A Twitter-based political party classifier

Balfre, T. & Barchiesi, D. & Hatzopoulos, V.
Capgemini
vasilis.hatzopoulos@capgemini.com — +44(0)7891 153336

Introduction

Over the last few years politicians have shown an increasing presence on Twitter, presumably as the platform provides the means to them to connect to their communities and voters in a direct way. In the run up to the 2015 UK general election a few of us from the rapidly growing Capgemini Data Science team decided to have a closer look at tweets coming from UK MPs' aiming to uncover certain themes and keywords that define the political language. We built a classifier system that is able to discriminate between labour and conservative MPs' with very high accuracy based on certain keywords appearing in the tweets. This implies a high uniformity of message within each party and divergent messages between them. The tweets by liberal democrats fall somewhere between the two.

Methodolgy

To start with we obtained the identities of 174 labour and 171 conservative UK MPs' from `http://tweetminster.co.uk/mps`. We then queried the Twitter API through the Python programming language to obtain the MP timelines for a range of days (roughly two months from 2015-02-25 to 2015-04-28) near the 2015 general election. We limited ourselves to 200 tweets per MP resulting in roughly 55k tweets. We then analysed the data through a combination of analytics tools in Python and R. We built a random forest classifier able to discriminate between labour and conservative MPs' on the basis of keywords in the tweets with accuracy over 95%. To extract the keywords standard text pre-processing techniques, such as removing stop words and urls', where first applied leaving us with sets of nouns, verbs, adjectives and adverbs. As using the individual tweets resulted in a significantly lower accuracy of 65%, tweets where aggregated by MP and a count over keywords was formed as a fingerprint characterising each of the politicians language.

Results

In Fig(1) we show the keywords most useful to the classifier in discriminating between the two parties. In particular it is the ratio of appearance of these words in the labour over the conservative tweets. If a keyword appears more often in labour tweets it will be plotted as a red point above the horizontal line. Conversely a ratio less than one means the keyword appears more often in conservative tweets and is plotted as a blue point below the line. We observe that:

Labour focus on words having to do with:

- Economic themes with negative sentiment such as **austerity, cuts, crisis, end.**
- Calls to arms such as **vote, register.**
- The conservatives - referred to as **tories**
- Themselves - **labour.**
- The PM - **Cameron.**

Conservatives focus on words having to do with:

- Economic themes such as **business, jobs, taxes, economy, income, unemployment.**
- Any negative effects of a coalition - **chaos.**
- Themselves - **conservative.**

To summarise the economy has dominated the MP posts on twitter with labour choosing to focus on urging people to vote so they as to end what they perceive as a negative economic environment. Conservatives focused on decreasing unemployment and what they perceive to be a better economic environment as well as the chaos of coalitions not involving themselves.

Finally we obtained tweets by 44 Liberal democrat MPs'. We then fed these to the classifier system wondering whether, on the basis of the same keywords, they would be predicted as labour or conservative. Perhaps surprisingly, there was an almost 50-50 split as from the 43 liberal democrat MPs' 21 where classified as conservative and 22 as labour.

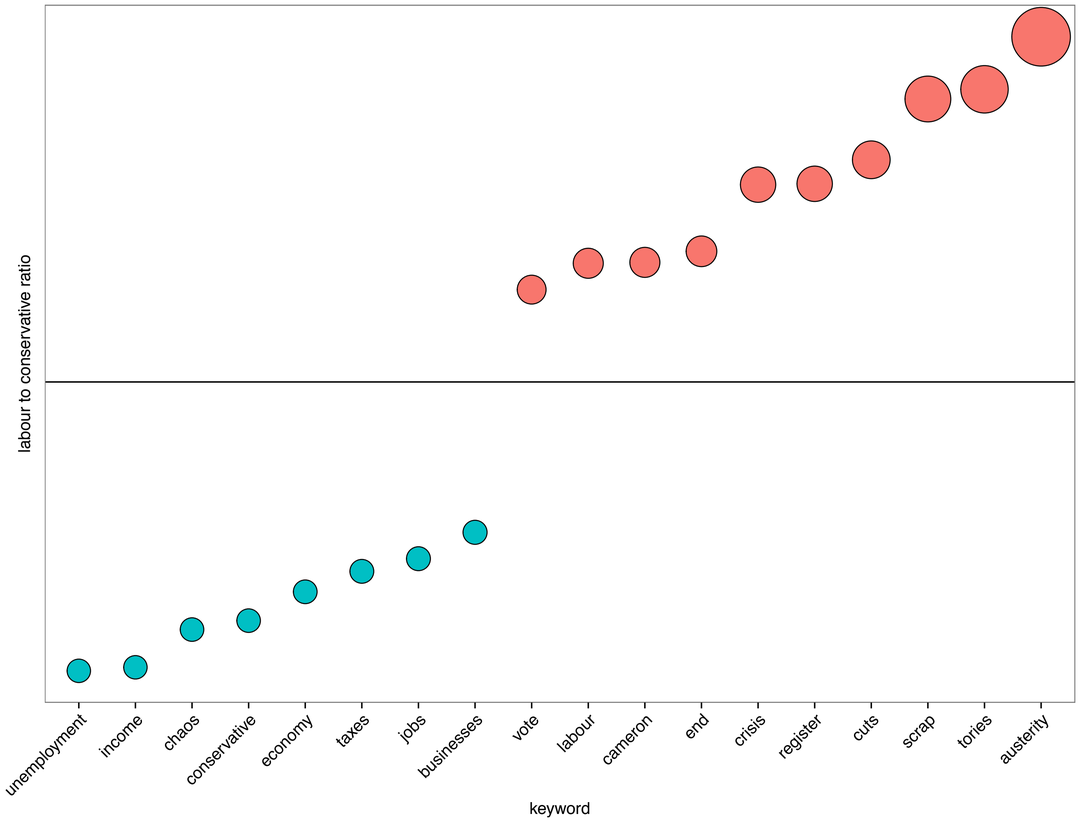

Figure 1: Figure caption

Conclusions

Over a sample of tweets per MP we see steady language usage patterns appear that are particular to each party. This suggests a uniformity of message within each party and a clear differentiation between them. The liberal democrats do not clearly fall into any of the two camps language wise and seem to use a more diverse keyword set, perhaps due to their part in the current coalition.

Acknowledgements