Inferring Political Opinion From Social Media Data

Angus Scott

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2015

Abstract

This is an example of `infthesis` style. The file `skeleton.tex` generates this document and can be used to get a "skeleton" for your thesis. The abstract should summarise your report and fit in the space on the first page. You may, of course, use any other software to write your report, as long as you follow the same style. That means: producing a title page as given here, and including a table of contents and bibliography.

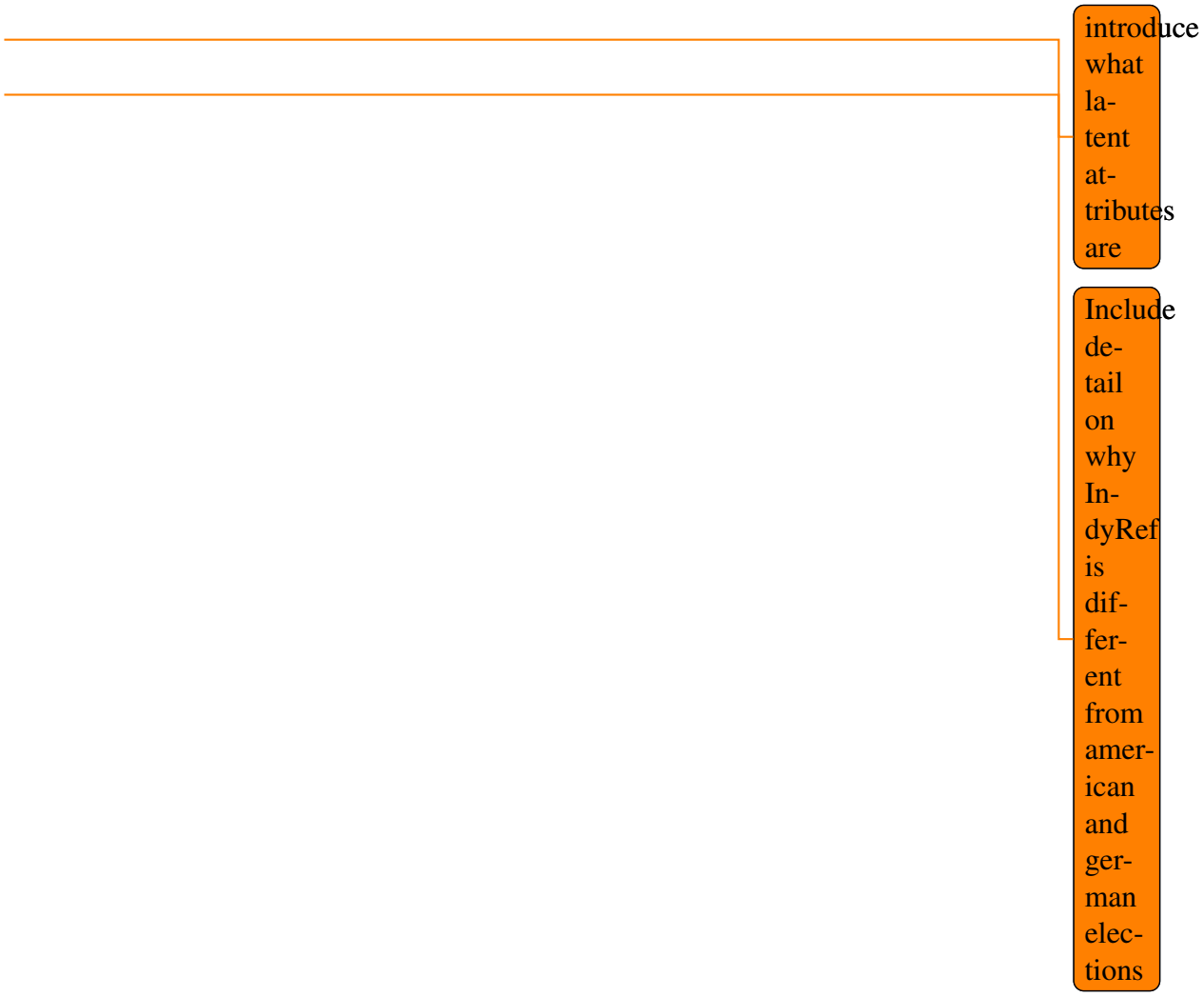
Acknowledgements

Insert Acknowledgements here.

