Assignment 1: Politics (Twitter)

Group 4 - Fundamentals of Data Science
Authors: Milos Dragojevic (10853456), Joris van der Vorst (10210717), Tycho Atsma (10791779), Andrew Paterson (12777153), Antonio Javier Samaniego Jurado (12799505).

Abstract

Twitter is a valuable source of data because it provides with massive amounts of information which can give insight into people's opinion. This project uses a topic modeling and sentiment analysis among other data analysis techniques on the 2016 US Election related tweets to extract valuable election insights and show that sentiment of tweets holds a relationship with the election results.

Introduction

In this project, a combination of time series, topic modelling and sentiment analysis is performed on a dataset of tweets collected during the 2016 US Presidential Election. For tweet sentiment extraction, a Naive Bayes classifier and the IMDb dataset was used, achieving ~76% accuracy. In addition, Latent Dirichlet Allocation (LDA) from gensim and an LDA model with Mallet algorithm was used for topic modelling. Combining those techniques with a series of data splits, signals from tweets metadata such as hashtags and user mentions, and different data analysis methods, a research question on whether the tweet sentiment addressed to candidates can predict election results is both defined and answered in detail, along with a set of insights.

Methodology

Data Acquisition and Pre-Processing

The twitter data is provided to analysts by the University of Amsterdam in the form of a JSON file containing 657307 tweets for the period 12/08/2016 - 12/09/2016. To allow for easier access and processing, the file is read with Python and stored in a MongoDB collection. Tweet text is processed using methods provided by the Natural Language ToolKit (NLTK) platform. A ‘sanitizer’ class is created and in this manner stop-words elimination, lemmatizing and tokenizing is applied. Punctuation removal is applied via Python’s string methods. Hypertext links as well as Twitter’s @ mentions are removed, again using Python. The application of the sanitization methods are applied against all tweet text.

Sentiment Analysis

Initial exploration of the sanitized twitter text is performed with NLTK's Naive Bayes Classifier trained with the 'Movie Review' IMDb corpus. A train / test split of 60:40 was found to deliver an accuracy of 0.76625. To perform a positive / negative classification against a tweet, the word features were extracted from the tweet's sanitized text and passed to the classifier. The results were saved together with tweet id and text allowing the sentiment to be used with other emerging data derived from the provided tweets. After sentiment analysis was performed, the tweets mentioning one of the candidates were selected. These tweets were aggregated by state to compare the sentiment with the actual votes cast [1]. For each state the difference between the positive and negative sentiment tweets was calculated and divided by the number of tweets mentioning the candidate in that specific state. This resulted in a number between -1 and +1, where -1 means all negative and +1 means all positive sentiment. Finally the lowest number for both candidates is added to this aggregate data to create a positive vector. The difference between these vectors is subsequently compared to the difference in percentage of votes in each state.

Tweet Sentiment Density Overview

An overview of tweet sentiment density across the US for the time period contained within the data was generated from the results of the sentiment analysis in order to provide the analysts with a notion of tweet weight per state. To generate the metric used in the overview the following formula was applied: 1 - (1 / (sum of all positive tweets + sum of all negative tweets)).

Topic Modeling

Topic analysis on the twitter dataset was performed using the Latent Dirichlet Allocation model. Hereafter, the results of the analysis, combined with a critical look at the limitations, but also opportunities for future work, are reported. After processing the dataset, a model is chosen. Two different LDA models are used, the LdaModel model provided by the `gensim` library, and the LdaMallet model. The LdaMallet model is a model that wraps around the default LdaModel model using Mallet's algorithm.

Testing and tuning

Before the model can be tuned, tests need to be run to lead to the selection of suitable values for the hyperparameters of the models. A model's coherence is used as an indication of how well a model performs.

