Week 2 & 3 assignment: Fundamentals of Data Science

Aswathy George, Antonios-Minas Krasakis, Taxiarchis Papakostas-Ioannidis, and Maarten Mol

University of Amsterdam, Science Park, Amsterdam, The Netherlands

Abstract. This report presents the results and visualizations of a Sentiment Analysis and topic modelling performed on a political tweets data set from a period during the 2016 US elections to find a correlation between the candidate preference and and the most discussed topics in each state (Sentiment analysis result and Topic modelling result). The following sections address the approach and findings of our research in detail.

Keywords: US Elections, Political tweets, Sentiment Analysis, Topic Modelling

1 Introduction

Ever since Twitter was founded in 2006, its users have been sending out 140-character text messages out to the internet. Often tagging the subject in hashtags, it provides many possibilities for computational analysis of people's interests and opinions. Given the polarization in the 2016 US elections between Hillary Clinton and Donald Trump, it is to be expected that the debate raged on through online platforms such as Twitter. In this research, we try to find whether the topics and opinions of Twitter users can be correlated to reflect a preference for a specific candidate in the elections. Using sentiment analysis and topic modelling, this research tries to correlate sentiment towards either presidential candidate to the topics most discussed in each state in order to gain more insight into the results of the 2016 election. This research is split up into the following three research questions:

- Election candidate preference based on the sentiment for each state.
- What are the most discussed topics in twitter for all the states and their relevance in individual states.
- Is there any correlation between the Candidate preference and topics discussed.

For the sentiment analysis and topic modelling, the source code can be found in the Github repository, `https://github.com/littlewine/USelections2016`.

2 Methodology

Our goal is to investigate the influence or correlation between political preference (as measured in our Twitter dataset) and topic discussions in each state. Towards this, we first have to measure the political preference in our tweet sample for each state. After achieving that, we will proceed with topic modelling to find which are the most discussed topics in each state. Finally, combining these two results we will allow their correlation to be calculated to answer our main research question.

2.1 Measuring political preference

In order to measure political preference on tweets, we have two different options. The first and most straightforward would be to use commonly used hashtags, such as (nevertrump, trumppence16, imwithher) which directly indicate a preference towards each candidate. However, we must note that many hashtags (such as trump, hillary, etc.) do not always indicate political preference. Additionally, misusing hashtags either on purpose or accidental is a common phenomenon.
Therefore, we chose to proceed by using sentiment analysis (or opinion mining), in order to define a sentiment class for each tweet (positive or negative) and then use user mentions, text and hashtag to categorize to which candidate does this tweet refer to.

The basic approach in sentiment analysis is identifying and classifying the polarity of the given text i.e whether the sentiment behind a text is positive, negative or neutral. The detailed approach is mentioned in the section 3.

2.2 Define the most discussed topics in each state

To approach this problem, we will use the technique of topic modelling. Topic modelling is a method that identifies and creates different clusters of words that often appear together within a set of documents/texts. We will label each cluster by hand according to the set of words that appear within it and then use this clusters to measure the distribution of topics appears within each corpus (tweet). Following, we will sum the topic distributions of each tweet corresponding to each state to extract the percentage of topic distribution within the whole state.

3 Implementation & Results

3.1 Datasets

- Dataset of approximately 0.5 million tweets, collected using the Twitter API, provided by the tutors of this course [1].
- US CENSUS dataset for demographics and population estimates [2].
- Twitter Sentiment Analysis training corpus , used to train a sentiment analysis classifier [3].

3.2 Preprocessing

After acquiring the twitter data, we imported them into a non-relational (MongoDB) database to allow for more efficient data handling and selection. Relevant tweets were selected and imported into a python pandas.DataFrame for further analysis. The tweets were filtered on language ("EN"), country ("US") and, finally, on the availability of their geolocation.

After selecting the relevant tweets for analysis, several text preprocessing steps were taken before performing the sentiment analysis. Text preprocessing is a crucial step in text mining and its importance on twitter sentiment classification is discussed by Singh et al. [5]. Part of the preprocessing scheme used by the authors in this research paper is used in our analysis.

Firstly, all characters were cast to lowercase and all html-links were removed, along with the twitter mentions and the hashtags and stored in a different column for future reference. Following, punctuation and stopwords were removed, while for the part of sentiment analysis we also used the NLTK english stemming library.

Also, tweets referring to either Donald Trump or Hillary Clinton (either in the form of plain text, mentions or hashtags) were flagged as such to later cross-reference their names and general sentiment towards them. Finally, each individual tweet was tokenized and stopwords were removed (using the english nltk.corpus.stopwords list).

