General Election Prediction

s1529180 Kuanli Duan
s1618756 Yiwei Sun
s1677968 Wenjing Chen

With the development of informatics technology, the scale of the industry application system is expanding rapidly. The data produced by these systems are explosively growing every day. Therefore, effective big data processing techniques or methods are becoming more and more important. Particular, the big data prediction technology is a hot research direction at present. The advantage of big data prediction is that it transforms a very difficult prediction problem into a relatively simple description. It can be used to help government, enterprises, and organizations to make strategic decisions. In this report, we analyzed 3 months BBC transcript data to find the hot news topic. Then we decided to choose the 2017 General Election topic to attempt to predict the result. Although the 2017 General Election had taken place on 8th June 2017, the prediction method is not outdated.

BACKGROUND

The United Kingdom general election of 2017 took place on Thursday 8 June. It resulted in a situation that no party won an overall majority. The Conservative Party won the largest number of seats and votes, taking 317 seats and 42.3% of the vote. The Labour Party won 262 seats, and 40.0% of the vote. The Liberal Democrats won 12 seats, got 4 seats, and 7.4% of the vote.

The Conservatives remained the biggest party, but lost their overall majority as voters returned a hung parliament. Mrs May is seeking to form a new government backed by her new Northern Irish allies, the Democratic Unionist party.

DATA PROCESSEING

1. At first stage, we used Web Crawler technology to grab the BBC news transcript online Json data which includes sentences, words, speech rate and transcript accuracy. There is more than 2G size for one-year Json data. We cannot spend too much time to processing all of these data. Finally, We decided to choose 3 months data in 2017 to analyze. Therefore, We obtained more than 4.5 million rows original data in total.

2. By using URLLIB, IJSON, ITERTOOLS functions to disassemble the raw Json data, we Transformed the Json data to an array.

3. At this time, we found that there are lots of invalid data, such as the words of “a, an, the, they and so on”. Then we chose Natural Language Toolkit (NLTK) function to clean the data. Firstly, we used the “stop words” to clean the common invalid words. However, the NLTK library of stop words can not satisfy our needs. We had to expand the stop words library by ourself. Secondly, in order to get the topic words, we classified the words property. Finally, we filtrated the noun words as the News topic.

4. At last stage, we got around 100 thousand effective topic words to analyze. we counted the News topic words’ frequency and drew the data visualization graphics to analyze. We found there are some interesting relationship between some special news topics. For example, the frequency of attack is positive correlation with the frequency of Trump.

DATA ANALYSIS

2017 General Election Warm-up Stage

Overall, the top5 key words in April, May and June are election, party, state, minister and Trump which indicates the people are continuous concerned about general election and the issues about general election have high exposure in this stage. According to this, we can have a deeper analysis on comparing the trends of media exposure between different candidates and parties.