Table of Contents

Chapter 1

Introduction



Chapter 2

Project Outline

Chapter 3

Related Work

3.1 Motivation

The inference of user attributes from social media, particularly with Twitter, has been considered in previous work. Examples typically include the detection of *gender*, *ethinicity*, *brand loyalty*, *regional origin* and *age*. [7] [9]

The detection of *Political Affiliation* has also been considered. Most of the research in this area has been focused on one of two political systems. First is the American system, with the 2010 Congressional Midterms [3] [4] [7] or the 2012 Presidential Elections [11]. Second is the German system, with the 2009 federal election to the national parliament [10] [5].

Typical approaches to the task of classifying a users political persuasion, using social media data, have generally focused around the use of Twitter. There are four main qualities used for classification:

Profile Features Features gained directly from a users profile, including: Bio information, usernames, number of followers and location

Tweeting Behaviour Features that capture how a user tweets, this includes: fraction of retweets, frequency of posting URLs and average number of hashtags

Linguistic Content Encapsulates the main points of interest/discussion as well as how they use language. Certain keywords may be highly indicative of a particular class, or certain topics of discussion may appear more frequently.

Social Network Who do users communicate with, you can construct graphical representations where users are a set of nodes and are connected to each other through twitter interactions such as retweets, mentions and follower/following relationships.

3.2 Profile Features and Tweeting Behaviour: "Who you are and how you tweet"

Most social media services, including Twitter, publicly share information about their users through their site or API such as their usernames, location, short bios, number of friends/followers or tweets/content. This could be used to infer some information about a user such as age of birth, or where the they live, which could all be used to help infer someones political persuasion.

As part of the work done by Rao et al. [9], they proposed six measures of tweeting behaviour:

Follower-following Ratio The ratio between the number of followers a user has and the number of users they are following

Follower Frequency The number of followers a user has

Following Frequency The number of users a person is following

Response Frequency Percentage of tweets from a user that are mentions

Retweet Frequency Percentage of tweets that are retweets

Tweet Frequency Percentage of tweets from a user that are not mentions or retweets

They then compared correlations between these six measures, and four different classification tasks: Age, Gender, Political Orientation and Regional Origin. They found there was no correlation between any of measures and the classes in the different tasks. This would imply that tweeting behaviour have little to no power for classification.

Pennacchiotti and Popescu [7] provided a method for obtaining profile features, and another for obtaining tweeting behaviours. They then built two Decision Trees, one for each method, and compared results to other Decision Tree based classifiers described later in this section.

To obtain profile features, the authors used regular expressions to strip users bio of key information such as gender and age using regular expressions. They also asked a pool of editors to identify ethnicity and gender based on the use of profile pictures. It was found that pictures can be misleading, and that bio fields do not contain enough high quality information to be used directly for classification purposes, and the location field is often to general such as states or countries or contains an imaginary place.

However, the authors still built a vector of profile based features including: length of username, number of numeric characters and alphanumeric characters in the username, use of avatar picture, number of followers, number of friends, friend/follower ratio, date of account creation, matching patterns from regular expressions and presence in the location field.

To capture tweeting behaviours, the authors built a vector of features including: number of tweets posted by the user, number and fraction of tweets that are retweets, number and fraction of tweets that are mentions, average number of hashtags and URLs per

tweet, fraction of tweets that are truncated, average time and std.dev. between tweets, average number and std.dev.of tweets per day, fraction of tweets posted in each of 24 hours.

The two methods performed weakly, with twitter behaviours having an accuracy of 61%, only 11% higher than random, which is inline with Rao's work and profile features having an accuracy of 68%. Profile features may appear to perform well, certainly compared to twitter behaviours, however it performs poorly compared to the 77% and 86% obtained with the ensemble of linguistic and network features respectively, which are discussed later in this section.

In conclusion, profile features and tweeting behaviours are poor measures to use to help classify users political leaning, or many other attributes such as gender and ethnicity. The data they produce is often too noisy and inconsistent to be used for classification purposes, and often don't have any correlation with different classifications at all. Therefore it is best to investigate classification methods based on linguistic and network based features, instead of using statistics of tweeting behaviours and profile features.