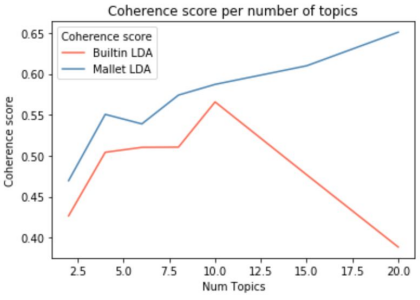

Fig. 1- Coherence scores

Each model is trained iteratively over a list of numbers, where each number represents *k* number of topics. The results of the iteration are shown in *figure 1*. This figure shows the coherence scores in relation to the number of topics. It also shows that Mallet's model shows the highest overall coherence scores. Based on this figure, further topic modelling is done using the Lda Mallet model with a value *k* of 20 as the number of topics.

Results and Discussion

Topic Modelling

Using the best model, the topics can be plotted in a graph to get an illustrative view of the model. Each bubble on the left-hand side plot represents a topic. The larger the bubble, the more prevalent that topic is. A good topic model will have fairly big, non-overlapping bubbles scattered throughout the chart instead of being clustered in one quadrant. This plot of a model trained using the optimal parameters mentioned earlier, is shown in *figure 2*.

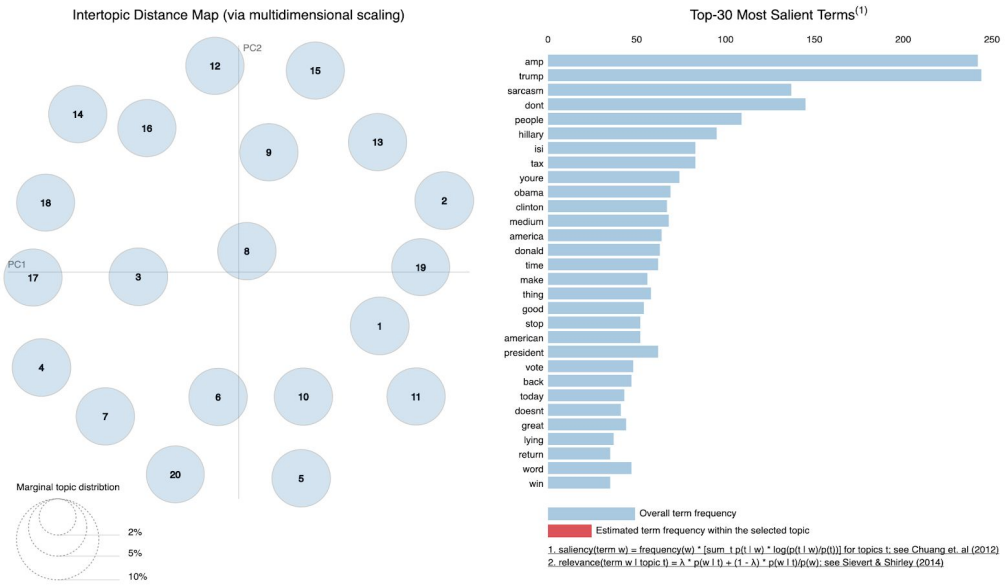

Fig. 2 - Intertopic Distance Map and Top 30 Most Salient Terms

Using the plot in figure 2, it can be seen that the topics are not overlapping. This implies that all topics are separate from each other. However, when looking at the top contributors, certain topics look similar. This could be because of the huge quantity of topics or their apparent complexity. Two new datasets are created to test if the topics are able to prove themselves useful between different timespans; a subset of tweets for the first week and the last week of the main data set.

Based on the analysis of the newly created models, one can argue that the top contributors to each topic do not change significantly. However, the portions of tokens that do not include the top contributor do change. Consider "trump" and "hillary". The topics where these contribute the most

contain different words between the models, that is, the first and last week. Also, “hillary” is a more salient term in the last week.

Relative Tweet Density

The relative density of tweets of either sentiment within the states provide insight into potential issues with relative weight of sentiment per state. This is shown in *figure 3* with brighter states having a higher density, darker a lower.

