How do you feel about Trump's tweets?

Fayez Mourad
fayez.mourad@epfl.ch

Yamane El Zein
yamane.elzein@epfl.ch

Abstract

In this report, we summarize our work and findings on the Applied Data Analysis (ADA) final project. This project relates to Donald Trump's tweets, more particularly the topics addressed by Donald Trump in his tweets, and the reactions of Twitter users to Donald Trump. Our work is divided into 4 main parts: (1) data collection and exploration (2) topic extraction (3) sentiment analysis on reactions to Trump and (4) analysis of results. We discuss each of these steps in this report.

1 Introduction

On November 8th 2016, Donald Trump, businessman and TV personality, was elected president of the United States. He is known for having controversial opinions and expressing them bluntly. One of the platforms through which Donald Trump conveys his views is Twitter, a social media platform where users can write (or 'tweet') short texts of maximum length of 140 characters, and broadcast them to their followers. President Trump is active on Twitter, and often voices his opinions on the platform. He uses '@realDonaldTrump' as his Twitter handle (username). In this project, we take a look at some of the tweets made by Donald Trump. We extract the topics from his tweets, and study which topics are most prevalent over the years. We also look at tweets by other Twitter users, which are replies or reactions to Donald Trump. We predict the sentiment (positive or negative) of these tweets by performing sentiment analysis on them, and we analyze the results with respect to the previously extracted topics, and with respect to time.

2 Related Work

Trump's tweets have been the center of attention of many before us. In the literature, a paper published by researchers from the political science and computer science departments of the University of Rochester, called *Catching Fire via Likes: Inferring Topic Preferences of Trump Followers on Twitter* (Wang et al., 2016) discusses topic extraction from Trump's tweets using Latent Dirichlet Allocation (LDA) in order to infer the topic preferences of Trump's followers on Twitter. They do this inference by looking at the number of likes that Trump's tweets have received. Our approach will be different from theirs, as we will be using Elasticsearch for topic extraction instead of LDA, and we will not only look at the likes of Trump's tweets, but also perform sentiment analysis on the replies to Trump.

3 Data Collection

For this project, we make use of two main datasets for our analysis, in addition to a helper dataset that we use to train a sentiment classifier.

The first dataset we use is the Trump Tweets dataset provided to us by the ADA course team as a link to the Trump Twitter Archive (www.trumptwitterarchive.com). This dataset contains tweets by Donald Trump from the years 2009 until 2017. We were able to download the archive from the link in comma separated format.

The second dataset is the Tweets Leon dataset, which is a substantial dump of tweets by a variety of users, which was made available to us on the ADA cluster. From this dataset, we filter out a subset of tweets that contain Donald Trump's Twitter handle (@realDonaldTrump) in order to get tweets that are either replies to Trump or that mention Trump. We use this subset of tweets as our second dataset.

Finally, we use a helper dataset of tweets which are labeled as either 'positive' or 'negative', meaning that they either convey a positive or negative sentiment respectively. This helper dataset is used only to train a sentiment classifier. This

dataset was provided to us by the Machine Learning course team.

4 Datasets Description

We start by describing the Trump Tweets dataset. This dataset contains the following fields: Source, Text, Date, Retweet count, Favorite Count, Is Retweet, and ID. For our analysis, we focus on the Text, Date, Retweet count, and Favorite Count fields.
The original dataset we acquired contained 29597 tweets by Donald Trump. Of these 29597 tweets, 5 contained NaN values for some of our fields of interest. We drop these tweets from the dataset.
We look at the Retweet count and Favorite count fields, and find that on average, a tweet by Donald Trump is retweeted 2662 times and favorited 8729 times by Twitter users.
We now describe the Trump replies dataset. As previously mentioned, this dataset was obtained by taking tweets that contained Donald Trump's twitter handle in them from a substantial Twitter dump. Originally, we obtained 500253 that contained Trump's Twitter handle. We remove tweets that are simply retweets of Donald Trump, by removing any tweet that starts with 'RT @realDonaldTrump', since they do not contain any valuable input by the author of the tweet. We also only retain tweets written in English. After this, we end up with a dataset consisting of 352151 tweets to work with. The fields in this dataset are: Language, Tweet ID, Date, User, and Text. For our analysis, we are interested mostly in the Text and Date fields.
Finally, we briefly discuss the dataset used to train our sentiment classifier. It contains 1250000 tweets labeled as positive and 1250000 tweets labeled as negative. No other fields apart from the text are present.