3.3 Sentiment Analysis

For the task of the sentiment analysis, a classifier needed to be trained to categorize each tweet with a sentiment. Classification is a technique that falls under the umbrella of Supervised machine learning, meaning that we have to provide a dataset of labeled tweets to our classifier, before it can produce any results. In our case though, we did not have information about the sentiment class of each tweet (ground truth). Hence, in order to avoid classifying some of the provided tweets by hand, we used a dataset which was acquired from the web [3]. The dataset contained 1.5 million tweets, along with their sentiment polarity.

A classifier can only work with numbers though, so we have to represent our corpus into a form of mathematical scheme instead of words. Several techniques are proposed in the literature, amongst them the simplest being the "bag of words" implementation". In this approach, we assign each word contained in our corpus a certain id and represent a corpus with the term frequency of all words. Another proposed methodology is to normalize the term frequencies using the *tf-idf* model, as having a high raw count does not necessarily mean that one specific word is more discriminative than another. However, in our case, due to the limited length of our corpus (140 characters limitation by twitter we did not use this transformation.

After testing several different classifiers (such as the Naive Bayes, Logistic Regression and Support Vector Classification) we found that all of them perform similarly as described on the literature as well and chose to proceed with the Naive Bayes model. Further, we used the chi-2 feature selection technique to define a subset of the 10% most descriminative words and use them as features. Our classifier achieved a 76% accuracy and 76% $F1 - score$ on our external dataset. After identifying the sentiment of all tweets, they were pivoted to every state using a mean aggregate function where a positive tweet gets a value of 1 and a negative value of -1. As can be seen in figure **??**, the general tendency for tweets in our data-set is to be slightly negative. The main outliers are Montana and Louisiana which are more positive, and Kentucky which appears to be overly negative.

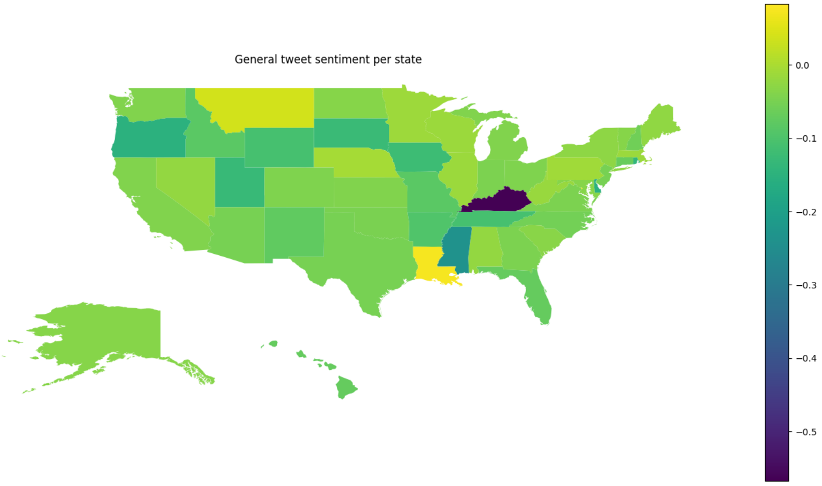


Fig. 1: Sentiment analysis of the USA

In figure 2 a prediction was performed based on the positive tweets each candidate had. Because, in every case, Donald Trump had way more tweets than Clinton, we needed to normalize the results and calculate the ratio between them, so as to categorize each state in a different color. Therefore, in figure 2 we colored the map based the highest number of normalized positives tweets. One possible interpretation could be that on Twitter, Trump seems to be more popular in the majority of the states in the US comparing to Hillary. This might be a political leverage because he is always on the top trend of the US elections, forming in that way the opinions about him.
However, figure 2 only shows an abstract picture of the predominant sentiment for each state. So, one step further, we included in our analysis a plot of a continuous color map based on the ratio between between the two candidates for their positive tweets, presented in figure 3. The same analysis also conducted for the negative tweets ratio in figure 3.

