MIE1624 Project Report: Fake News Detection

Submitted by Group 8

Kunal Bambardekar, 1005618002
Renu Indapurkar, 1005659039
Amr Mohamed, 999593079
Rafael Perez, 1006594641
Collin Wilson, 1002258175

1. Problem Statement

With the aim of helping citizens make more informed political decisions during federal elections, many initiatives and organizations have recently launched their fact-checking services. To highlight the importance of such services: Politifact, a leader in verifying the statements of political candidates during elections, has found that 70% of 740 statements uttered by US president Donald Trump during the last US election campaign lacked any truth in them, compared to Hillary Clinton's 25% of 240 statements[1].

Traditionally, fact checking has been done by journalists and editors by verifying the content of claims. This project has been undertaken to study if and how Machine Learning (ML) can be utilized to verify the content of claims in support of these fact checking efforts. The data used in the research behind this report has been obtained from data web-scraped from 9 of the leading fact-checking website and provided to participants of Kaggle's 2019 "Leaders Prize" competition.

The question which we seek to explore is: *'Can certain words or combinations of them help us predict the truthfulness of associated claim?'* With this, the rest of the report is structured as follows: We perform some preliminary data analysis to form a few hypotheses in Section 2. In Section 3 we review the results of 2 of the studied ML algorithms before discussing a possible improvement in Section 3. We end the report by discussing the findings of the paper.

2. Preliminary Data Analysis

Web-scraped data often contains html characters and punctuation symbols that do not affect the meaning of a sentence. These characters are removed. Words are stemmed according to their part-of-speech using a lemmatizer. What this does is stem plurals to their singular form (eg. cats to cat) and verbs to their roots (eg. played to play), amongst other things. We lower-case all letters and remove stop words which are just common insignificant words (eg. the, we) from the text corpus. Now the data consists of instances/observations corresponding to a clean claim, and its label. The labels are {0: false, 1: partly true, 2: true}.

Figure 1 shows the label counts in the data. Observe that there are much more false labelled (blue) claims that true labelled (green) claims. We will see later in the report that a classifier trained on this data will be able to more accurately classify false claims as compared to true ones due to the larger training set available for false labelled claims.

Figure 2 shows that the length of claims after cleaning is now more or less normally distributed around a center of 7 words. Based on this, a reasonable combination of words in a claim is 2 or 3 words. Larger combinations will be very specific to their associated claims.

[1] "Comparing Hillary Clinton, Donald Trump on the Truth-O-Meter", [online] https://www.politifact.com/truth-o-meter/lists/people/comparing-hillary-clinton-donald-trump-truth-o-met/

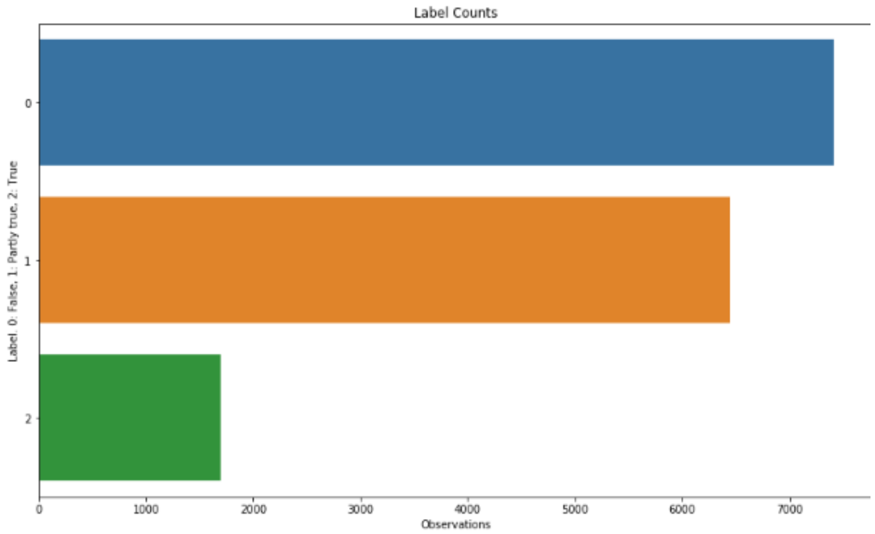


Figure 1. The label counts of claims in the data. The labels are false (blue), partly true (orange) and true (green)

