Detecting Fake News Tweets from Twitter

Sheryl Mathias
sheryl14@umd.edu

Namratha Jagadeesh
namratha@umd.edu

College of Information Studies
University of Maryland, College Park
Maryland, USA

Abstract

Social media has increasingly become central to the way billions of people experience news and events, often bypassing journalists—the traditional gatekeepers of breaking news. Events in the real world create a corresponding spur of posts (tweets) on Twitter. This casts a lot of importance to the credibility of information found on social media platforms like Twitter. We have used various supervised learning methods like Logistic Regression, Naïve Bayes, Decision Trees and Support Vector Machines on the data to differentiate tweets between real and fake news. For our machine learning models, we have used tweet and user features as our predictors. We achieved an accuracy of 74% using Random Forest classifier and 68% accuracy using Logistic Regression. However, we believe that analyzing user accounts would definitely increase the precision of our models.

Introduction

Twitter is a micro-blogging service, which has gained popularity as one of the prominent news source and information dissemination agent over the last few years. Each post on Twitter is characterized by two main components: the tweet (content and associated metadata) and the user (source) who posted the tweet. Rumors/fake news posted on twitter during real world events can result in damage, chaos and monetary loss. Today, online social media plays a vital role during real world events such as earthquakes, hurricanes, elections and social movements.

The aim of this paper is to classify fake and real news by using various features of tweets such as text mining of tweet content along with tweet and user based features. In this paper, we want to answer our main research question which is: Can we use various user and tweet features to distinguish between fake and real news? To classify a tweet as fake or real, we will be using supervised machine learning classification techniques. The year 2017 has witnessed major natural disasters such as hurricane Harvey, Irma, Maria, California Wildfires and shootings in Las Vegas and Texas. We have considered 2 major events for our project: Hurricane Harvey and Las Vegas shooting.

Hurricane Harvey: Hurricane Harvey was the costliest tropical cyclone on record, inflicting nearly $200 billion (2017 USD) in damage primarily from widespread flooding in the Houston metropolitan area. This was a category 3 hurricane and occurred between August 17, 2017 and September 3, 2017.

Las Vegas Shooting: On the night of October 1, 2017, a gunman opened fire on a crowd of concertgoers at the Route 91 Harvest music festival on the Las Vegas Strip in Nevada, leaving 58 people dead and 546 injured. Between 10:05 and 10:15 p.m. PDT, 64-year-old Stephen Paddock of Mesquite, Nevada, fired more than 1,100 rounds from his suite on the 32nd floor of the nearby Mandalay Bay hotel.

Related Work

Gupta et al. have highlighted the role of Twitter during Hurricane Sandy (2012) to spread fake images about the disaster. In this paper, the authors have used classification models to distinguish fake images from real images of Hurricane Sandy by using Twitter specific features like the content of the tweet and user details. They used a *Compute_overlap* algorithm and found an overlap of 11% between the retweet and followers graphs for users who tweeted fake images of Sandy. The Decision Tree classifier achieved 97% accuracy in predicting fake images from real. They also found that tweet based features are very effective in distinguishing fake images from real, while the performance of user based features are very poor. [1] Gupta et al. analyze Twitter for content generated during the event of Boston Marathon Blasts to understand what factors influenced in malicious content and profiles becoming viral. They have answered one of the most important research questions to understand if impact of users who propagate fake content be used to estimate how viral the content would become in future by using linear regression.

To understand roles of user attributes in fake content identification they have calculated the overall impact of a user as a linear combination of social reputation, global engagement, topic engagement, likability and credibility. They have then used this calculated *Impact* of all previously active users to predict how many users will get activated in the next time segment. The results of the regression analysis can be used to predict how viral the fake content would become in future based on attributes of the users currently propagating fake content. [2]

According to Gupta A. and Kumaraguru P. linear logistic regression analysis on various Twitter based (content and user) features indicated that the most prominent content based features were number of unique characters, swear words, pronouns and emoticons in a tweet; and user based features were the number of followers and length of username. They also showed that automated algorithms using supervised machine learning and relevance feedback approach based on Twitter features can be effectively used in assessing credibility of information in tweets. [3] Researchers have used the following variables as predictive variables for determining fake news:

1. Number of original tweets, retweets, replies
2. Average length of original tweets, retweets, replies
3. Number of words in original tweets, retweets, replies

These words combined with linguistics helped them uncover words and phrases which indicate whether an event will be perceived as highly credible or less credible. Developing a theory driven, parsimonious model working on millions of tweets corresponding to thousands of events and their corresponding credibility annotations, they unfold ways in which social media text carry signals of information credibility. [4] According to Kouloumpis et al. (2011) explain that part-of-speech features (i.e., count of number of verbs, adverbs, adjectives, nouns and any other parts of speech) may not be as useful as the microblogging features. (i.e., the presence of intensifiers and positive/negative/neutral emoticons and abbreviations) The authors have concluded that using n-grams together with the lexicon features and microblogging features have been useful for Twitter Sentiment Analysis. [8] According to Proposed Approach for Sarcasm Detection in Twitter the use of TextBlob package to determine the polarity of a tweet was very efficient and reliable. The steps include tokenization, part of speech tagging and parsing in Python [5].