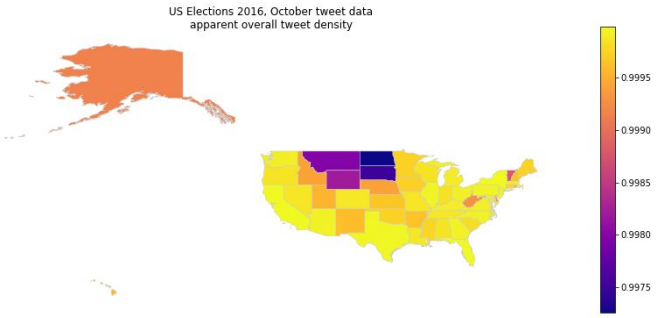

Fig 3. - Tweet Density per state

Naive Bayes sentiment per state and candidate

In order to determine the sentiment towards each candidate all tweets with the feature “country.code = US” and valid location data were selected. These tweets were filtered on whether the username of one of the candidates was mentioned. This resulted in 325834 tweets mentioning Donald Trump, 140740 mentioning Hillary Clinton and 28253 tweets mentioning both candidates.

For the entire country the sentiment score of Hillary Clinton was -0,113 and the sentiment score for Donald Trump was -0,130. *Figure 4* shows the sentiment scores for each candidate and each state. *Figure 5* shows the difference in sentiment score compared to the vote count. Although no numerical score can be calculated, these seems to be an overlap between the state sentiment per candidate and the difference in percentage of votes.

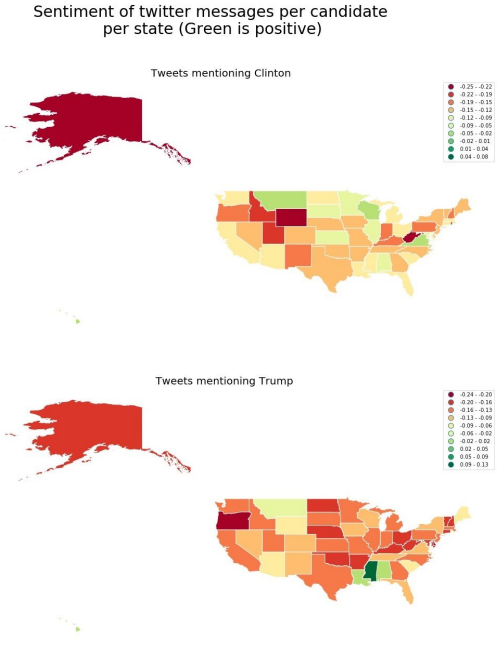

Fig. 4 - Sentiment scores

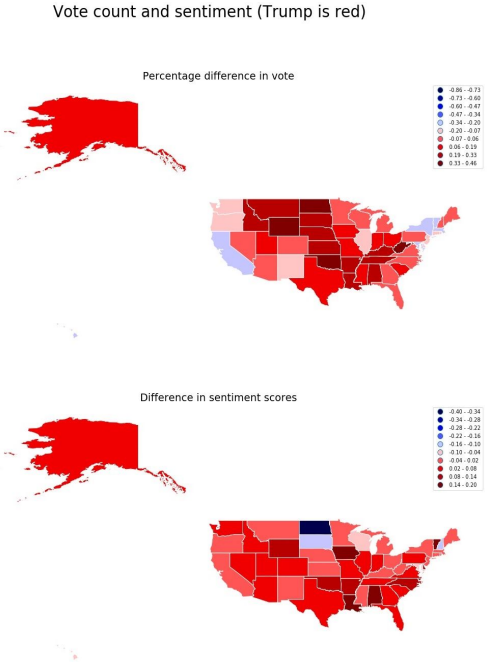

Fig. 5 - Vote count and Sentiment score

Hashtags and @ Mentions

We analyse the count of hashtags and @ mentions to see which of the two candidates is the most trending candidate and whether the hashtags are mostly positive or negative. In the data there are 86421 different users of which 37648 use hashtags and 68785 use @ mentions.
We see that the general hashtag “trump” was the most popular hashtag and more popular than “hillary” and “hillaryclinton” combined. It is interesting to note that although “nevertrump” is the second most mentioned hashtag, “maga” is still is more popular than “imwithher”. The most popular @ mention was @realdonaldtrump which refers to Trump’s Twitter account. @realdonaldtrump was mentioned twice more than @hillaryclinton. These numerical facts allow us to infer that Trump was the most trending candidate.