3.3 Linguistic content: "What you tweet"

Tweets contain linguistic information which directly conveys the thoughts and opinions of users. By analysing what the user says/discusses, we can observe similarities between users and from this, group users into different classes. It would be reasonable to assume that users of social media are likely to discuss topics that are of importance/interest to them, and people who share political ideologies/opinions would be likely to discuss similar topics.

Conover et al. (2010) [3] propose the use of linear support vector machines (SVMs) to classify 1,000 'highly connected' users into 'left' and 'right' political classes. The users were manually annotated into 'left', 'right' and 'ambiguous' classes. The tweets were collected over the six weeks preceding the 2010 U.S Midterm Elections. The authors compute accuracy for SVMs built with:

Full-text TF-IDF weighted unigrams of tweets (hashtags, mentions, URLs and stop words removed)

Hashtags Bag of words representation of hashtags in tweets (removed all text that wasn't hashtags and hashtags only used by one user)

Latent Semantic Analysis of Hashtags (LSA) Applied to the hashtag-user matrix to identify latent factors that correspond to political alignment.

It was found that the SVM trained on the full-text corpus obtained an accuracy of 79.2%, but when trained on hashtags the classifier obtained 90.8%. The addition of LSA to hashtags with the first three dimensions had a minimal improvement on accuracy of 0.1%, and the addition of subsequent features only decreased performance.

The paper also makes use of the social network, and is discussed in section 3.4 along with further discussion of what it means to be a 'highly connected' user and why the results in this paper may be overly optimistic.

Tumasjan et al. [10] is one of the most highly cited of the papers discussed here, and establishes the use of sentiment analysis as a means of detecting political persuasion. The paper uses a corpus of 100,000 tweets referencing the German federal election, collected over the 5 weeks preceding the election. The authors use the LIWC text analysis software and focus on 12 dimensions they feel are relevant to political sentiment: Future orientation, past orientation, positive emotions, negative emotions, sadness, anxiety, anger, tentativeness, certainty, work, achievement, and money.

The paper establishes that the tweets in their corpus were dominated by a small number of users, implying that a small number of users control the majority of the opinion on Twitter. The authors also establish that candidates who are further from the centre politically have more exaggerated sentiment scores amongst the 12 dimensions mentioned above. The paper concludes by attempting to show that the number of tweets about particular parties reflect the election outcome, and that co-occurrences of two parties reflect real life coalition partners.

Despite the paper being highly cited, there are two primary criticisms of this paper. First is that tweets were translated from German to English using some machine translation system, and sentiment scores were then calculated from the English translations, even though the LIWC software has a German dictionary. This is likely to introduces a considerable amount of noise, particularly in tweets where slang and incorrect/non-standard grammar is used. Balamurali et al. [1] found that sentiment analysis performed on translated text performed poorly compared to native text, they attributed this to the failure of the translation system to capture cultural divergence between languages with respect to expressions of sentiment.

The second criticism is the conclusion that number of tweets reasonably reflects the outcome of the election. The authors only compare against one election, and one data point is clearly not enough for any meaningful evaluation. In fact, in the Scottish Referendum corpus, constructed as part of this project, the number of 'Yes' leaning tweets exceeds the number of 'No' leaning tweets significantly, even though the No Campaign ended up being victorious.

Pennacchiotti and Popescu [7] make use of gradient boosted decision trees (GBDT), combining a collection of weak leaners to form a strong learner. The authors built a balanced corpus of 10,338 self identifying Democrat and Republican voters, with political leanings scrapped from various sources.

Their weak leaners are a combination of linguistic features and network features, which will be discussed in section 3.4. The authors use the following linguistic features:

Prototypical Words Collection of the 200 most class indicative words per class. Scored by measuring the ratio between the number of occurrences in a class and the number of occurrences in all classes. Similar to Full-Text above, but with less noise.

Prototypical Hashtags Collection of the 100 most class indicative hashtags per class. Scored by measuring the ratio between the number of occurrences in a class and the number of occurrences in all classes. Similar to Hashtags above, but with less noise.

Generic Latent Dirichlet Analysis (GLDA) 100 course grained topics built from corpus of 4 million users. The model is then applied to each test user in order to obtain their topic distribution, which are the features used for classification.

Domain Specific Latent Dirichlet Analysis (DLDA) Similar to the above, but model is built on users in training set. This produces more fine grained topics that are more class discriminative.