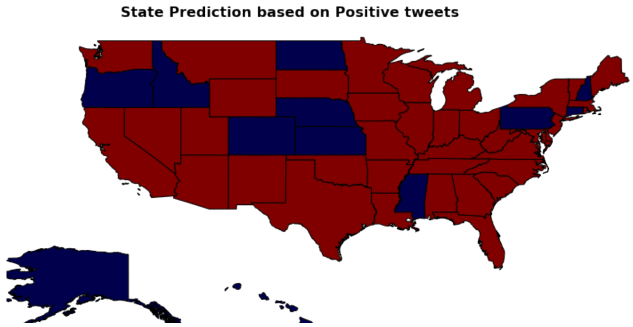


Fig. 2: Estimate of political preference in each state based on our sample of tweets

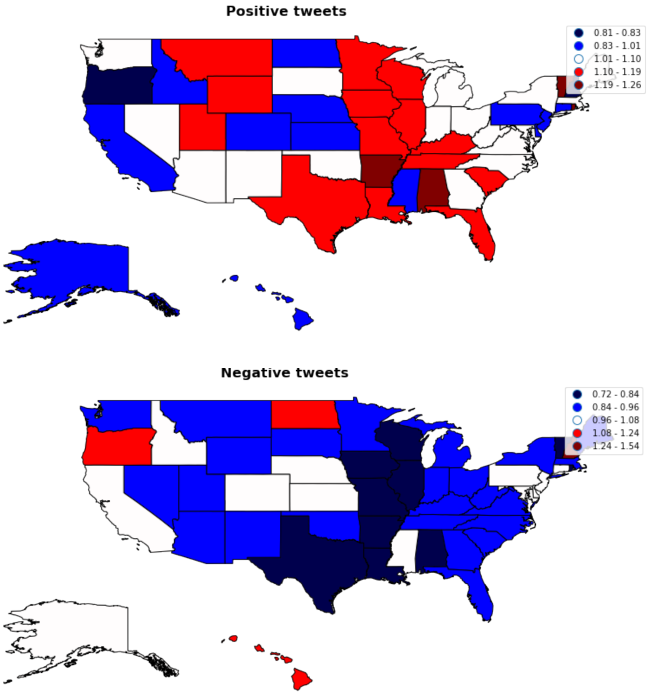


Fig. 3: Differences between the two candidates translated in percentages ratio for positive and negative tweets.

3.4 Topic Modelling

For the topic modelling part, Latent Dirichlet Allocation model was used produce the clusters of words. Two different approaches were tested to extract the topics:

- We trained a model to generate topics for the whole dataset of US tweets.
- We trained different models for each US state with the full dataset, classified into different datasets, based on states.

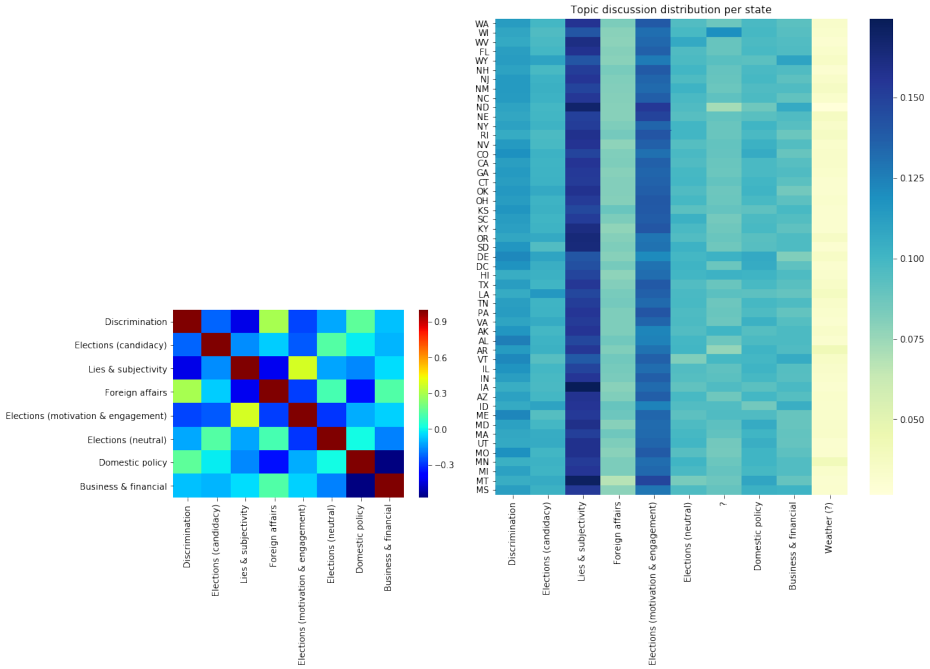
Also, a parameter of the LDA model is the number of topics. After experimenting with combinations of the options above, we concluded on choosing **10** topics on the whole US dataset, as the word sets were more coherent in that case.
Table 1 presents the top 10 topics from all US tweets dataset from the results of the LDA topic model output.