Hashtag Sentiment Analysis

In order to perform the sentiment analysis on the hashtags, preprocessing was performed. Hashtags composed of more than one word were split. For example: “nevertrump” becomes “never trump”. A text file containing 466k words used in the english language was used in this procedure. Election specific words such as “maga” or “hillary” added were added to the corpus. Note that in the English language, “trump” means “to be better than”, and could be considered a positive word and was therefore categorised as neutral to avoid influencing scores.
Hashtags were classified into positive or negative sentiment using positive / negative corpus sources [3]. Every word in the hashtag was examined for presence in the sources and scored as positive minus negative score.

Time Series - Hashtag Sentiment

Following hashtag sentiment extraction, time series analysis was performed on both hashtag sentiment and mentions, aiming to extract findings of both candidates. Figures 6 and 7 show how hashtag frequency varies over time, for both positive and negative hashtags, on Trump and Clinton.

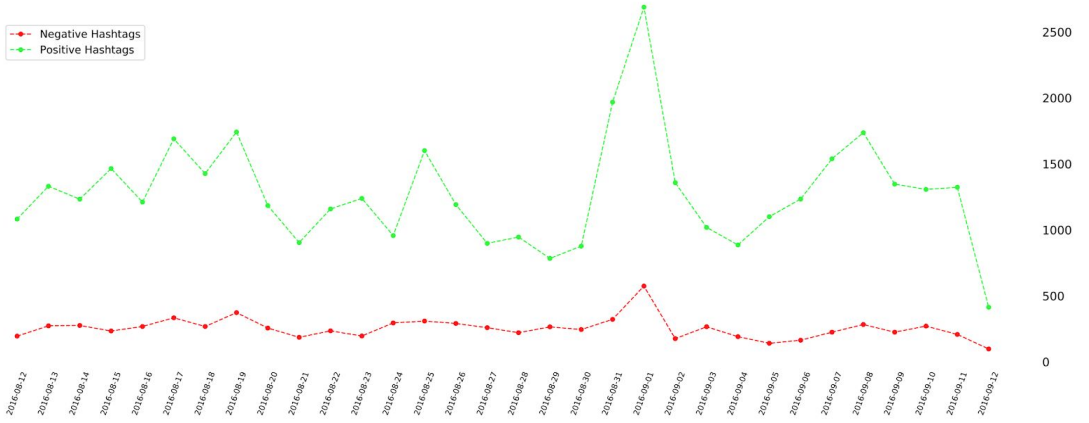

Fig. 6 - Hashtag Frequency Over Time, per Sentiment: **Donald Trump**

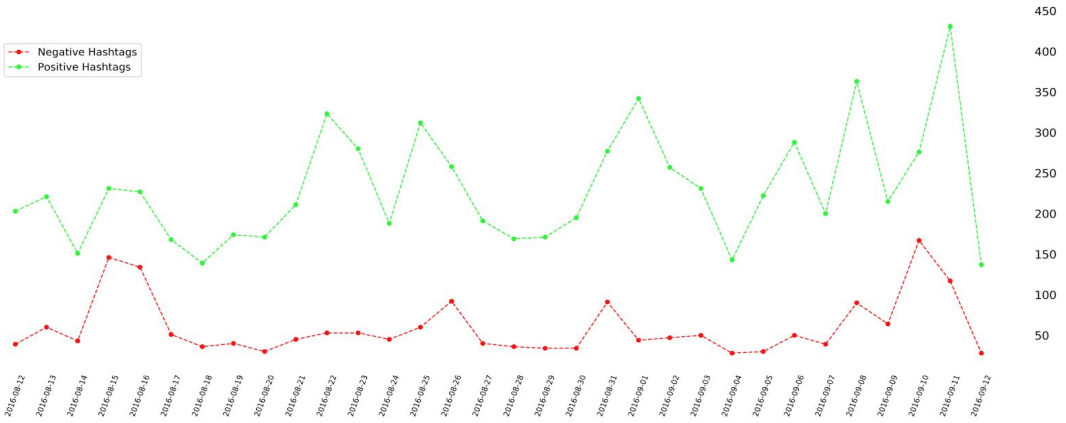

Fig. 7 - Hashtag Frequency Over Time, per Sentiment: **Hillary Clinton**

This analysis takes all positive and negative hashtags for each candidate. It is clear from *figure 6* and *figure 7* that more positive hashtags were written towards each candidate than negative over time, meaning more good reactions from their supporters than bad ones from opponents.