Dataset

We collected tweets for our dataset based on events such as the Las Vegas shooting which took place on October 1, 2017 and Hurricane Harvey that occurred in Houston, Texas from August 17, 2017to September 3, 2017using Twitter's streaming API. We have a total of 250 tweets for the Las Vegas mass shooting and Hurricane Harvey events. The figure below represents the distribution of fake and real news in our dataset. We had a total number of 94 fake and 156 real tweets in our dataset. We have considered only the original tweets, which means retweets are not included and all the tweets written in English. There were a total of 21 variables/ features which were mostly of categorical and numerical types. The last column represented the outcome variable which is the Final Label for the tweet. By extracting 13 features out of the 21 final features and labelling the 250 tweets from 2 different events we were able to create our final dataset.

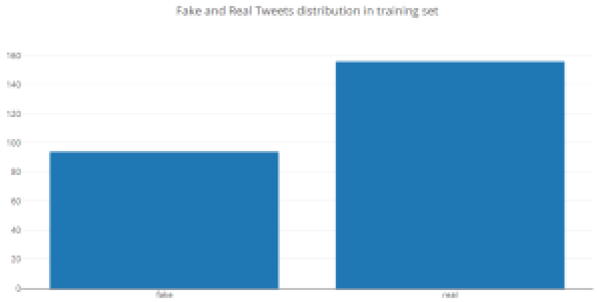

Figure 1

Exploratory Analysis

We analyzed individual features for both fake and real tweets to understand if any of these features are different for both types of labels. Using ggplot2 package, the following 2 bar charts were created. Figure 2 represents the sum of the age of the user's account who have tweeted on the Y axis and the 2 main groups: fake and real tweets on the X axis. Likewise, figures 3 and 4 represent the sum of the number of friends and statuses for both fake and real tweets respectively. It can be seen from figure 2 that the account of users who post real tweets on Twitter are older than those who post fake tweets. Although, the difference is only 300 days. Figure 3 explains a huge difference in the sum of friends for these users, where users who post fake tweets have as less as 85,000 friends in total while those who tweet real news have as large as 165,000 friends. Figure 4 explains the difference in the number of statuses a fake and real users tweet about events. It can be seen that fake users have a greater count of statuses compared to their counterparts. (Real users)