5 Methods

5.1 Text preprocessing

As we are working with text data, we perform cleaning on the tweets by removing word contractions, urls, user tags, stopwords, special characters, numbers and punctuation, and by performing word lemmatization. This cleaning is helpful for both the topic extraction and sentiment analysis which are both discussed below.

5.2 Topic Extraction

We start by extracting topics from Trump's tweets. We experiment with three different methods of topic extraction: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and Elasticsearch.

5.2.1 Latent Semantic Analysis

We first try Latent Semantic Analysis on Trump's tweets in an attempt to extract the 16 most relevant topics from the tweets. LSA creates a vector based representation of the tweets, analyzes the relationships between these vectors, and produces a set of concepts or topics related to these tweets (Wiemer-Hastings, 2006). Using LSA, we end up with 16 concepts, with a set of keywords under each concept. However, after examining the topics, we find the results of LSA on Trump's tweets unsatisfactory, since the keywords that fall under the topics are not always well grouped.

5.2.2 Latent Dirichlet Allocation

We then try Latent Dirichlet Allocation on Trump's tweets. This approach is similar to the LSA approach. It models tweets as a mixture of topics with certain probabilities, and extracts common topics from the corpus of tweets (Blei et al., 2003). We also aim to extract the 16 most relevant topics with LDA. As in the case of LSA, the results we obtain from LDA are not relevant enough for our analysis. We therefore abandon the LSA and LDA approaches and focus on using Elasticsearch for topic extraction.

5.2.3 Elasticsearch

Finally, we perform Elasticsearch on the set of Trump's tweets in order to extract topics from them. In order to do that, we build a dictionary of predefined topics, with keywords relevant to each topic, in order to perform Elasticsearch queries on the corpus of tweets. We list the topics we chose to work with in the table below:

Abortion	Economy	Civil rights
Corporations	Crime	Drugs
Education	Environment	Foreign policy
Trade	Gun	Healthcare
Immigration	Jobs	Tax
Technology	Religion	War

Table 1: Topics used for Elasticsearch.

Elasticsearch works as a search engine over the corpus of tweets, and extracts all tweets which are relevant to a certain query from the corpus (Divya and Guyal, 2013).
We make queries relating to the keywords of the topics listed above on Trump's tweets, and label the tweets that result from each query with the query's topic. We do the same on the tweets which are replies to Trump.
We are left with 13118 by Donald Trump and 45719 tweets in reply to Trump, that fall under one of our topics.

5.3 Sentiment analysis

In order to see how Twitter users feel about Trump's tweets, we perform sentiment analysis on the replies to Trump dataset, by using a Fasttext (Joulin et al., 2017) classifier trained on the helper dataset mention in the Datasets Description section. This labels each of the replies to Trump with either a positive or negative tag, which means that the tweet conveys a positive sentiment or negative sentiment respectively.

6 Analysis and Results

In this section, we use the results of topic extraction and sentiment analysis to find trends in topics mentioned by Donald Trump and the reactions of Twitter users to these topics.