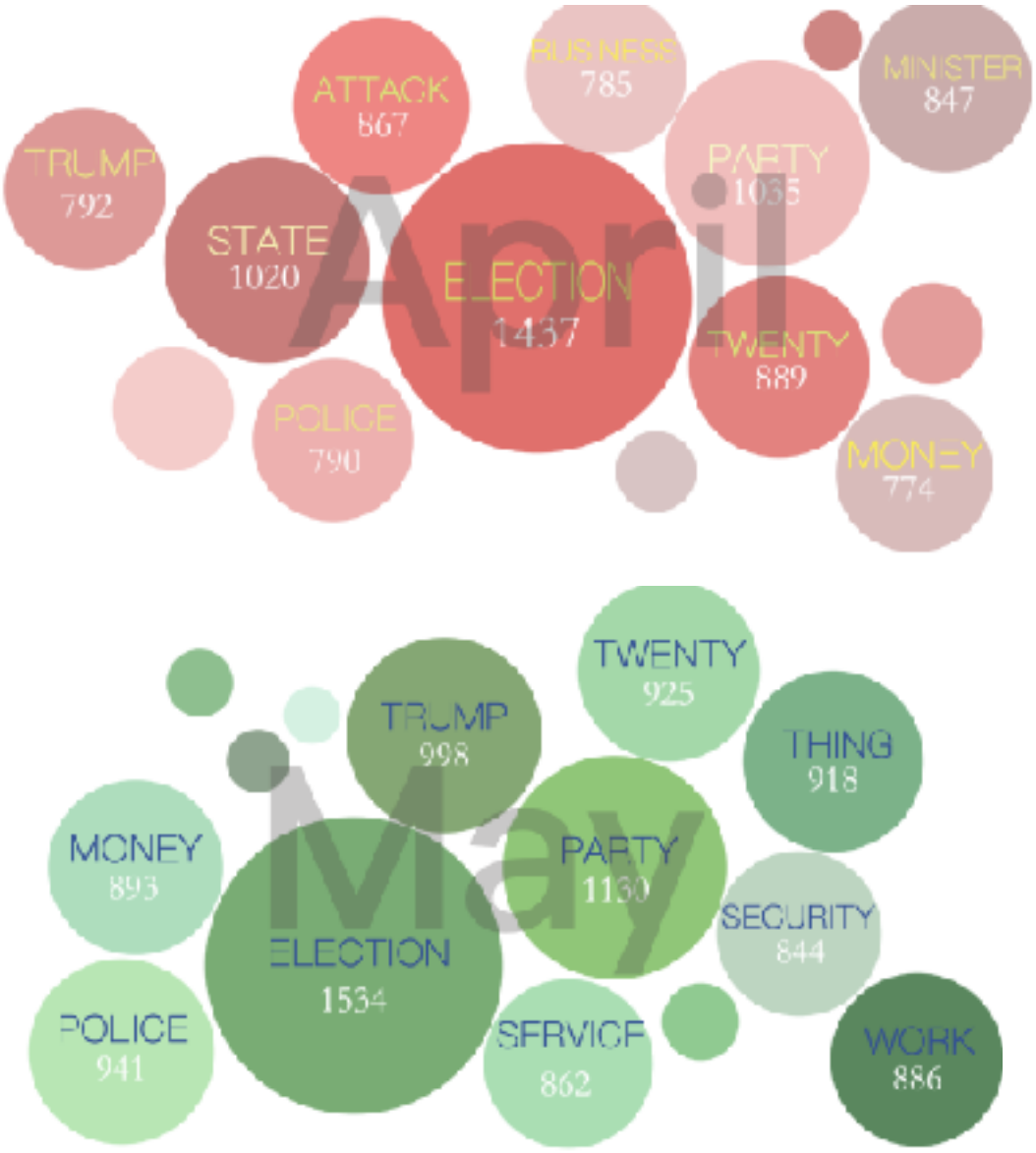


Chart1

April: The top3 key words in April can be found in chart1, they are election, party and state.

