Predicting the Outcome of the 2017 UK General Election

By Rory Birtwistle, Rachel Cross, Wasim Malek, Ivan Tchebykine
and Monika Zajac

7th June 2017

Contents

1   Introduction

On 18th April 2017, Theresa May announced a snap general election, due to take place on 8th June 2017. This report outlines our attempt to predict the outcome of the election by analysing the sentiment of tweets along with other data sources such as voting polls, betting odds and Google Trends.

2   Possible Approaches

Two high-level approaches were identified:

- A: Identify how the UK would vote at a national level.

- B: Identify constituencies representative of the rest of the UK, then identify how those constituencies would vote.

Approach B appeared to be more complex, yet more sophisticated, especially when constituencies were identified that had successfully predicted the outcome of the last ten or more elections consecutively and only had a marginal majority in the 2015 general election. However, less than 1% of tweets contained reliable and accurate data on location.

An alternative was to check the sentiment of tweets mentioning the MPs in each local constituency. Unfortunately, only 1,500 tweets were collected in 48 hours using this approach and the quality of those tweets was poor; many tweets referred to people who shared the same name of the local MP. In addition, local MPs often visit other constituencies to help other MPs with their election campaigns. Therefore, tweets mentioning an MPs name were often sent from people in other constituencies, hence skewing the results. Approach B was therefore rejected, and approach A was chosen.

3   Chosen Approach

Research into previous election predictions using Twitter sentiment highlighted one recurring problem. The Twitter population is inherently biased, with the vast majority of its users consisting of young people that hold left-leaning views.

In an attempt to balance this bias, other data sources were used. Google Trends had been used to successfully predict previous elections. This was then supplemented with data from polls and odds from betting companies. Articles from major news corporations were not used, as each news corporation promotes its own political preferences, and naturally attracts readers who share the same political preferences. Therefore, we felt this data source lacked value. However, a log was kept of major news headlines to see how they affect sentiment on twitter, betting odds, polls and Google Trends results.

The scope of the analysis was limited to the UK, meaning tweets and Google searches from other locations were not considered. Furthermore, since poll data was limited to six parties (Conservative, Labour, SNP, Liberal Democrats, Green and UKIP), other political parties were not considered. Finally, users on the Twitter have the option to 'retweet'; research highlights that the action of retweeting indicates interest, trust and agreement in a tweet, along with trust in the originator. Therefore, retweets were included in the analysis.

3.1 Twitter Sentiment

Tweets were collected from the Twitter Streaming API and sentiment analysis was applied to the tweets producing values for polarity and subjectivity. Polarity values range from -1 to +1, and indicate the level of positivity or negativity in the tweet. Meanwhile subjectivity values range from 0 to 1 and indicate whether the tweet was a fact, opinion, or mixture of both.

The tweets collected were limited to the track terms in table 1.

Table 1: Twitter Track Terms

Party	Track Terms
Conservatives	'conservatives', 'theresa may', 'conservative party', 'theresamay', 'tories', 'tory', 'conservativeparty', 'theresa_may','votetory', 'toriesout','VoteConservative'
Labour	'labour party', 'jeremy corbyn', 'jeremycorbyn', 'labour', 'labourparty', 'corbyn', '@uklabour', 'jc4pm', 'corbyn4pm', 'forthemany', 'strongandstable','Votelabour'
Scottish National Party	'nicola sturgeon', 'nicolasturgeon', 'scottishnationalparty', 'scottish national party', 'snp','VoteSNP'
Liberal Democrats	'tim farron', 'timfarron', 'liberal democrats', 'liberaldemocrats', 'libdems', 'libdem','Votelibdem','Votelibdems'
Green	'Caroline Lucas', 'CarolineLucas', 'Jonathan Bartley','JonathanBartley', 'Green Party', 'GreenParty', 'jon_bartley','VoteGreen2017','VoteGreen'
UK Independance Party	'Paul Nuttall', 'PaulNuttall', 'UK Independence Party', 'UKIndependenceParty', 'UKIP', 'paulnuttallukip','VoteUKIP'

The track terms were designed to avoid using obscure terms such as 'conservative', so that tweets not related to the election were not collected, for example:

```
"After a conservative start to the game, Harry Kane has burst into
life by bagging a hat-trick in the last five minutes at the Emirates!!"
```

Tweets upto two days after special events, such as the launch of party manifestos and TV debates were also collected using event-specific hashtags, such

as: #labourmanifesto for the launch of the Labour Manifesto.

3.2 Google Trends

Data was also collected from the Google Trends API. The data is returned in the form of an index score, which ranges between 0 and 100 and indicates how frequently a search term was used in the Google search engine. Data was collected for the search terms in table 2. As with Twitter sentiment, the search terms were designed to avoid using obscure terms.

Table 2: Google Trends Search Terms

Party	Search Terms
Conservatives	'conservatives', 'conservative party', 'tories', 'theresa may'
Labour	'labour party', 'jeremy corbyn'
Scottish National Party	'nicola sturgeon', 'scottish national party', 'snp'
Liberal Democrats	'tim farron', 'liberal democrats', 'lib dems', 'lib dem'
Green	'caroline lucas', 'jonathan bartley', 'green party'
UK Independance Party	'paul nuttall', 'paulnuttall', 'uk independence party', 'ukindependenceparty', 'ukip'

3.3 Voting Polls

Results from a number of polls were collected from organisations such as The Guardian, Telegraph, Evening Standard, Kantar Public and Business Insider. The polls question a sample of individuals, representative of the UK's population, about their voting intention at various points in time throughout the year, and are of special interest during the run up to the general election.

3.4 Betting Odds

Betting companies publish odds/probabilities of each party winning the election. Betting companies need to publish odds that would not result in them paying out excessively to individuals, yet ensure the odds are attractive enough for individuals to place bets. Therefore, they assign significant level of resources to ensure the odds/probabilities are appropriate at any point in time. They are, therefore, an important data source.

3.5 Prediction

The prediction for the election can be broken down into two steps.

Firstly, the trend over time of betting odds, Google Trends, Twitter sentiment and polls needs to be extrapolated until the day of the election.

Secondly, the weightings for each extrapolation needs to be calculated and applied so that one score is available for each party. The party with the highest score will be our prediction for the election.

4 Visualisations

5 Prediction

6 Summary