An interesting finding is the positive peak in hashtags directed towards the candidates. On September 1, 2016, when he tweeted: *"Mexico will pay for the wall!"* [4]. This led Trump's supporters to be more active on Twitter than his opponents, increasing the positive reaction. Similarly and among other peaks, on September 11, 2016, Hillary Clinton left the 9/11 Memorial early after feeling 'overheated,' being diagnosed with pneumonia [5]. This likely explains why the hashtag sentiment directed towards her switched from negative to positive on that day.

Sensitive Tweets

A further analysis was also performed on sensitive tweets. These are tweets that Twitter itself labels as *possibly_sensitive*, meaning their text or media (e.g. image, GIF) contains sensitive material.

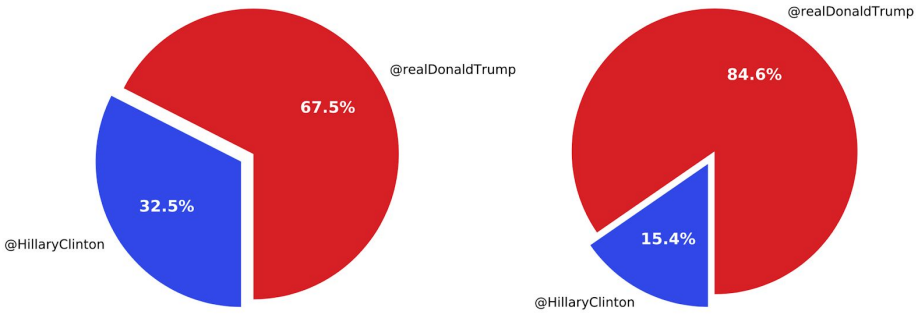

Fig. 8 - (**Left**) Proportion of sensitive **tweets** that mention each candidate.
(**Right**) Proportion of sensitive **users** that mention each candidate.

Sensitive material is mostly on tweets mentioning @realDonaldTrump (67.5% of sensitive tweets against 32.5% that mention @HillaryClinton) as seen in *figure 8*. In terms of sensitive users (authors of at least one sensitive tweet), the proportion is 84.6% Trump vs. 15.4% Clinton. This shows that not only is Trump popular on Twitter, he also deeply divides the political online debate which is reflected by him being linked with sensitive material much more than Clinton.

Conclusion

This research was performed with the purpose of examining the link between tweets written before the 2016 US Presidential Election and the actual election results. To that aim, a sentiment analysis as well as a topic analysis of the tweets was performed. Following a description of the context of the US election, the data and methodology used throughout the study were presented. Based upon the data and the election results, a hypothesis was formulated. The hypothesis examines if there is a relationship between the sentiment of tweets addressed to the major candidates and the outcome of the election. The results show that the sentiment of tweets addressed to the major candidates does reflect the election outcome.

One of the limitations of this study is the number of parameters studied. A tweet contains more than 100 attributes which can be used in various ways to perform a sentiment analysis. For example, tweets these days are not limited anymore to 140 characters. If users had more space to express their sentiment would this have resulted in more negative or positive sentiment, or could this have lead to a more balanced stance? Further, tweets may not capture the sentiment and vote intentions from less tech literate demographics such as the eldery or from lower socioeconomic status.
Future research could focus itself on analysing whether the number of tweets addressed to a candidate is more crucial than the sentiment of the tweets themselves. Is a lot of bad publicity better than a few of positive one?

References

[1] http://history.house.gov/Institution/Election-Statistics/Election-Statistics/
[2] https://github.com/dwyl/english-words
[3] https://gist.github.com/mkulakowski2
[4] https://www.nytimes.com/2016/09/02/us/politics/transcript-trump-immigration-speech.html
[5] https://www.nbcnews.com/politics/2016-election/hillary-clinton-falls-ill-9-11-memorial-n-y-n646376