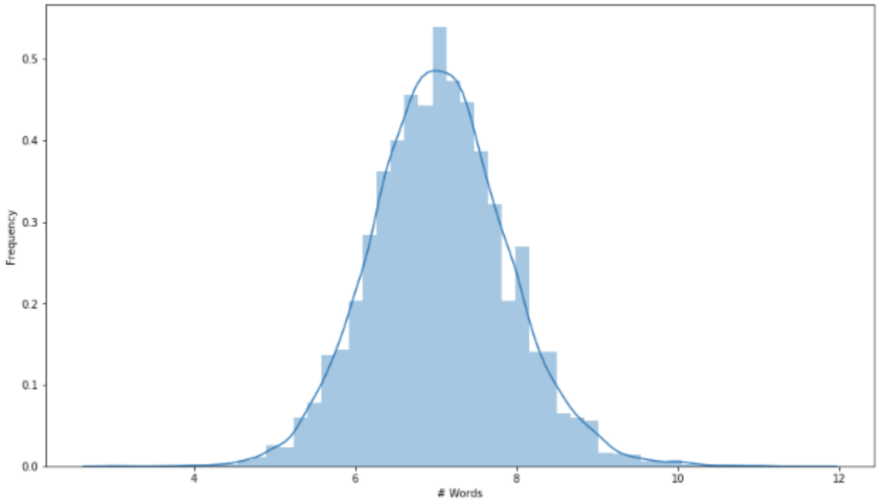


Figure 2. The distribution of the length of claims in the data after cleaning

Next we want to study whether we can visualize words that tend to mostly be in false claims. We can learn of such words from Figure 3. It looks like claims about *Hillary* Clinton, *Muslims*, *Barack Obama* and the *Democrats* tend to be false. A similar figure can be produced for words that tend to mostly be in false claims, however, it has been left out due to lack of space.

It is our aim to train an ML model to be able to identify such words or combinations of them, and to predict the truthfulness of a claim given the words that it is contained. Note that the aim is not to build a model which memorizes every claim that it has seen previously and then to classify an instance by recalling the label it had memorized for it. That would require a lot of data, and an abundance of memory to be able to 'remember' each and every claim. However, the model should just make a prediction based on the trends of the data. Those trends include, as we have discussed here, how likely a claim is to be false when it contains some specific word(s). From now on, we will refer to the words or their combinations as features.