Nr	**Topics**	**Sample Words per Topic**
1	Lies/Subjectivity	[lies, believe, crooked, corrupt]
2	Elections(Campaign)	[hope, win, america, better]
3	Discrimination	[black, racist, Muslim, immigrants]
4	Finance/Business	[tax, health, security, classified]
5	Elections(Candidacy)	[president, leader, chief, candidate]
6	Home Affairs	[government, jobs, corruption, mexico]
7	Elections(Poll)	[polls, election, rally, rigged]
8	Adjectives/Pronouns	[funny, dumb, best, omg]
9	Foreign Affairs	[Putin, war, ISIS, Iraq]
10	Weather	[wind, temp, rain, forecast]

Table 1: Topic labels and sample words in each topic

Following, each word was represented as a distribution of the topics in which it appeared and likewise, each tweet was represented as a distribution of different topics. After adding up all distributions of tweets corresponding to each state, we concluded to a topic distribution per state.
We can also point out that topic set 1 (referring to truth, lies, corruption etc.) is the most popular. Following is subject 2 which is mostly about convincing people to vote. Hence, one useful insight is that Twitter is often used by campaign managers and supporters to engage and encourage people to vote.

In figure 5 we can see the Intertopic Distance Map between each topic. It is clear that the topic containing words related to weather is irrelevant to the rest and topics ”Home Affairs” and ”Foreign Affairs” share some common words. The rest of the topics are close to each other and share a lot of words with each other.



(a) Correlation heatmap between topics

(b) Topic distribution per state

Fig. 4: Topics correlation and distribution per state

In figure 4 we can see the correlation heatmap that states that talk more about domestic policy usually tend to talk less about Foreign affairs and Financial financial matters. On the contrary, we can see that in states where there are a lot of tweets trying to engage people to vote, we have a lot of talk about what is right or wrong, corruption and attacks to the opposite candidacy (”Lies and subjectivity” labeled topic).

In addition, we can see that topic ”Lies and Subjectivity” is the most discussed topic, while trying to engage voters, or engage supporters and volunteers also plays a major part in Twitter. Finaly, we can also notice that the irrelevant topic about weather which came up is by far the least discussed.

More advanced approaches on topic modelling in Twitter, along with a comparison and discussion for these methods can be found on ”Empirical Study of Topic Modeling in Twitter” by Liangjie Hong et al. [6]

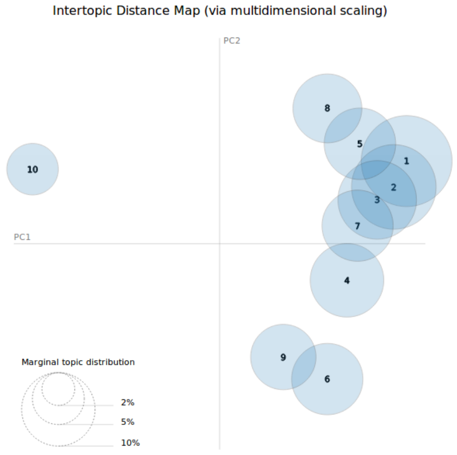
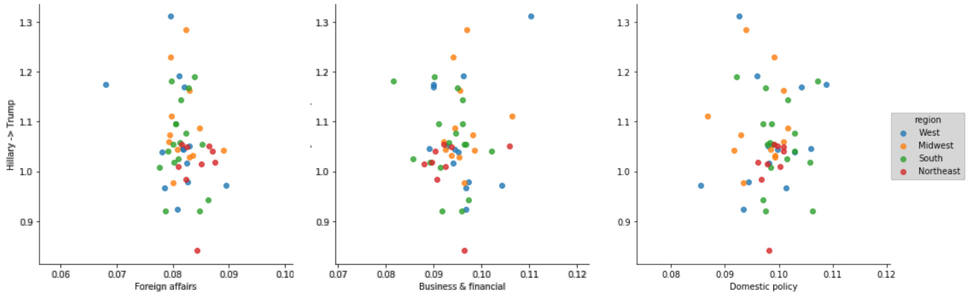


Fig. 5: Topics produced by Latent Dirichlet Allocation. Interactive version available at `http://bit.ly/2wWtsAf`

4 Final results - Discussion

The following scatterplots in 6 are our final results. Unfortunately, we cannot conclude to a concrete result on how each topic correlates with the political preference at this point. Therefore, it seems that either there is no correlation between the latter two, or that the models we used introduced a lot of error (noise) and hence we are unable to identify any underlying patterns. Another notable result is, that there seems to be no correlation between these scatterpoints (representing states) and the region that they belong to.