Sentiment Words Authors created a small collection of terms t that divide opinions between Republicans and Democrats, and used the tool Opinion Finder 1.5 to find cases where a word carrying sentiment was used with respect to a term in t. This was used across all the users tweets, and their scores tallied, to produce a score which indicates your most likely class.

The authors created a decision tree for each of the features above, and created a function to combine these features together which is a GBDT, an ensemble of the above features. Prototypical Words have an accuracy of 73%, Prototypical Hashtags 70%, GLDA 67%, DLDA 76%, Sentiment Words 70% and the ensemble method with 77%.

In conclusion, linguistic data can provide a powerful means of classification, with sentiment scores, topic modelling and prototypical phrases, we can build classifiers that out perform methods that use profile features and tweeting behaviour to classify a users political affiliation.

3.4 Social network: “Who you tweet”

Many social networks exhibit the concept of homophily - the tendency for users to associate/connect with those who are similar to themselves - and virtual social networks are no exception. As a consequence of this, information about the structure of the network can provide insight, to help infer properties about users that associate with one another, which motivates the use of social networks as a method of user classification.

Conover et al. (2011) [4] investigates the structure of retweet and user-to-user mention networks. The authors built a corpus of 250,000 tweets over the six weeks leading up to the 2010 U.S congressional midterms. With users given a class label ‘Republican’ or ‘Democrat’. The authors used a combination of network clustering algorithms, statistical analysis of political tweet content and manually-annotated data.

Using these tools, the authors found that retweet networks are highly polarised, taking the form of two large clusters of users who distribute content to others in their cluster, with very little overlap between the two clusters. Conversely, it was found that no such structure existed in the user-to-user mention networks, and instead was formed as one

large politically heterogenous cluster, where ideologically opposed users interact at a much high rate than in the retweet network.

The partisan nature of the discourse in retweet networks makes it ideal for the use of classification, hence many of the studies in this area have focused on the use of retweet networks for the purpose of classification.

In addition to their use of linguistic features, Conover et al. (2010) [3] make use of the social network to classify users. They argue that as many social networks exhibit homophilic properties as described above, they can be used to infer properties about users who tend to associate with one another. The authors form a retweet network composed of the 1,000 most connected users, in other words the 1,000 node subgraph, such that the sum total of the node's connections to the other nodes of the 1,000 node subgraph is maximised.

To classify users the authors used a greedy label propagation method. Each of the 1,000 nodes was given a class 1 or class 2 label using Newman's eigenvector modularity maximisation technique. The algorithm then begins iteratively assigning each node to the class shared by the majority of it's neighbours, and randomly assigning when a tie occurs. This was done until the algorithm converged on a solution.

The authors found a strong associated between cluster membership and political alignment, and suggested a simple classifier of 'Left' if in class 1 or 'Right' if in class 2. The method had an accuracy of 95%, higher than using semantic information. The authors also attempted to combine the two methods by combining cluster assignments and 19 hashtag features selected using Hall's feature selection algorithm and an SVM, but found it performed no better than network clustering alone.

Whilst the paper provides a solid approach to tackling the problem of detecting political alignment, there is a concern over a biased selection of users for training and testing. By selecting the 1,000 most highly connected users, it is likely that the results are biased, because as Tumasjan et al. [10] note, a small number of users dominate the communication in Twitter, and it is very likely that these highly connected users are part of that small group as retweets represent connections, which are observed through communication. Therefore the users who are used for training/evaluation, have an unusually high amount of linguistic content, and connections when compared to most other Twitter users. This leads to the possibility that the results are overly optimistic, if the aim is to use this as a means of classification for general twitter users.

Pennacchiotti and Popescu [7] make use of gradient boosted decision trees (GBDT), combining a collection of weak leaners to form a strong learner. The authors make use of linguistic features (discussed in section 3.3) and the following network features:

Follower Network Collection of the most exemplar users per class, based on follower relationships. Scored by measuring the ratio between the number of occurrences in a class and the number of occurrences in all classes. Then for each exemplar user in the collection, a boolean feature is set to '1' if a user follows them, and '0' otherwise.

Mention Network Collection of the 200 most exemplar users per class, based on user-

to-user mentions. Scored by measuring the ratio between the number of occurrences in a class and the number of occurrences in all classes.