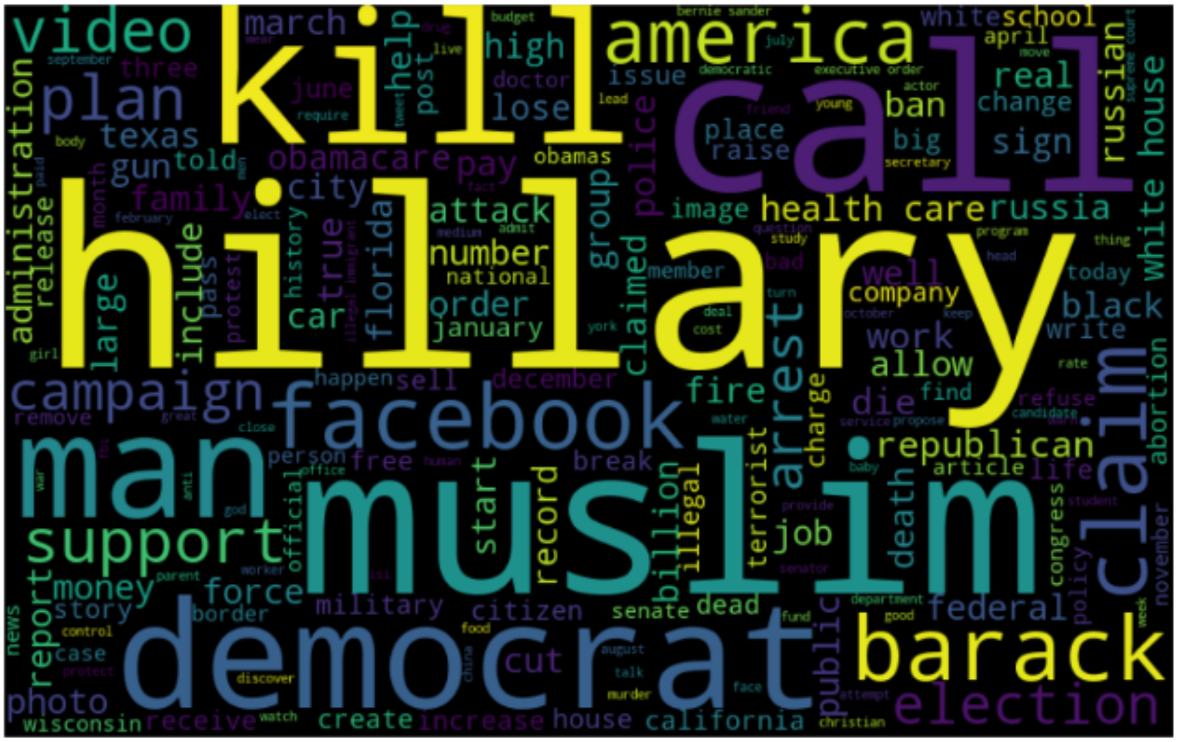


Figure 3. Word cloud of words which tend to be in false claims

3. Machine Learning Model

7 ML models were tested and tuned in the research using 3 methods of encoding the data numerically. 3 of these models had very similar performances. We avoid the technical details in this section, and only discuss the main findings of analyzing these models after training them on the data.

The 3 best models were Logistic Regression (LR), Linear Support Vector Machine (SVM), and Random Forest classifiers which all achieved 59% average accuracy (with 10-fold cross validation). The models are able to correctly classify 1 in every 2 claims they encounter.

3.1. Random Forests

Due to being an ensemble, random forests (RF) are usually black boxes when it comes to interpreting them. However, we can assess the features that the model deemed most important. The features are visualized in Figure 4 with higher feature weights signifying more importance.

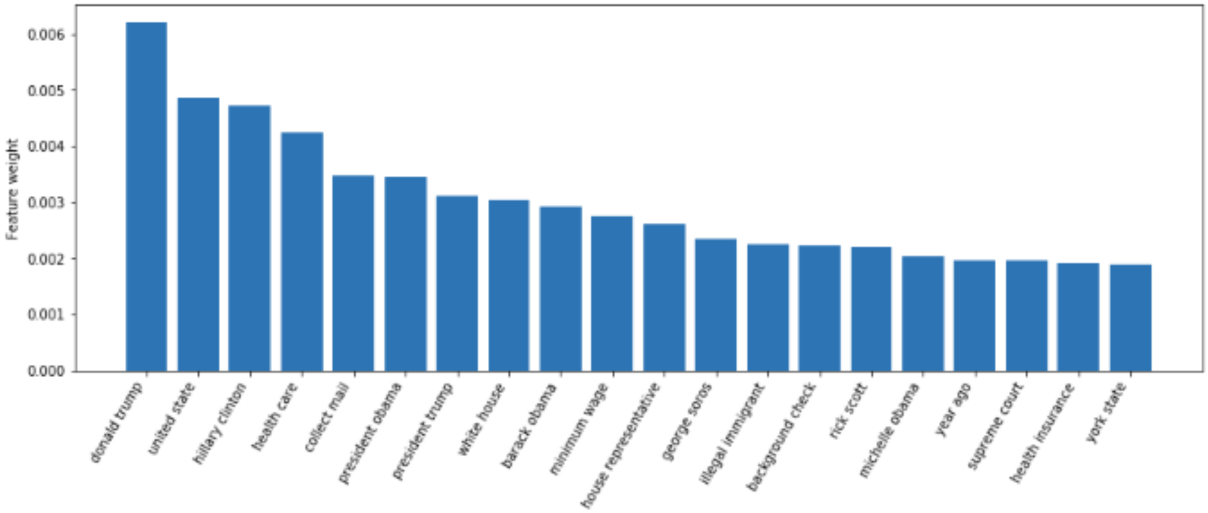


Figure 4. Important Features according to RF classifier

The importance of a feature according to a RF is a measure of how frequent it was in the corpus. We can learn about the most frequently discussed topics from Figure 4.

For example, we can narrate the following story from the data: *Health care,* specifically *health insurances,* as well as *illegal immigration* and the need for *background check*s were two of the most discussed topics during the *United States* election, with *Donald Trump, Hillary Clinton* and *Barack Obama* being the most discussed politicians. All the italicized words were extracted from the figure.