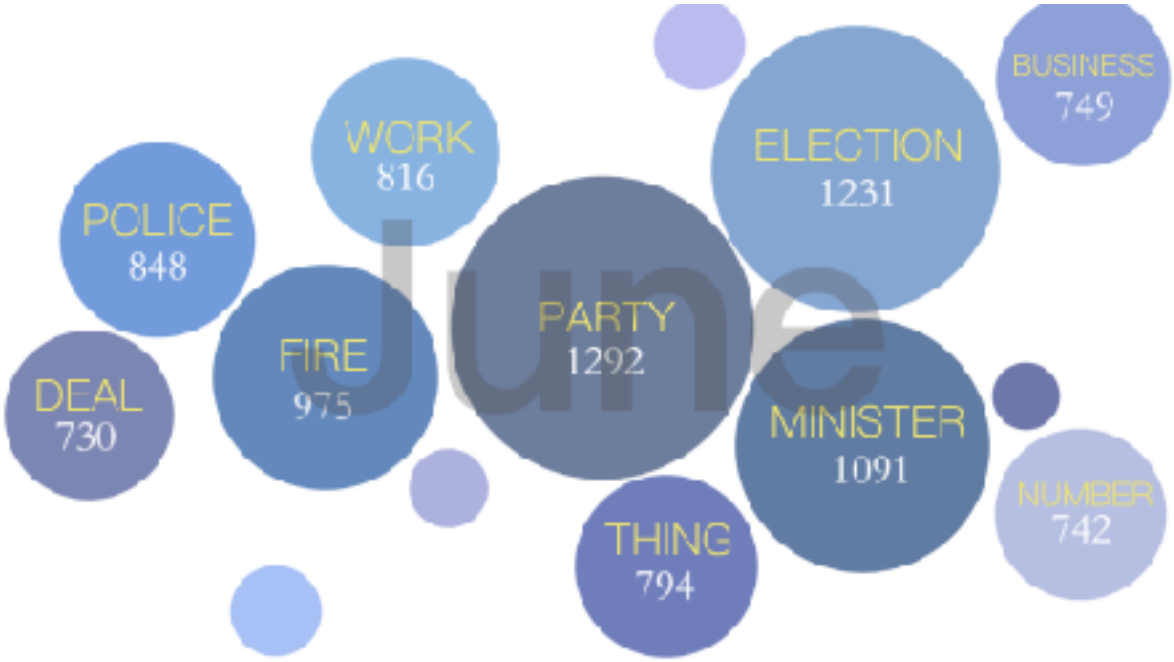


Chart2

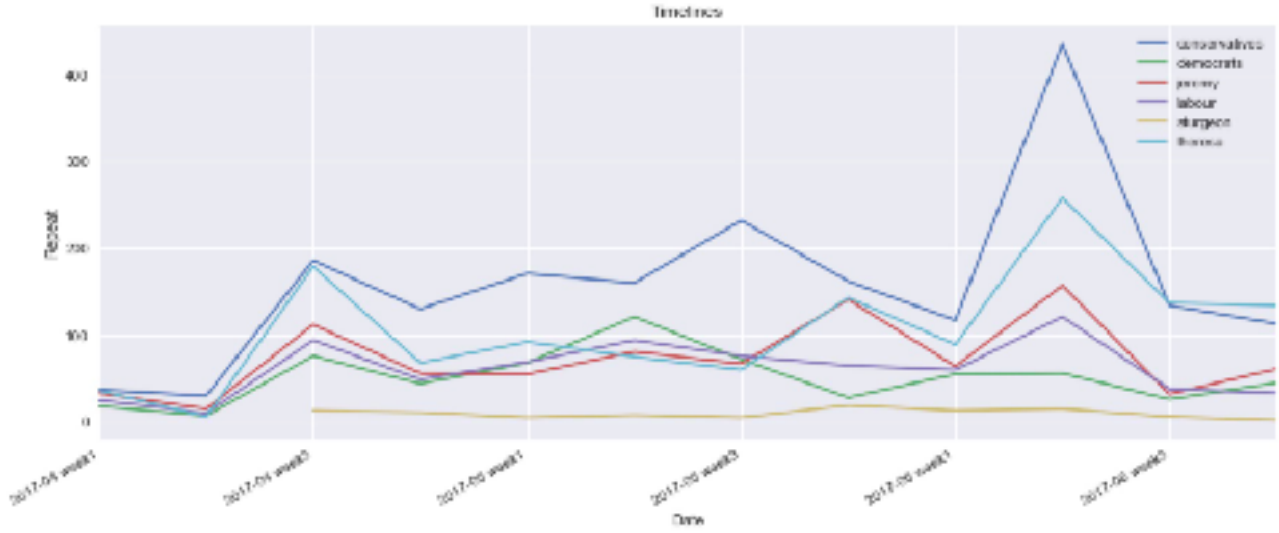
May: The top3 key words in May are election, party and trump from chart2.

Chart3

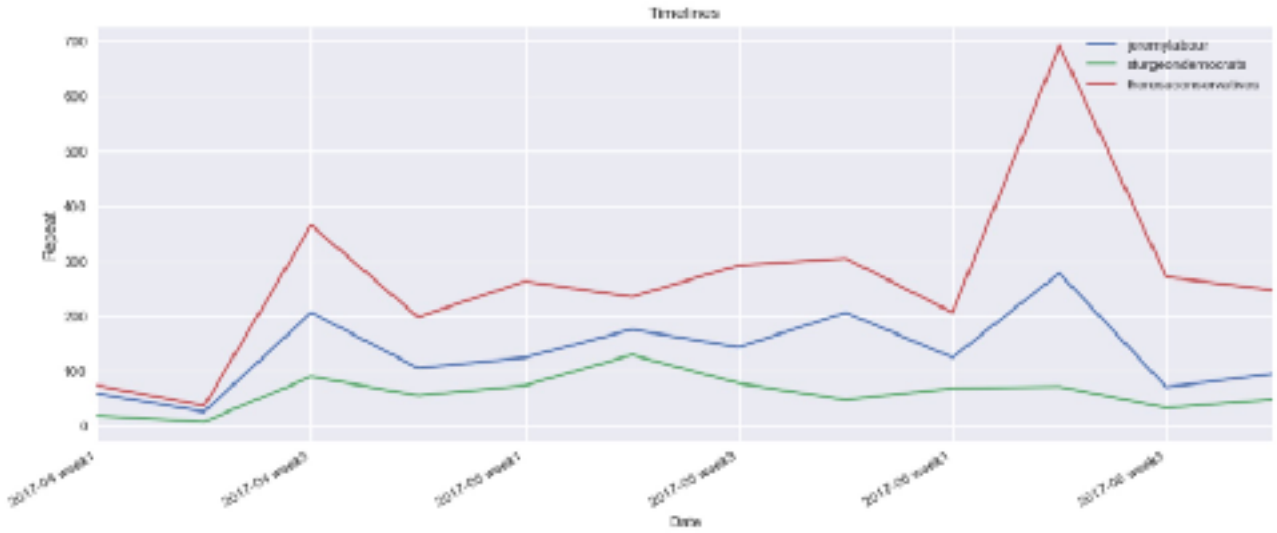
June: Party, election and minister are the top3 key words in June which can be seen in chart3.