6.1 Trump Topics over time

We look at the number of times topics are mentioned by Donald Trump over the years. For the years 2009 to 2011, the number of tweets by Trump in our dataset is very low compared to other years. We exclude these years from our analysis and focus on the years 2012 to 2017.
We start by looking at the number of occurrences of the topics for all years, and summarize the results in the tables below:

Topic	Number of occurrences
Foreign Policy	2269
Corporations	2127
Jobs	1647
Economy	1032
Crime	838

Table 2: Top 5 most mentioned topics

Topic	Number of occurrences
Abortion	7
Drugs	105
Technology	147
Civil rights	243
Education	308

Table 3: Top 5 least mentioned topics

We then try to see if the pattern of occurrence of a topic differs from one year to another. We show the results for the years 2012 and 2017 only in the graph below, since the results for other years are very similar (the plots for years 2013 to 2016 can be found in the notebook).

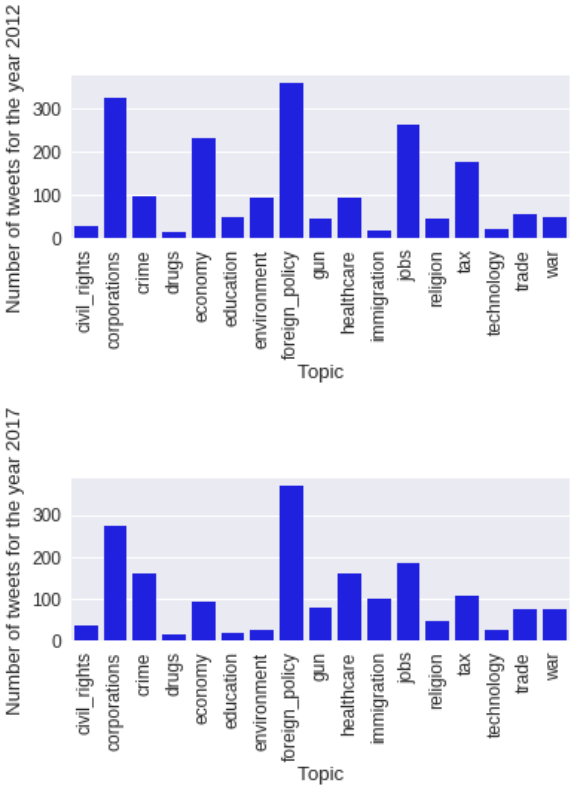

Figure 1: Topic occurrences in Trump's tweets for 2012 and 2017

Therefore, the distribution of the topics is similar from one year to the other. The topics of Foreign policy, Corporations and Jobs are systematically more mentioned by Trump than other topics. For the topics of Foreign policy and Corporations, we could say that these topics seem to occur more because they are broad topics. However, the Jobs topic occurs often as well, but is not a broad topic.

6.2 Reactions to Trump's tweets

6.2.1 Preliminary analysis based on number of retweets and favorites

We first assess the popularity of a topic mentioned by Trump based on the average number of retweets and favorites per tweet for this topic. We show the results in the figure below:

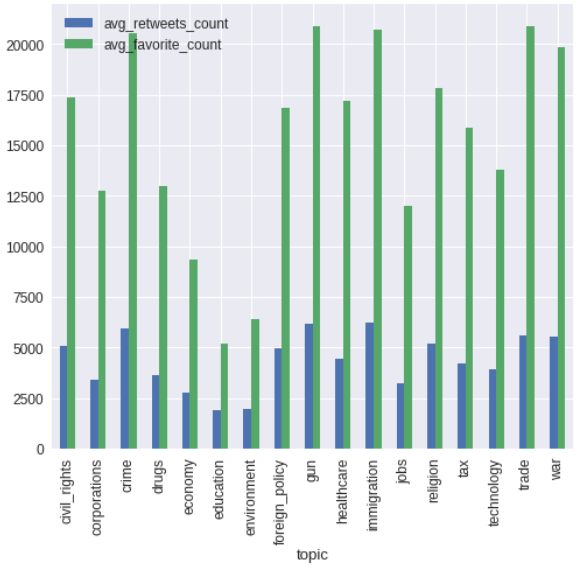

Figure 2: Average number of retweets and favorites of Trump Tweets per topic

Based on these two metrics, the most popular topics among Trump's followers seem to be Guns, Immigration, Crime, Trade and War.

