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

introduce what latent attributes are

Include detail on why IndyRef is different from american and german elections

Chapter 2

Related Work

The inference of user attributes from different sources of media has been well considered both in the context of traditional media and new media.

include links to papers that have perofrmed tasks in different media

A natural extension to this task has been to apply similar methods to the microblogging service Twitter. Previous work includes detection of *gender*, *ethinicity*, *brand loyalty*, *regional origin* and *age*. [10] [11]

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] [**?**] [10] or the 2012 Presidential Elections [14]. Second is the German system, with the 2009 federal election to the national parliament [13] [6].

Approaches to the task of inference of latent attributes from twitter can be broadly split into two categories:

1. Use of the networked structure of the data
2. Analysis of user communication streams

2.1 Network Structure

It has long been known that individuals form information networks corresponding to their own political preferences [8], and this has continued into online social networks [7]. This demonstrates that people who participate in particular networks, are likely to share similar political values or opinions, which could be used to infer someones political beliefs.

There are four main types of networks that have previously been exploited to infer political persuasion:

- *Follower* -
- *Mention* -

- *Retweet* -
- *Hyperlink* -

2.2 User Communication

Sentiment Analysis Topic Modelling Support Vector Machines Decision Trees

Chapter 3

Related Work

Broadly split into three approaches:

3.1 Profile features and Tweeting behaviour: "Who you are and how you tweet"

3.2 Lexical content: "What you tweet"

Tweets contain lexical 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 3.3 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. [13] 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 [10] 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 3.3. 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 an All function 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 All with 77%.

The above paper describes a range of linguistic methods, all of which are reasonable individually, but combine to become a fairly good classifier. When you introduce social network features - discussed in section 3.3 - the model performs strongly with an accuracy of 89%.

3.3 Social network: “Who you tweet”

Introduce why networks work, i.e. homophilic properties

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. [13] 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 [10] 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 3.2) and the following network features:

Follower Network Collection of the most exemplar users per class. 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

Retweet Network

Chapter 4

Data and Methodolgy

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 Glasgow

4.2 The Twitter Platform

Twitter is a popular microblogging and social network site that allows users to send and read short 140 character messages commonly known as *tweets*. Tweets as used in this report include the 140 character message, along with the associated metadata. This includes features such as: creation dates, user follower counts, coordinates and language.

As well as tweeting[1] to an audience of *followers*, Twitter users can interact publicly

[1] Tweeting - The act of broadcasting a tweet

through two main approaches: *retweets* and *mentions*. Retweets often indicate agreement [9], allowing users to rebroadcast content from other users, having the effect of spreading tweets to a larger audience [5]. Alternatively, mentions work by allowing someone to refer to a particular Twitter user by including their `@username` in a tweet, creating a public dialogue between the referrer and referee.

Hashtags are another important feature of Twitter, allowing users to tag tweets according to topic and their intended audience. For example `#IndyRef` was often used to reference the topic of the Independence Referendum, or `#bettertogether` which generally indicated that someone was No Voter or was addressing No Voters.

Twitter has several benefits over other social networking sites [12] [14] including:

1. Twitter users retweet notable events and participate in the spread of realtime news.
2. The 140 character constraint forces tweets to be concise and to the point.
3. Users are highly reactionary and discussed events tend to have happened in the recent past.
4. Tweets are media rich, and include content like video, image and hyperlinks along with text.

4.3 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 4.2, 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 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.

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

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

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 [2].

4.4 General approach to user profiling

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

4.5 Models

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

4.5.1 Naive Bayes

4.5.2 Topic Modelling

4.5.3 SVM

4.5.3.1 One class SVMs

4.5.3.2 Alternative SVMs, standard?

4.5.4 Gradient Boosted Decision Trees

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

Chapter 5

Evaluation

Chapter 6

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] Scott Golder Danah Boyd and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. *HICSS-43, IEEE:Kauai, HI, January 6*, 2010.

[6] 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.

[7] Itai Himelboim, Stephen McCreery, and Marc A. Smith. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *J. Computer-Mediated Communication*, 18(2):40–60, 2013.

[8] Robert Huckfeldt and John Sprague. Networks in context: The social flow of political information. *American Political Science Review*, 81:1197–1216, 12 1987.

[9] Panagiotis Takis Metaxas, Eni Mustafaraj, Kily Wong, Laura Zeng, Megan O'Keefe, and Samantha Finn. Do retweets indicate interest, trust, agreement? (extended abstract). *CoRR*, abs/1411.3555, 2014.

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

[11] 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.

[12] Claire Warwick Shirley Ann Williams, Melissa Terras. what people study when they study twitter: Classifying twitter related academic papers. *Journal of Documentation, 69*, 2013.

[13] 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.

[14] 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.