Retweet Network Collection of the 200 most exemplar users per class, based on retweets. Scored by measuring the ratio between the number of occurrences in a class and the number of occurrences in all classes.

The authors built a decision tree for each of the network features above, and created an ensemble function which is a GBDT, an ensemble of the above features. Follower had an accuracy of 86%, Mention 68% and Retweet 66%.

It is worth noting that contrary to the work of Conover et al. (2011) [4], mentions have performed better than retweets. This is most likely attributable to the scoring mechanism used in this paper to find prototypical mentions. Mentions may generally be much less polarising, but the scoring still finds terms that maximise the use in one class, but minimise in other classes. So whilst mentions generally may be more heterogenous, the particular mentions selected are still prototypical of a class.

The above paper describes a range of linguistic and network based methods, all of which are weak/reasonable individually, but combine to become a strong classifier with an accuracy of 89%.

In conclusion, network based methods for classification can provide excellent means of classification. Retweet networks are politically more partisan than user-to-user mention networks, but does not stop them from an effective means of classification. Follower networks appear to be the best performing networks for classification, this is likely attributable to the fact that you are likely to follow someone because you are interested in what they say, it's a more active form of agreement that simply retweeting something.

3.5 Conclusion

Of the four qualities discussed, it was found that linguistic content and the social network provided the most accurate means of classification, whilst profile features and in particular tweeting behaviour before poorly. However, one issue of note is that for network approaches, the corpuses tend to be a larger or have been select in order to maximise performance by selecting most highly connected users. This approach is therefore unsuitable, or unlikely to yield as positive results as they have done in these studies if they were performed on a smaller corpus. Therefore the work of this thesis has focused on linguistic features, and will be discussed in Chapter 5.

Chapter 4

Scottish Indepdendence Referendum Corpus

4.1 Scotland Independence Referendum Background

The Scottish Referendum on Independence was a referendum on Scottish Independence with the intention of establishing Scotland as an independent sovereign state, independent of the rest of the United Kingdom. The election took place on the 18th September 2014, with record turnout of 84.59% and an electorate of 4,283,392. The referendum concluded with a "No" side victory of 55.3%, over the "Yes" side with 44.7%.

Yes Scotland was the primary campaigning group behind independence, whilst Better Together were campaigning to maintain the union with the rest of the United Kingdom. Many other campaigning groups, including: political parties, businesses, newspapers and famous individuals, were also involved in the campaign. The campaign was widely discussed on social media, with estimates by YouGov stating that 54% of the electorate got information, regarding the campaign, from social media and other websites. There was much debate during the campaign, on issues including: currencies and monetary policy, public expenditure, EU membership and North Sea oil revenues.

There was also considered to be several key events on the run-up to the campaign, including The Commonwealth Games that were hosted in Glasgow and the Ryder Cup held at Gleneagles both being hosted in 2014. The year also coincided with the 700 year anniversary of the Battle of Bannockburn, a landmark victory for the Scottish over the English in the First War of Scottish Independence.

4.2 Introduction

4.3 Existing Corpora

One of the issues of working with Twitter is that under the Twitter API Terms of Service [1] you may not share Twitter content, including datasets of tweet text and follower relationships. This therefore makes finding publicly available Twitter corpora very difficult. Because of this, many researchers build their own datasets using the Twitter API or buy them from third party companies such as DataSift [2] who pay Twitter to use their data.

Nonetheless some Corpora have have existed which have subsequently been taken down, or remain up either because they have explicit permission from Twitter to keep their datasets available, or Twitter haven't issued or enforced a take down notice. Such examples include the School of Informatics and Computing at Indiana University's "truthy" dataset [3], The Edinburgh Twitter Corpus [8] and Stanford's SNAP Twitter Dataset [6]. However, due to the time sensitive nature of the task, the truthy and SNAP datasets was not appropriate. However some of the infrastructure that were used as part of the Edinburgh Twitter Corpus were used, and is discussed in Section 4.4.

Before beginning the process of building my own corpora, I contacted DataSift asking if they would be willing to sponsor a sample of data over the referendum for the project. They very graciously offered to donate a one month sample over the course of the referendum, depending on the details of my project. However this was done in September, when the project was still in the early stages of planning, and there was concerns about how the one month window may not be adequate for some tasks that may have been completed as part of this project. I therefore thanked them and came to the conclusion that the loss in flexibility was too great for a project that was, at that time, in the planning stages. However given the same opportunity again, I would strongly consider working with a data provider, as it provides the opportunity to spend more time modelling and less time mining the data.