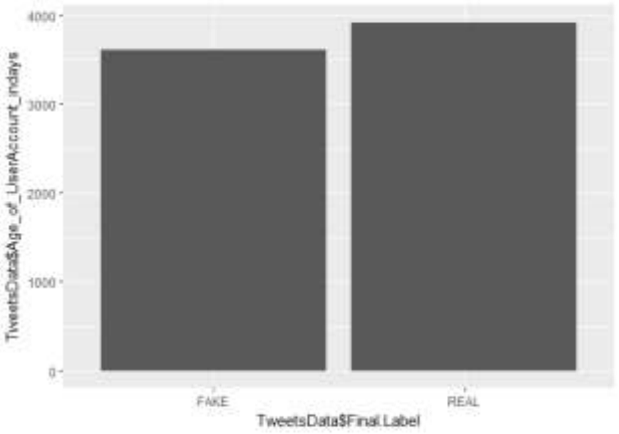

Figure 2

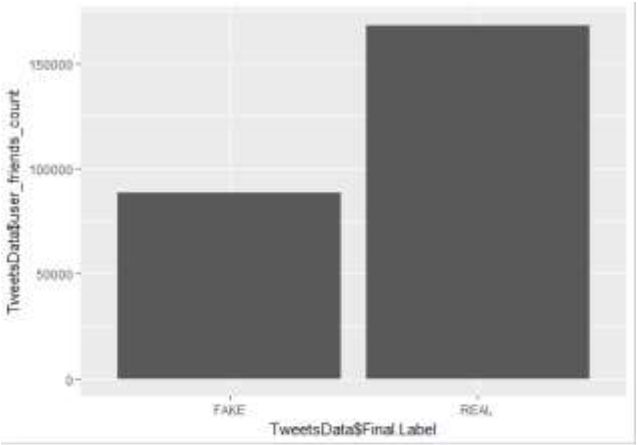

Figure 3

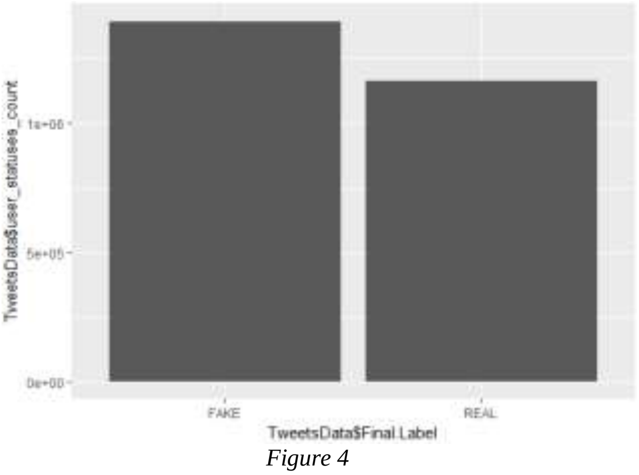

Figure 4

Methodology

For creating the dataset, we used search parameters such as ‘Las Vegas shooting’, ‘Mandalay Bay’, ‘Stephen Paddock’ and ‘Hurricane Harvey’ related to these events to extract the tweets. Once we extracted the tweets we labelled each tweet manually to classify them as real or fake news by referring to websites such as YouTube and Snopes.com as well as the Internet to get facts. There were two annotators for labelling the tweets. Both of us labelled the tweets separately and then cross verified our labels and kept only those tweets for which our labelling matched. We ensured to retain only original tweets and not any retweets as it would cause redundancy in our dataset.

We extracted many features for determining real and fake news. These features can be identified as user and tweet features. Using nltk, regex and textblob packages in Python we extracted the following features:

- Number of Hashtags
- Number of Question Marks
- Number of Mentions
- Number of Exclamation marks
- Number of URLs in tweet
- Polarity of the tweet – Positive, Negative and Neutral
- Number of First Order Pronouns
- Number of Second Order Pronouns
- Number of Third Order Pronouns

The text of the tweets were tokenized using NLTK package. We then removed stopwords from the tweets. We used regex package in Python to extract symbols such as ‘#, @, ?, !, ’ to count the number of hashtags, mentions, question marks and exclamation marks respectively. Later, we cleaned the text of the tweets by removing URLs and punctuation marks using regex package and then counted the occurrence of colon symbols and the number of words in the tweet.

After removing all the unwanted characters and words we had the cleaned text with us. This text was analyzed using the TextBlob package to determine the polarity of the tweet. Internally, this package has a corpus of negative and positive words. The text data is analyzed to see whether the words match with the words in the corpus and based on that the polarity is assigned. The context of use of words in also taken into consideration while assigning the Polarity to a text. This package is more efficient that comparing each word in our tweet to a dictionary of positive and negative words.

The other derived tweet features are:

- Ratio of number of statuses/followers
- Ratio of number of friends/followers
- Source of the tweet
- Age of the user account

We categorized the source of the tweet into 5 categories namely: Mobile, browser, news channel, Facebook and Others to classify them as discrete sources. The age of the user account was calculated in days. We considered user and tweet features which are as follows:

- User Followers count
- User Friends count
- User status count
- User Verified
- Favorite count
- User's profile contains a URL
- Retweet count
- Length of the tweet

These tweets were extracted in CSV format with each column representing a feature of the tweet and the final column which forms the predictor. (Final Label of the tweet) We divided the dataset into training and testing datasets by splitting the dataset in the ratio of 80:20 for applying supervised machine learning classifiers. We use stratified sampling technique to ensure the both training and testing datasets maintain the same proportion of real and fake tweets.

Results

The classifier models were trained on the dataset which contains 250 tweets with a distribution of 94 fake tweets and 156 real tweets. The results in the Table 1 gives the model performance for the mentioned classifiers. Random Forest model with the above mentioned features gave us the best accuracy of 74% and precision of 58%. The second best model was the Decision Tree model after applying 10 fold cross validation which gave us an accuracy of 67.6% and a precision of 58.2%. The third model was the Logistic Regression model which gave us an accuracy of 68% but with a precision of only 50%.

The right part of figure 5 shows the top 10 variables that have the maximum importance in the Random forest model. GINI importance measures the average gain of purity by splits of a given variable. The importance of variables is related to the ability of each variable to classify the data appropriately at each tree split. As seen in figure 5, the most important variable in the Random Forest model is the ratio of friends/followers count and the least important variable is polarity.

