Detecting Fake News Tweets from Twitter

CHERLAKOLA ABHINAV REDDY

Post graduation scholar, School of Business, Woxsen University, Telengana, India

SAI NITESH GADIRAJU

Post graduation scholar, School of Business, Woxsen University, Telengana, India

Dr. Samala Nagaraj

Assistant Professor, School of Business, Woxsen University, Telengana, India

Abstract : Online media has progressively obtained integral to the route billions of individuals experience news and occasions, frequently bypassing writers—the conventional guardians of breaking news. Occasions,in reality, make a relating spike of posts (tweets) on Twitter. This projects a great deal of significance on the validity of data found via online media stages like Twitter. We have utilized different managed learning techniques like Naïve Bayes, Decision Trees, and Support Vector Machines on the information to separate tweets among genuine and counterfeit news. For our AI models, we have utilized tweet and client highlights as our indicators. We accomplished a precision of 88% utilizing the Random Forest classifier and 88% utilizing the Decision tree. Notwithstanding, we accept that breaking down client records would build the accuracy of our models.

Introduction

Twitter is miniature writing for a blog administration, which has picked up prevalence as one of the unmistakable news source and data spread specialist in the course of the most recent couple of years. Each post on Twitter is described by two fundamental segments: the tweet (content and related metadata) and the client (source) who posted the tweet. Bits of gossip/counterfeit news posted on Twitter during genuine occasions can bring about harm, disarray and money related misfortune. Today, online web-based media assumes a crucial job during true occasions, for example, tremors, storms, races and social developments. The point of this paper is to group phoney and genuine news by utilizing different highlights of tweets, for example, text mining of tweet content alongside tweet and client-based highlights. In this paper, we need to respond to our principle research question which is: Can we utilize different client and tweet highlights to recognize phoney and genuine news? To order a tweet as phoney or genuine, we will utilize regulated AI grouping methods. The year 2020 has seen serious fiascos COVID-19. We have thought about this as a significant occasion for our undertaking.

Related Work

Gupta et al. have featured the job of Twitter during Hurricane Sandy (2012) to spread phoney pictures about the catastrophe. In this paper, the creators have utilized characterization models to recognize counterfeit pictures from genuine pictures of Hurricane Sandy by utilizing Twitter explicit highlights like the substance of the tweet and client subtleties. They utilized a Compute_overlap calculation and found a cover of 11% between the retweet and adherents diagrams for clients who tweeted counterfeit pictures of Sandy. The Decision Tree classifier accomplished 97% exactness in foreseeing counterfeit pictures from genuine. They additionally found that tweet based highlights are viable in separating counterfeit pictures from genuine, while the exhibition of client-based highlights is extremely poor. [1] Gupta et al. examine Twitter for content produced during the occasion of Boston Marathon Blasts to comprehend what elements affected malignant substance and profiles getting viral. They have addressed quite possibly the main exploration inquiries to comprehend if the effect of clients who spread phoney substance be utilized to gauge how popular the substance would get in future by utilizing straight relapse.

To comprehend jobs of client ascribes in phoney substance recognizable proof they have determined the general effect of a client as a direct blend of social standing, worldwide commitment, point commitment, amiability and believability. They have then utilized this determined Impact of all already dynamic clients to foresee the number of clients that will get actuated in a whenever fragment. The aftereffects of the relapse examination can be utilized to foresee how popular the phoney substance would get in the future dependent on properties of the clients right now proliferating counterfeit substance. [2] According to Gupta A. also, Kumaraguru P. direct strategic relapse investigation on different Twitter-based (substance and client) highlights demonstrated that the most noticeable substance-based highlights were several exceptional characters, swear words, pronouns and emojis in a tweet; and client-based highlights were the number of supporters and length of username. They likewise demonstrated that mechanized calculations utilizing directed AI and an importance criticism approach dependent on Twitter highlights can be adequately utilized in surveying the validity of data in tweets. [3] Researchers have utilized the accompanying factors as prescient factors for deciding phoney news: 1. Several unique tweets, retweets, answers 2. The normal length of unique tweets, retweets, answers 3. Several words in unique tweets, retweets, answers These words joined with etymology assisted them with revealing words and expressions which demonstrate whether an occasion will be seen as profoundly dependable or less sound. Building up a hypothesis-driven, tight-fisted model chipping away at a great many tweets comparing to a large number of occasions and their relating validity explanations, they unfurl manners by which web-based media text convey signs of data believability. [4] According to Kouloumpis et al. (2011) clarify that grammatical form highlights (i.e., check of several action words, modifiers, descriptors, things and some other grammatical forms) may not be as valuable as the microblogging highlights. (i.e., the presence of intensifiers and positive/negative/impartial emojis and shortened forms) The creators have presumed that utilizing n-grams along with the vocabulary highlights and microblogging highlights have been helpful for Twitter Sentiment Analysis. [8] According to Proposed Approach for Sarcasm Detection in Twitter, the utilization of the Text Blob bundle to decide the extremity of a tweet was effective and solid. The means incorporate tokenization, grammatical feature labelling and parsing in Python [5]