4.4 Corpus Construction

As

[1] https://dev.twitter.com/overview/terms/agreement-and-policy
[2] http://datasift.com
[3] http://truthy.indiana.edu/

Topic Indicators:	`#indyref, #scotland, #scotdecides, #scotlanddecides, #independence`
Yes Campaign:	`#voteyes, #yesscotland, #yesscot, #youyesyet`
No Campaign:	`#voteno, #nothanks, #bettertogether, #letsstaytogether`
Political Parties/Politicians:	`#snp, #labour, #conservative, #tory, #libdem, #salmond, #alexsalmond,#davidcameron, #gordonbrown, #alistairdarling`

Table 4.1: Set of hashtags related to the campaign

4.5 Corpus Evaluation

4.6 Corpus Statistics

Include timeline, most popular hashtags, URLS and mentions

4.7 Dataset

The dataset was collected following the referendum, and is composed of X tweets from Y users, spanning from the 1st February 2014, to 17th September 2014, the day prior to the Independence Referendum. The corpus also includes set of user classifications obtained through annotation of users tweets from the 18th of September.

Add number of users here

The first steps in building the corpus, was to obtain tweets from the 18th of September 2014. This first set of tweets were obtained from a Twitter Mining system developed by Sasa Petrovic and Miles Osborne as part of their efforts to create the Edinburgh Twitter Corpus. From this I was given a collection of tweets that matched a set of hashtags commonly used by users involved in the Yes and No campaigns.

Hashtags, and their associated campaign are given in Table 4.1, which were obtained using the hashtagify.me tool. It is worth noting that this is a set of non-event specific hashtags. Whilst hashtags such as `#PatronisingBTlady` [2] are politically divisive, they have been excluded as they have a short lifespan as discussed in Section **??**, and are therefore unlikely to be used on the final day of the referendum.

These tweets are then used as evidence for the annotators to classify a user as Y (Voted Yes in the referendum), N (Voted No in the referendum) or U (Unknown or Undeterminable). We make the assumption that tweets on the final day are the most emblematic of how a user voted, as the opportunity to flip their opinion is minimised andTwitter

Include graph showing strength of hashtag, and how many tweets it

activity surrounding the event was at it's peak. This is done as there are both technical and ethical issues about obtaining classifications more directly.

After providing a class for each user, we then collect all their publicly available tweets, available from 1st February to 17th September, and their associated meta data using the Twitter API [4].

[4]Due to a limitation of the Twitter API, you can only collect from the last 3200 published tweets.

Chapter 5

Methodology

5.1 Dataset

5.2 General approach to user profiling

5.3 Models

5.3.1 Naive Bayes

5.3.2 Topic Modelling

5.3.3 SVM

5.3.3.1 One class SVMs

5.3.3.2 Alternative SVMs, standard?

5.3.4 Gradient Boosted Decision Trees

Detail what can be measured in Twitter and how we could use it

Models I plan on implementing, may be moved to eval section

Chapter 6

Evaluation

Chapter 7

Conclusions and Further Work

Bibliography

[1] AR Balamurali, Mitesh M Khapra, and Pushpak Bhattacharyya. Lost in translation: viability of machine translation for cross language sentiment analysis. In *Computational Linguistics and Intelligent Text Processing*, pages 38–49. Springer, 2013.

[2] Libby Brooks. Scottish independence: no campaign's new ad convinces some to vote yes, August 2014. [Online; posted 27-August-2014].

[3] M Conover, B Gonçalves, J Ratkiewicz, A Flammini, and F Menczer. Predicting the political alignment of twitter users. In *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom)*, 2011.

[4] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Political polarization on twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[5] Albert Feller, Matthias Kuhnert, Timm Oliver Sprenger, and Isabell M. Welpe. Divided they tweet: The network structure of political microbloggers and discussion topics. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.

[6] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[7] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification, 2011.

[8] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. The edinburgh twitter corpus. 2010.

[9] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.

[10] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

[11] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets and retweets. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.*, 2013.