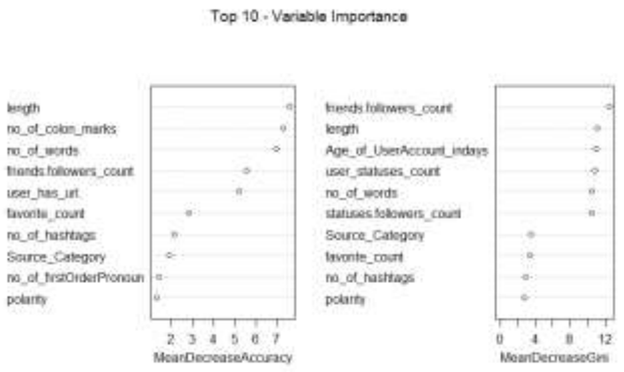

Figure 5

Our final step was to evaluate the model's performance on a dataset with a distribution of real and fake tweets in the ratio of 90:10 to replicate a real time scenario. Our dataset had 16 fake tweets and 156 real tweets for this new ratio. However, the classifier was not able to detect fake news and classify tweets as fake and real. One of the reasons could be the distribution of fake tweets or the classifier had too little data to model for such a distribution. The strange/difficult distribution of samples for the fake class could be a reason for the classifier not being able to detect fake tweets.

The limitations of this model is that it cannot analyze categories or types of words which can be done for both the text of the tweet and user description column. We can use n-gram model in future to better understand the frequency of words used in the text.

Classifier	**P**	**R**	**S**	**Acc**
RF	**58%**	**62.5%**	**79.4%**	**74%**
Logistic Regression	**50%**	**56.2%**	**73.5%**	**68%**
Decision Tree(10 fold CV)	**58.2%**	**48.9%**	**78.8%**	**67.6%**

Table 1

Discussion

With nearly 38% fake tweets in the data set, the Random Forest model achieved an accuracy of 74% while the Logistic Regression model achieved and accuracy of 68%. After using 10 fold cross validation, the Decision Tree Model achieved an accuracy of 67%. As we increased the proportion of real tweets out of the total tweets to 90%, that is, the ratio of real and fake tweets is 90:10, the classifier was unable to classify the tweets as fake and real. This was a major drawback of our model because for our dataset we have only considered all tweets relevant to the news and all tweets that considered user's opinions about the event were omitted. The classifier will need some more extra information such as the description of the user who tweeted about the event. A future scope would be to involve text mining of the user description column as well as use n-gram models and combine it with the above 3 classifiers to achieve a better accuracy and precision. Another limitation is the size of the dataset which contains only 250 tweets. This is a very small sample size. Future scope would be to increase the sample size to at least 1000 tweets.

References

[1] Gupta A., Lamba H., Kumaraguru P. and Joshi A. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. *In Proceedings of the 22nd International Conference on World Wide Web* Pages 729 – 736, New York, NY, USA, 2013. ACM.

[2] Gupta A., Lamba H. and Kumaraguru P. $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. *eCrime Researchers Summit (eCRS), 2013.*

[3] Gupta A. and Kumaraguru P. Credibility Ranking of Tweets during High Impact Events. *In Proceedings of the 1st Workshop on Privacy and Security in Online Social Media* Article No. 2 PSOSM '12, New York, NY, USA, 2012. ACM.

[4] Mitra T., Wright G. and Gilbert E. A Parsimonious Language Model of Social Media Credibility across Disparate Events. *In Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing* Pages 126 – 145, CSCW '17, Portland, Oregon, USA, 2017. ACM

[5] Saha S., Yadav J. and Ranjan P. Proposed Approach for Sarcasm Detection in Twitter. *Working Paper in Indian Journal of Science and Technology, July 2017.*

[6] HaCohen – Kerner Y. and Badash H. Positive and Negative Sentiment Words in a Blog Corpus Written in Hebrew. *In Proceedings of 19th International Conference on Knowledge Based and Intelligent Information and Engineering System, September 4, 2016.*

[7] Fast E., Chen B. and Bernstein M. Empath: Understanding Topic Signals in Large-Scale Text. *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* Pages 4647 – 4657, CHI '16, San Jose, California, USA. May 07 – 12, 2016.

[8] Kouloumpis E., Wilson T. and Johanna M. Twitter Sentiment Analysis: The Good the Bad and the OMG! *In Proceedings of the Fifth International Conference on Weblogs and Social Media,* Barcelona, Catalonia, Spain, July 17 – 21, 2011.

[9] Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau. Sentiment Analysis of Twitter Data. *Working paper in Department of Computer Science, Columbia University, July 2017*

[10] Hana Anber , Akram Salah, A. A. Abd El-Aziz. A Literature Review on Twitter Data Analysis. *In proceedings of the International Journal of Computer and Electrical Engineering*.

[11] Yanchang Zhao. Analyzing Twitter Data with Text Mining and Social Network Analysis. *In Proceedings of the 11-th Australasian Data Mining Conference (AusDM'13), Canberra, Australia.*