3.2. Logistic Regression

LR assigns weights to each of the features. The higher the magnitude of a weight that more it factored in classifying a claim. To elaborate on this, consider the figure below. The features under study are single words, and the features shown are the ones having the largest and smallest weights. The top figure shows how much a feature factors into determining if a claim is false, and the bottom figure shows how much a feature factors into determining if a claim is true.

For example, a claim about the Ukrainian billionaire, George *Soros*, is very likely to be false (2nd from left in top figure), and very unlikely to be true (12th from the right in bottom figure). A claim is also very likely to be true if a *photograph* (1st from left in bottom figure) is mentioned.

We use these two figures to verify some of our preliminary visualization observations. The blue colored features of the top figure show that claims containing *Muslims* and *Democrats* tend to be false. The red colored features of the bottom figure shows that claims containing *Hillary Clinton* tend to be less likely to be true.

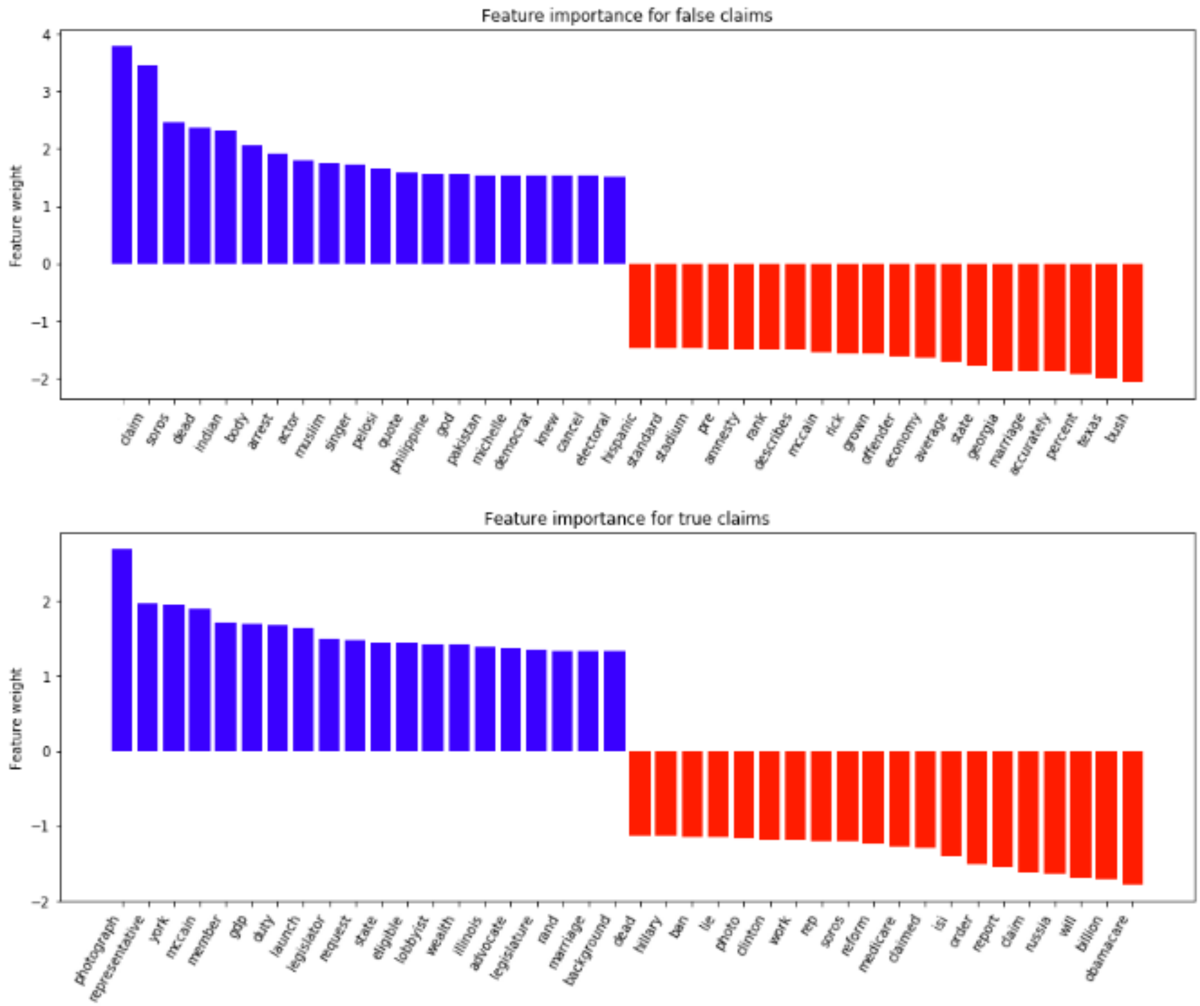


Figure 5. Weights of features (1 word) for the LR classifier

4. Improvements

On the left of each pie chart in Figure 6 is the true label according to the data. Each pie chart shows the proportion of label classifications made by the LR model. Observe that the model accuracy for false/0 claims is highest, given the size of the blue portion of the pie. The LR model also performs well on the partly-true/1 claims. The model, however, hardly makes any correct predictions for true/2 claims.

This goes back to our observation in the preliminary data analysis section when we mentioned that due to the lack of true/2 labelled claims, the model might not be able to train well enough to get a high accuracy in predicting true labels.

To improve the accuracy of the model, it is advised that there be an equal proportion of false, true and partly true claims in the training data.