Dataset

We gathered tweets for our dataset dependent on the occasion of COVID-19 utilizing Twitter's streaming API. We have an aggregate of 100 tweets. The figure underneath speaks to the dispersion of phoney and genuine news in our dataset. We had an all outnumber of 94 phoneys and 156 genuine tweets in our dataset. We have thought about just the first tweets, which implies retweets are excluded and all the tweets written in English. There was an aggregate of 21 factors/highlights which were generally of unmitigated and mathematical sorts. The last section spoke to the result variable which is the Final Label for the tweet. By extricating 13 highlights out of the 21 last highlights and marking the 100 tweets from 2 distinct occasions we had the option to make our last dataset.

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 the ggplot2 package, the following 2 bar charts were created. Figure 1 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 2 and 3 represent the sum of the number of friends and statuses for both fake and real tweets respectively.

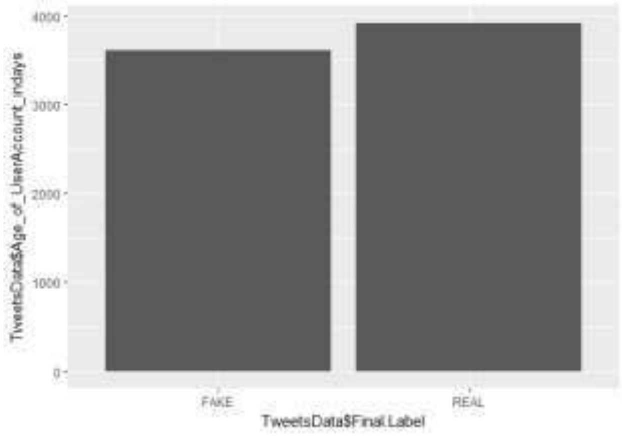

Figure 1

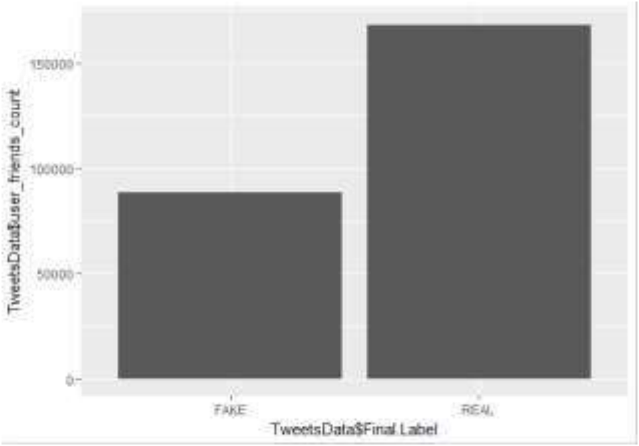

Figure 2

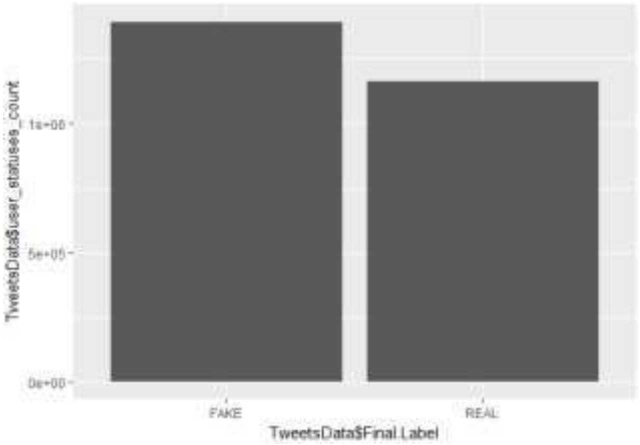

Figure 3

Methodology

For creating the dataset, we used search parameters such as “covid-19”, ‘Carona positive to extract the tweets. Once we extracted the tweets we labelled each tweet manually to classify them as real or fake news. 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 text blob 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 a tweet
- The 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 was tokenized using the NLTK package. We then removed stopwords from the tweets. We used the 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 the 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 the use of words is also taken into consideration while assigning the Polarity to a text. This package is more efficient than comparing each word in our tweet to a dictionary of positive and negative words.

The other derived tweet features are:

- The ratio of number of statuses/followers
- The 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 a user and tweet features which are as follows:

- User Followers count
- User Friends count
- User status count
- User Verified
- Favourite count
- The 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 a stratified sampling technique to ensure both training and testing datasets maintain the same proportion of real and fake tweets.

