analysis library that is used and how data gathering is done. It also contains a lengthy discussion about the reliability and validity of the project, as well as a short discussion about the ethical considerations.

Chapter 3 describes the implementation details and illustrates the flow of the scripts using sequence diagrams. Since the application relies on Node.JS scripts, class diagrams are not of as much help to describe the application and are therefore not included.

Chapter 4 presents the results of the study using graphs and tables, while also describing the details of the results, which are then analysed in Chapter 5. Analysis. Chapter 6 discusses the findings of the thesis and the final chapter, Chapter 7, attempts to draw a conclusion and recommend future work that can be done to expand on this study.
2 Method

This project will attempt to solve the problem with the Controlled Experiment approach, to collect quantitative data from Twitter via their API. This method is best suited for the project since it will use a large quantity of data gathered programmatically from Twitter. This data will then be analysed in a controlled environment to determine the ranking of the parties according to the number of votes. The independent variables are the tweets and the betting odds, while the dependent variable is the result of the analysis which is then compared to the opinion polls.

With the election day taking place far beyond the end date of this project, the polls are the best metrics available to determine the validity of the results.

2.1 Sentiment Analysis Library

The sentiment analysis will be performed with the sentiment-swedish [17] JavaScript library. The library translates the Swedish word to English, then uses the AFINN list of English words, developed by Finn Årup Nielsen [2], that have been rated with an integer between negative five and positive five.

2.2 Data Gathering

All data is stored in JSON format for later analysis. Twitter data is gathered using the Twitter Search API. The betting odds data is gathered manually in connection with running the Twitter Search script, this is due to the technologies on the two selected betting odds sites. They are both built using front-end frameworks with pure client-side rendering, meaning a crawler doesn’t get access to the same data a human sees in the browser.

2.3 Reliability and Validity

The biggest reliability and validity problems for the project are related to the classification and sentiment analysis of the data. The algorithms can’t promise 100% accuracy, however that problem should remain the same in repeat studies on the same data. Further, there are far fewer tools to perform sentiment analysis on Swedish text and even fewer that don't first translate the text into English. A problem with automatically translated text is the fidelity to the source – translated texts are far too often incorrect, though this
has been somewhat countered by the application of machine learning to translation in Google Translate services. However, even the original text from tweets present a problem for the sentiment classifiers with misspelled words and the use of abbreviations due to the maximum character restrictions on tweets. On the other hand, misspelled words would still be incorrectly labelled by the sentiment analysis.

A library that uses a Swedish sentiment lexicon would be preferred. Researchers at the University of Gothenburg have created a Swedish sentiment lexicon [18], however, there seems to be a lack of libraries that implement it.

The reliability and validity of the dataset are increased due to the use of the Premium search of the Twitter API rather than just the standard, or even crawling the regular website, as the latter two only provide a sample of the tweets whereas the former provides access to search all tweets since 2006. However, as only data from Twitter is collected and no other social media sources are used, the data might be skewed by a limited demographics. On the other hand, that is one of the points of interest for this study – to compare the sentiment analysis of the Twitter data to the results of national polls.

The sentiment analysis might also be skewed if – for example – a supporter of Sverigedemokraterna tweets negatively about immigration and then proceed to tag Sverigedemokraterna in the tweet. This could give a negative score to Sverigedemokraterna even though it should give them a positive score. Hopefully, this will be counteracted by the tagging of other parties in the same tweets, or similar cases with the other parties as well. To increase the validity of the result a context-aware sentiment analysis could also be performed on the data and compare that result to the sentiment of each party before giving a score. However, that is beyond the scope of this thesis project and perhaps a suggestion for further studies.

The betting odds data will be gathered from Unibet [19] and Betsson [20] and an average score will be given to each party. Since the betting odds are not fixed but rather actively change due to external factors the exact day the data was collected will be noted as well as the respective odds from each of the betting sites.
2.4 Ethical Considerations

Due to the gathering of Twitter data and sentiment analysis of this data, the project has Ethical considerations to make. Specifically, that Twitter usernames or full names can be mapped to a certain political opinion. To protect the Twitter users’ privacy no personally identifiable information will be stored or presented during the project.
3 Implementation

The software in this project consists of several scripts with specialised functionality. The scripts are written in the JavaScript programming language and use the sentiment-swedish NPM package for sentiment analysis. All data is stored in JSON format. This chapter describes each of the scripts in this implementation and presents the flow with sequence diagrams.

3.1 Implemented Scripts

The following is a description of the scripts that have been developed for the project. All scripts are executed with NPM using NodeJS from the terminal.

3.1.1 TwitterSearch

The TwitterSearch script calls the Twitter API with the selected queries: “sdriks”, “socialdemokrat” and “nya_moderaterna” to get all mentions of the three selected parties. Due to the limitation in the number of queries that can be made each month (250) the first implementation allowed for a maximum
of 75 calls to the API per party. This was intended to be done once per month during the project. However, since it was discovered that this only gave results for a short time period (due to the high volume of tweets) the limit was lowered, and calls were instead made twice per month to get a better temporal spread in the gathered tweets.