6.2.2 Analysis based on tweets sentiment

We strengthen our analysis by looking at the sentiments that Twitter users convey when replying to Trump's tweets for a certain topic. We find that for all topics except Drugs, Twitter users convey more negative sentiments toward Trump's tweets than positive ones.
The topics for which the discrepancy between the number of negative tweets and positive tweets is most prominent are Economy, Immigration, War, Healthcare and Foreign policy.
The figure below summarizes our findings:

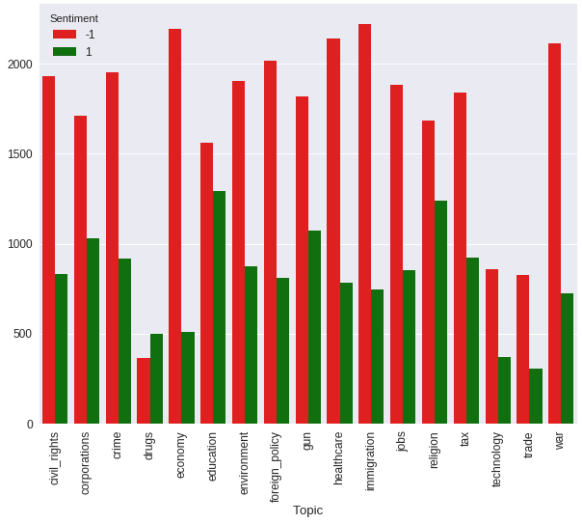

Figure 3: Sentiments of Twitter users towards topics of Trump

Finally, we look at the percentage of positive tweets among replies to Trump for the years 2012 until 2016 (the replies dataset does not have any replies for 2017), to see how sentiments of Twitter users towards Donald Trump evolve with respect to time.

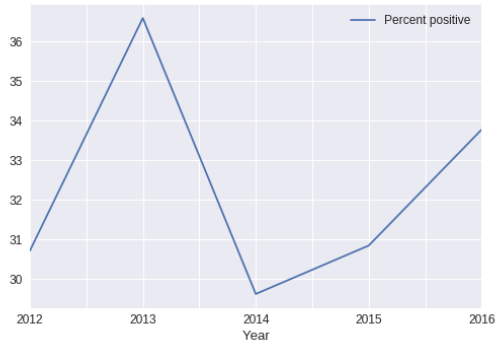

Figure 4: Sentiments of Twitter users towards topics of Trump

The percentage of positive tweets fluctuates within a narrow range (31-38%). It reaches its lowest in 2014, one year before the electoral campaign starts, and is at its highest in 2013. With the electoral campaign in 2015, it increases slightly and keeps increasing until 2016.

7 Conclusion

We have looked at Trump's tweets and seen that his most mentioned topics are consistent over the years, and are mainly Foreign Policy, Corporations, Jobs, Tax and Economy. We have also seen

that Twitter users generally disagree with Trump and convey negative sentiments towards him in their Tweets, especially when it comes to the topics of Economy, Immigration, War, Healthcare, and Foreign policy. In 2014, the percentage of positive replies to Trump's tweets is at its lowest, but increases in from 2015 to 2016 with the electoral campaign.

References

Blei, D. M., Ng, A. Y., and Jordan, M. A. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, page 9931022.

Divya, M. S. and Guyal, S. K. (2013). Elasticsearch: An advanced and quick search technique to handle voluminous data. *COMPUSOFT*, 2(6):171175.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

Wang, Y., Luo, J., Niemi, R., Li, Y., and Hu, T. (2016). Catching fire via likes: Inferring topic preferences of trump followers on twitter. *ICWSM 2016*.

Wiemer-Hastings, P. (2006). Latent semantic analysis. *Encyclopedia of Language and Linguistics*, page 706709.