Results

The classifier models were prepared on the dataset which contains 100 tweets with an appropriation of 23 phoney tweets and 77 genuine tweets. The outcomes give the model execution for the referenced classifiers. The arbitrary Forest model with the previously mentioned highlights gave us the best exactness of 74% and accuracy of 58%. The subsequent best model was the Decision Tree model in the wake of applying 10 crease cross approval which gave us an exactness of 67.6% and an accuracy of 58.2%. The third model was the Logistic Regression model which gave us an exactness of 68% however with an accuracy of just half. The correct piece of figure 5 shows the best 10 factors that have the greatest significance in the Random woodland model. GINI significance gauges the normal increase of virtue by parts of a given variable. The significance of factors is identified with the capacity of every factor to order the information fittingly at each tree split. As found in figure 5, the main variable in the Random Forest model is the proportion of companions/supporters tally and the most un-significant variable is extremity. Figure 5 Our last advance was to assess the model's exhibition on a dataset with a dispersion of genuine and phoney tweets in the proportion of 90:10 to repeat an ongoing situation. Our dataset had 16 phoney tweets and 156 genuine tweets for this new proportion. In any case, the classifier couldn't identify counterfeit news and group tweets as phoney and genuine. One reason could be the dissemination of phoney tweets or the classifier had too little information to demonstrate for such a dispersion. The peculiar/troublesome dispersion of tests for the phoney class could be an explanation behind the classifier not having the option to identify counterfeit tweets. The impediments of this model are that it can't examine classes or sorts of words which should be possible for both the content of the tweet and client depiction section. We can utilize the n-gram model in future to more likely comprehend the recurrence of words utilized in the content.

Discussion

With almost 38% phoney tweets in the informational index, the Random Forest model accomplished a precision of 74% while the Logistic Regression model accomplished an exactness of 68%. In the wake of utilizing 10 overlay cross approval, the Decision Tree Model accomplished an exactness of 67%. As we expanded the extent of genuine tweets out of the complete tweets to 90%, that is, the proportion of genuine and phoney tweets is 90:10, the classifier couldn't group the tweets as phoney and genuine. This was a significant downside of our model because for our dataset we have just viewed all tweets about the news and all tweets that considered the client's conclusions about the occasion were discarded. The classifier will require some more additional data, for example, the depiction of the client who tweeted about the occasion. A future extension is to include text mining of the client depiction segment just as use n-gram models and consolidates it with the over 3 classifiers to accomplish a superior exactness and accuracy. Another constraint is the size of the dataset which contains just 100 tweets. This is a minuscule example size. The future degree is to increment the example size to at any rate 1000 tweets.

References

[1]Faking Sandy | Proceedings of the 22nd International Conference on World Wide Web. (2021). Retrieved 4 March 2021, from https://dl.acm.org/doi/10.1145/2487788.2488033

[2] $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. (2021). Retrieved 4 March 2021, from https://ieeexplore.ieee.org/document/680

[3]Credibility ranking of tweets during high impact events | Proceedings of the 1st Workshop on Privacy and Security in Online Social Media. (2021). Retrieved 4 March 2021, from https://dl.acm.org/doi/10.1145/2185354.2

[4]Mass Participation During Emergency Response | Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. (2021). Retrieved 4 March 2021, from https://dl.acm.org/doi/10.1145/2998181.299

[5] Saha, S., Yadav, J., & Ranjan, P. (2017). Proposed Approach for Sarcasm Detection in Twitter. *Undefined*. Retrieved from https://www.semanticscholar.org/paper/Proposed-Approach-for-Sarcasm-Detection-in-Twitter-Saha-Yadav/1dd26b5fa90f8cce27315baf619b268

[6] Positive and Negative Sentiment Words in a Blog Corpus Written in Hebrew. (2021). Retrieved 4 March 2021, from https://www.researchgate.net/publication/307913780_Positive_and_Negative_Sentiment_Words_in_a_Blog_Corpus_Written_in_Hebrew

[7] How to identify positive, negative, and neutral polarities in subjective sentences?. (2021). Retrieved 4 March 2021, from https://www.researchgate.net/post/How-to-identify-positive-negative-and-neutral-polarities-in-subjective-sentences

[8] Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG!. *AAAI Press*, 538-541. Retrieved from https://www.research.ed.ac.uk/portal/en/publications/twitter-sentiment-analysis-the-good-the-bad-and-the-omg(2c78550c-2683-4aef-a6da-e102714cb1dc).html

[9] (2021). Retrieved 4 March 2021, from https://www.irjet.net/archives/V4/i12/IRJET-V4I12240.pdf

[10] (2021). Retrieved 4 March 2021, from http://www.ijcee.org/vol8/931-IT015.pdf

[11](2021). Retrieved 4 March 2021, from https://crpit.scem.westernsydney.edu.au/confpapers/CRPITV146Zhao.pdf

Mining Conference (AusDM'13), Canberra, Australia.