THE RELATION BETWEEN GENERAL ELECTION RESULTS AND THE MEDIA EXPOSURE

The relation between the general election results and media exposure can be analysed from two aspects. One is the relation between the general election results and media exposure of candidates and another one is the relation between the general election results and media exposure of parties.

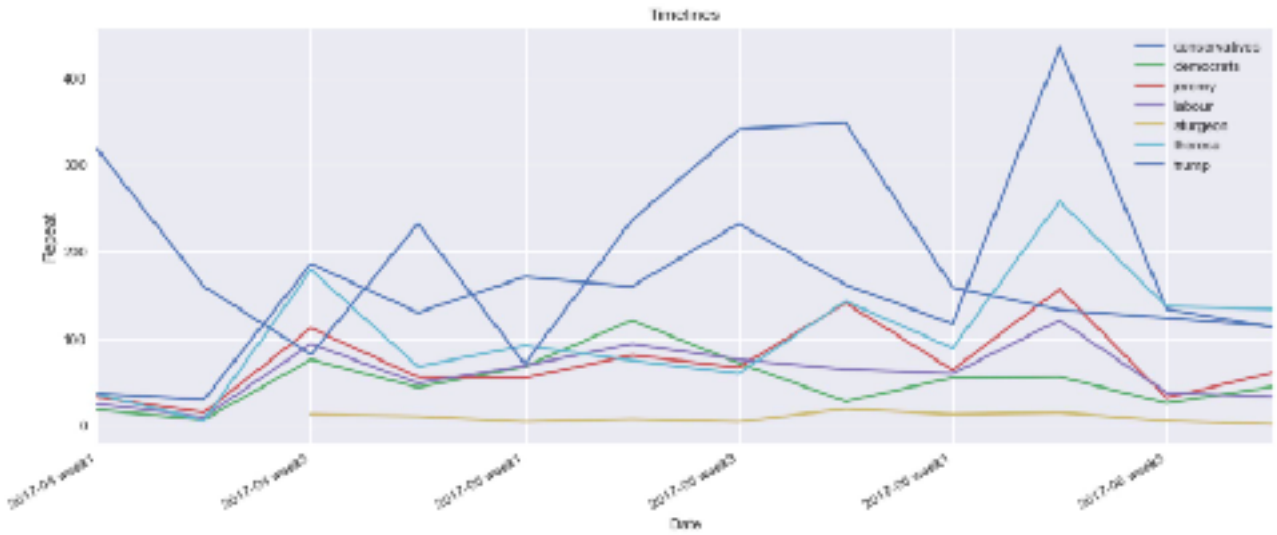


From this chart, we can find that before the 8th June, the three presidential candidates had the different rates of media exposure. In general, the media exposure rate of Theresa was higher than other two candidates. Theresa's party— the Conservative Party, won the largest number of seats and votes. As for the second aspect, we can find that one week ahead the general election the Conservative Party has the highest media exposure and the Labour Party has the second highest exposure and the Democrats has the least exposure which is as the same as the vote results. However, between week3 of April and week3 of May, there are come coincident points and overlap areas, so the relation between the general election results and media exposure of parties in this stage is not clear. To find out more information about the relation between the general election results and media exposure of parties and candidates, we combined the exposure of parties and candidates together and generate another chart.



Obviously, in this chart, the Theresa and her party—the Conservative Party has the highest exposure and Jeremy and Labour has the second highest exposure and the Sturgeon and Democrats has the least exposure which indicate there is a relation between the media exposure of each candidate and his/her party and the general election results

According to the analysis, we speculate that there is a relationship between the media exposure and the general election results. The higher the media exposure the higher the possibility of a candidate and his/her party winning. As for the relation between the general election results and media exposure of parties, there is no obvious relation between them.



WHAT'S MORE

Interestingly, during the period of UK general election, the American president Trump's media exposure rate was much higher than British future president. Probably, Trump made some attractive global news at that time.

Reflection

There is still a problem if we use media exposure to predict the general election. We cannot know whether the prediction result will affect the general election result and the media exposures of each candidate and party will be affected by the prediction result? There is still need more research.