Further due to the rate limit of the Twitter API the script pauses for 60 seconds after every 29th request to the API. The script also saves the gathered data in JSON format after stripping away any non-essential data (i.e. any data that is not the time-stamp of the tweet, the text and the mentioned party) by using the CleanData script (explained in the next section) developed for that purpose.

However, the second execution of the script showed that pausing the script for 60 seconds after 29 requests introduced a bug as it allowed the script to reach the rate limit of requests per minute. To counter that the limit was lowered to 21 requests every 65 seconds to allow for a safe buffer.

3.1.2 CleanData

The CleanData script receives the response data from the Twitter API and sanitises it by removing any sensitive and non-essential information relating to the project. That means the only information that is kept is the time-stamp, the tweet’s text content and the party that was mentioned. The reason is to ensure the privacy of Twitter users as outlined in Chapter 2.4 Ethical Considerations.
3.1.3 SentimentAnalysis

Figure 3.2: SentimentAnalysis Sequence Diagram shows the execution of the SentimentAnalysis script and reliance on external Node Module sentiment-swedish.

The SentimentAnalysis script uses the sentiment-swedish NPM package to perform the sentiment analysis on the tweets. The script reads the data from the saved files and scores each tweet with either a positive or negative integer depending on the outcome of the sentiment analysis. The sentiment analysis itself sums the words in the tweet by using a pre-existing sentiment lexicon where words are given a score based on their weighted positive or negative nature.

Once all tweets have been scored the results are saved in time-stamped files for each party.
3.1.4 PollCrawler

The PollCrawler script crawls the data used on val.digital [21]. The script uses two NPM libraries: request-promise and cheerio to request the HTML, process and mine it for the relevant data.

This script is run once per month during the project to get the polling results of that month. The average of the polls is then calculated and saved in a time-stamped file to compare with the results of the sentiment analysis and the betting odds data.
3.1.5 ScoreCalculator

The ScoreCalculator script reads the data files containing the results of the sentiment analysis, the average score of the polling results and calculates the average score of the betting odds data that is manually gathered twice a month during the project. This data is then compared to test the results of the sentiment analysis and the betting odds data compared to the results of the opinion polls. The result of the comparison is then displayed in the user’s terminal.
4 Results

The Twitter data was analysed 3 times during the project. In total 68,400 
tweets were analysed. Finally, when the project concluded the average for all 
data was calculated. The analysis provided the results presented in this 
chapter.

4.1 Data points

The results of each run are presented with the following data points:

4.1.1 Total number of tweets

Presents the total number of tweets gathered for each party during each 
execution as well as the total during the entire project.

4.1.2 Sentiment analysis average score

The average score is calculated from the total score of the sentiment analysis 
divided by the total number of tweets. The sentiment analysis score is 
calculated by adding the scores for all the words in each tweet. The resulting 
score is either a positive or negative integer. Higher is better.

4.1.3 Comparative sentiment average

The average comparative score, calculated by dividing the total comparative 
score by the total number of tweets. The comparative score is calculated by 
dividing the total score for each tweet with the total number of words in that 
tweet, as some words are neutral and give a score of 0. Higher is better.

4.1.4 Percentage positive

The percent of all tweets that received a positive score for each party. Higher 
is better

4.1.5 Betsson and Unibet odds

The odds for which party will get the highest percentage of votes in the 2018 
election, gathered from the betting sites Betsson and Unibet. Lower is better.

4.1.6 Polls

The average score from opinion polls scraped at val.digital during that month.
4.2 Result Presentation

4.2.1 Total number of tweets

Figure 4.1: Total number of tweets gathered from Twitter API

4.2.2 Sentiment analysis average score per tweet

Figure 4.2: Average score of Twitter sentiment analysis March-May 2018. Higher is better.
4.2.3 Comparative average score

Figure 4.3: Average comparative score of Twitter sentiment analysis March-May 2018. Higher is better.

4.2.4 Percent of all tweets that were positive

Figure 4.4: Percent of all tweets that were positive March-May 2018. Higher is better.
4.2.5 Betsson odds

Figure 4.5: Betting odds for which party is most likely to get the most number of votes in the election. From Betsson March-May 2018. Lower is better.

4.2.6 Unibet odds

Figure 4.6: Betting odds for which party is most likely to get the most number of votes in the election. From Unibet March-May 2018. Lower is better.
4.2.7 Opinion polls

![Opinion polls March-May 2018](image)

Figure 4.7: Opinion polls March-May 2018
5 Analysis

As can be seen in the previous chapter, the sentiment analysis differs from one month to the next. In the first execution, the parties ranked as could be expected according to the polls. However, in the second execution, Sverigedemokraterna had the highest percentage of positive tweets as well as the highest average sentiment score. The move forward for the party can also be seen in the betting odds at Betsson and Unibet, which both rank Sverigedemokraterna higher than Moderaterna. However, according to the polls, the party actually lost ground that month. Moderaterna were the only party, out of those analysed, to gain a higher percentage in the polls from the month of March to the month of April. However, their ranking in the sentiment analysis dropped and the betting odds remained the same.

