Team Regexer at Constraints@AAAI 2021 - COVID-19 Fake News Detection in English: Fighting an Infodemic Using Machine Learning and Deep Learning

Navneet Agarwal, Anmol Gupta, Sudhir Attri, and Naman Jain

Indraprastha Institute of Information Technology Delhi, New Delhi, Delhi 110020
I. navneet18348@iiitd.ac.in, II. anmol18329@iiitd.ac.in, III. sudhir18267@iiitd.ac.in, IV. naman18347@iiitd.ac.in

Abstract. This paper is a part of the submission by the team *Regexer* in the competition *Constraint@AAAI2021 - COVID19 Fake News Detection in English* hosted on CodaLab. We proposed and implemented a novel approach that includes a novel feature, it consists of the unique websites which were quoted in a particular tweet. This feature was used along with the conventional TF-IDF word n-grams. We tried various machine learning and deep learning models including Linear Regression, Decision Trees, Support Vector Machines and MultiLayer Perceptron . We also compared the performance of our extracted features with a word embedding CNN model using word2vec. Our final submission was an MLP model which achieved an F1-score of 97.8%.

Keywords: Covid-19, Corona, Infodemic, Fake News Detection

1 Introduction

There has been an exponential rise of the internet and social media in the past few years. Although this rise is beneficial in many ways, we do face some challenges arising out of this perpetually growing online network. One such challenge is the circulation of fake news. Fake news spreads rapidly on social media as most people are unaware of the importance of verifying credibility of the news and the source, and often end up propagating it. Fake news is often political, but such news about a disease can lead to adverse effects. The emergence of COVID-19 pandemic has inevitably led to a sudden spurt of misinformation on the issue. For example, a news that COVID-19 is not worse than flu spread so much that even the US president compared them on his official Twitter account, when in reality COVID-19 is far worse than the flu with long-lasting symptoms. Exposure to such false information at such a large scale can have adverse affects because it influences people in certain ways, the above tweet made people underestimate the seriousness of the virus.

In this paper, we describe the system that we used to try and classify news related to COVID-19 into two categories: fake and real. We developed this system

for the *Constraint@AAAI2021 - COVID19 Fake News Detection in English* task. The news were shared on Twitter, various fact-checking websites were leveraged for labelling the fake tweets and real tweets were taken from verified Twitter handles. Dataset details and description can be found in (Parth et al, 2020, p. 2).

2 Related Work

(Kelly, 2018, p. 2) defines fake news as the deliberate spread of misinformation on social media and it focuses on discussing the reasoning behind the existence of fake news and its impact on social media. The author also proposes methods to detect fake news using Naive Bayes and Support Vector Machine (SVM) classifiers, but the implementation and results are not provided. The paper emphasizes on the methods of detecting fake news, i.e. linguistic cue methods and network analysis methods. Linguistic cue methods utilize differences in patterns of communicative behaviour between fake and real news. These differences arise from the inexperience and lack of domain knowledge of the person creating fake news. Network analysis methods, on the other hand, use the truthfulness of actual statements by building up a knowledge graph with existing facts and deduce whether new news is fake.

In the context of fake news in the times of COVID-19, (Debanjana, Bhardwaj, Samanta, Azad, 2020, p. 2) used 3 different data sources for tweets related to COVID-19 in English, Hindi and Bengali, of which the Bengali data was unsupervised. It was labelled by the authors using the definition - 'any tweet that does not contain a verifiable claim is a malicious or fake tweet' resulting in a total of 800 labelled tweets. Along with text-embedding of tweets, they extracted various hand-crafted features by using information form the twitter statistics such as retweet count and user name, and used some external tools to construct a fact verification score and a bias score. For classification tasks, they used BERT and mBERT models to get sentence embedding which were further classified using SVM, RFC and MLP classifiers, the authors achieved 0.89 F1-score using the mBERT model, but it excluded other hand-crafted features.