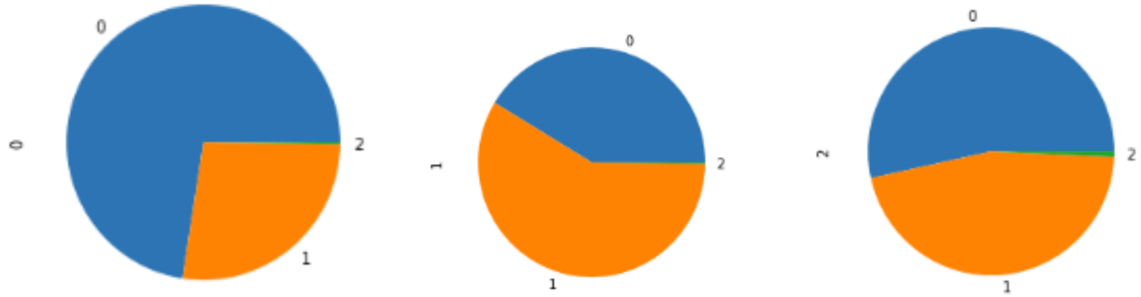


Figure 6. Pie charts showing the proportion of classification labels predicted for each label, left: for False, center: for Partly True, right: for True. Colors: Blue(0): False, Orange(1): Partly True, Green(2): True.

5. Discussion

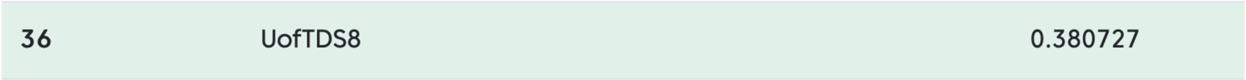
Given a dataset with much more false claims than true claims, we saw that an ML model would be very wary of classifying a claim as true. Interestingly, in a World that has more lies than truths, we, human beings, can also grow to be very wary of believing any claims as well.

We use the features deemed most important by a Random Forest model to get an understanding of the most discussed topics in the text corpus. We see that the corpus was mostly extracted from the period of the last US elections. This is as expected since a lot of the fact checked data is available from that period, when many initiatives were active with regards to fact checking the claims of the candidates and news.

We then use the weights of a Logistic Regression model to understand which entities are likely to be targeted by false claims. We saw that claims containing Hillary Clinton, Barack Obama, and the Democrats were associated with much falsehood. From experience, one remembers that many of the right wing news outlets were very active in targeting false accusations towards these entities during the elections. The elections period also marked the rise of some extremist news outlets such as Breitbart. Muslims, being victims of Islamophobia in the US, and actors and singers, being the meal for Paparazzi, have also been target to many false claims in the data. These findings, if informed to the public, can help the public be wary of claims targeted at these entities.

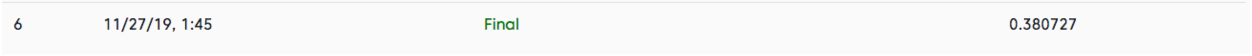
The results of the model accuracy as returned by DataCup is 0.380727. The name of our group in the competition is UofTDS8 (see print screens).

Ranking, Name, Score



36	UofTDS8	0.380727

#, Date, Status, Score



6	11/27/19, 1:45	Final	0.380727