It is interesting to note that Sverigedemokraterna received a much higher average sentiment score in the second execution than in the first, just to drop to the lowest score in the third execution.

The second execution of the analysis also gathered 2,000 tweets fewer about Socialdemokraterna than the other two parties. The reason for this was due to a software bug, as the calls to the Twitter API are done asynchronously for all the parties and a rate limit was set to 30 requests per minute and 10 requests per second, the software hadn’t left enough buffer in the interval method which paused the execution for exactly 60 seconds. It also didn’t take the limit of 10 requests per second into account. This was rectified in the third and final execution of the script.

While the length of this study is not sufficient to draw any real conclusions from the data, an interesting pattern has nevertheless emerged. The Twitter analysis showed an increase in Twitter users’ positive sentiment towards Sverigedemokraterna the month before they showed an increase in the polls. At the same time, both Socialdemokraterna and Moderaterna showed a decrease in the polls the month after they received a lower score in the sentiment analysis. Further, the betting odds have shown a steady incline for Sverigedemokraterna while the other two parties saw a decline (positive) or no change at all. It would be interesting to see if this pattern would continue, or if it was just a coincidence.

The average across all the executions seems to correlate better with the polls, so it’s possible that the more stable long term gathering of data for analysis would give a more reliable score. However, as noted before, the study is far too short to draw any real conclusions about that.

There is one thing worth keeping in mind when analysing the results of this study, namely the problem mentioned in section 2.3 Reliability and Validity, where the overall negative attitude of Sverigedemokraterna towards immigration is discussed. This could account for a large part of the negative
score the party received in the first Twitter analysis. It would be interesting to see if *Sverigedemokraterna* still placed third if more parties were analysed using the same method, or if the negative sentiment towards immigration (which is one of the party’s focus points) would drag them down compared to the smaller parties.
6 Discussion

The findings of this study show that sentiment analysis on Twitter data may differ wildly and, depending on which data is investigated, may be ambiguous and point to different results. As seen in the first execution *Moderaterna* gained a higher average score than *Socialdemokraterna* but the latter had a higher percentage of positive tweets. The percent of all tweets about a party that were positive seem to be most indicative of the results in the polls. However, this study would need to include more parties to test the reliability as noted in the previous chapter.

The betting odds also mostly placed *Sverigedemokraterna* higher than *Moderaterna*. The decreasing odds of *Sverigedemokraterna* was also reflected in both the Twitter analysis and the opinion polls. The change in the other two parties is not significant enough to draw any real conclusions.

However, the sentiment on Twitter differs a lot more from one month to the next than the opinion polls, or even the betting odds. This is expected due to the nature of microblogging platforms like Twitter, which reflect users’ reactions to current events. Opinion polls, however, are gathered over time and allow participants a longer time to reflect before responding.

The analysis could have benefited from using a context-aware approach, where the content is analysed in a larger context. For example, as was mentioned in Chapter 2.3, if a user tweets negatively about immigration and mention *Sverigedemokraterna* the score will be negative even though it should probably give a positive context score for the party. Twitter threads are another context that could be taken into consideration, as a user might be responding to someone else but mention their party in the tweet.

Another problem is the limitation to only the top three largest single parties without looking at their coalition partners and the other major parties. This study was, however, limited in the number of requests that could be done to the API, yet another limitation to this study.

These findings seem to correlate to the studies that were successful at predicting the election results [9, 10, 12], at least based on the short few executions that were run during the short time frame of the project. However, the fluctuating results also show that the method is unpredictable as it differs far more than the polls and the short time frame of this study could mean that what seems to be a pattern is only a coincidence.
7 Conclusion

The results of this study lead to the conclusion that while using sentiment analysis on Twitter data to predict an election could be possible, it becomes more difficult in a multiple-party system, compared to a simpler two-party system. The sentiment analysis also needs to be context-aware to draw any real conclusions from the analysis, as well as extend further in time. Further, the other major parties of the Swedish system should be included to gain a more reliable result.

The main relevance of this study is to serve as a foundation for further, large scale, studies. This study could be applied to any political election, at least if there are betting odds available for analysis.

7.1 Future work

As mentioned in the above chapter, the scale of this study was too small to conclusively answer the problem of using Twitter data to predict election results. With more resources, the study should expand to include all the major parties of the Swedish political system over a much longer time span, as well as a context-aware sentiment analysis model using machine learning technologies to score the sentiment of each tweet as it relates to the mentioned parties, or within a thread of tweets. The results should also be compared with the final election results and not only the opinion polls to get a far more accurate result.
References