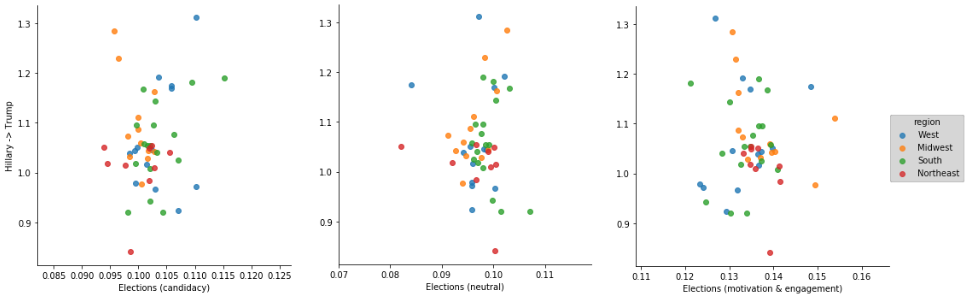


Fig. 6: Scatter plot of topics discussed and candidate preference per state

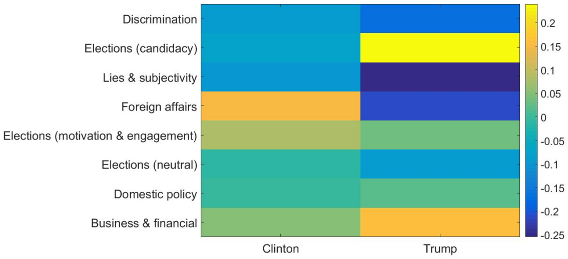


Fig. 7: Topics and their correlation to the general sentiment towards candidate

Discussion

Before going into a discussion of the applied methods, it should be noted that there are heavy criticisms on trying to use twitter data to predict election results. The most important criticism is the selection bias that occurs when using twitter data; not everybody uses twitter, and of those that use twitter, far from everybody tweets about politics [4]. In addition, twitter is not only the domain of citizens, as there are companies, bots and other institutions sending out tweets that might not be directly related to voting results.

Another limitation about using twitter to describe the political landscape is that political discussion are often rife with rhetoric. Sarcasm, humor and implied statements are commonplace, and very difficult to identify using standard sentiment analysis methods such as the naive Bayes classifier [4].

Another fact that has to be taken into consideration, the classifier for the sentiment analysis was trained on a data set not specifically concerning political topics, which may have impacted the actual accuracy. When this is combined with the LDA, which will also have less than perfect accuracy, these errors may add up to skew the final results.

We must note here that the actual accuracy of our sentiment analysis classifier could not be measured on this dataset, as we were unable to know the ground truth of the tweets. Hence, it might be promising to investigate the correlation of topic discussion and the election results themselves or approach our political preference question in a different way, such as directly drawing results from supportive hashtags etc.
More advanced methods in regards to that matter are discussed in a scientific papers

5 Conclusion

The sentiment analysis and topic analysis resulted in a number of interpretations. The first outcome of the analysis was the political preference of each state based on the positive tweets classified by our trained model for performing sentiment analysis. Arriving at the most discussed topics was the second task performed as part of this research and an LDA trained topic model was used for this purpose. Correlation plots were used to arrive at and present the correlations between the preferences and topics. But during the process we have also identified the difficulties/limitations that comes along with the sentiment analysis and topic modelling on tweets like sentiment classifier not being 100% accurate and training a topic model for tweets using LDA does not generate a good quality output as the texts are short and might require further text preprocessing. From the correlations in figure 7, it can clearly be seen that where the discussion is centered around the elections candidacy or business and financial topics, people tend to have a better sentiment towards Trump, whereas Clinton's approval ratings are high where the discussion revolves around foreign affairs.

References

1. Web scrapped tweets provided by instructors.(n.d.). Retrived September 13 2017, https://www.dropbox.com/sh/9vqjouzdroj45dt/AABeoQwVfwEMrvgA1qDtDE3ia?dl=0
2. US Census population statistics 2010-2016. (n.d.). Retrieved September 12, 2017, from https://www2.census.gov/programs-surveys/popest/datasets/2010-2016/state/asrh/
3. Twitter sentiment hand labeled corpus.(n.d.) Retrieved September 20, 2017, from http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/
4. Gayo-Avello, D. (2012). " I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"–A Balanced Survey on Election Prediction using Twitter Data. arXiv preprint arXiv:1204.6441.
5. Singh, T., Kumari, M. (2016). Role of Text Pre-processing in Twitter Sentiment Analysis. Procedia Computer Science, 89, 549-554.
6. Hong, L., Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics (pp. 80-88). ACM.
7. Prusa, J. D., Khoshgoftaar, T. M., Dittman, D. J. (2015, May). Impact of Feature Selection Techniques for Tweet Sentiment Classification. In FLAIRS Conference (pp. 299-304).