(Parth et al, 2020, p. 2) Developed a database of real and fake news tweets from twitter containing news related to COVID-19 which were collected from verified twitter accounts or fact verification sites. They ran 4 classifiers on TF-IDF features to get a benchmark scores of 0.93 F1-score.

A data mining approach to the generalised problem of fake news detection is given by (Kai et al, 2020, p. 2), they aimed to classify news articles taking into account the publisher, content and the social news engagement, i.e. the process of spreading of news through users. Along with linguistic features (n-grams, lexicons, POS tags), they extracted social context features representing proliferation of the news.

3 Methodology and Data

3.1 Dataset Description

The dataset was curated by (Parth et al, 2020, p. 2) for the competition *Constraint@AAAI2021 - COVID19 Fake News Detection in English* hosted on CodaLab. COVID-19 related tweets with fake news were collected from various fact-checking websites such as Politifact and NewsChecker. Real news was collected from verified sources like WHO and ICMR. The collected data was partitioned into training, validation and testing sets for the purpose of the competition keeping the class distribution similar across the sets. The distribution of data is given in *Table 1.*

Category / Set	**Training**	**Validation**	**Testing**
Real	3360	1120	1120
Fake	3060	1020	1020
Total	6420	2140	2140

Table 1. Distribution of data across classes and sets

3.2 Preprocessing

For feature construction, we first made two copies of the dataset and applied different pre-processing steps on them.

In the **dataset copy 1** we did the following pre-processing steps:

- Converted every Unicode emoji present in the into text. For example, 🙂 got converted into slightly_smiling_face and 😎 got converted into smiling_face_with_sunglasses. This helps in including the the unicode emojis as text in the tweets, which we will use as a text feature (n-grams).
- Converted the tweets into lower case.
- Removed all the punctuation marks from the tweets.
- Removed all the URLs from the tweets.
- Removed stop words from the tweets.
- Used Porter Stemmer to stem the words of all the tweets.

In the **dataset copy 2**, we removed everything from the tweet except the URL links. The URL links are shortened automatically by Twitter and they appear as https://t.co/random_string_here.

Tweets / Set	Training	Validation	Testing
Total Tweets	6420	2140	2140
Tweets With URL	4392	1496	1494

Table 2. Distribution of tweets with URLs

3.3 Feature Extraction

– We extracted n-grams for n from 1 to 4, and applied a TF-IDF (Term Frequency – Inverse Document Frequency) transformer on it **(Using the preprocessed dataset part 1 for this feature)**. TF-IDF for a word increases as the frequency of that word increases and it decreases as the number of documents containing that word increases (Parth et al, 2020, p. 2).

– We constructed one feature for each of the unique websites whose web pages were quoted in the tweets. We used these websites as a feature because some of the tweets include URLs like:

1. The Spoof News (https://www.thespoof.com/)

   This is a parody and satire news website. If a tweet quotes a web page on this website, then the information quoted by this tweet is likely to be fake. The following are two examples of such tweets, both of which are labelled as fake.

   (a) *"Politically Correct Woman (Almost) Uses Pandemic as Excuse Not to Reuse Plastic Bag https://t.co/thF8GuNFPe #coronavirus #nashville"*

   The URL in the tweet redirects to https://www.thespoof.com/spoof-news/us/148337/politically-correct-woman-almost-uses-pandemic-as-excuse-not-to-reuse-plastic-bag

   (b) *"Obama Calls Trump's Coronavirus Response A Chaotic Disaster https://t.co/DeDqZEhAsB"*

   Again, the URL redirects to https://www.thespoof.com/spoof-news/us/148271/obama-calls-trump-s-coronavirus-response-a-chaotic-disaster which is a page on The Spoof News.

2. Government of Maharashtra Public Health Department (https://arogya.maharashtra.gov.in)

   This is the official website of Government of Maharashtra Public Health Department. So, the tweets quoting this website are likely to be real. The following are two examples of such tweets, and both of them are

real.

(a) *"As of today there are 23917 active COVID19 cases in Mumbai District Maharastra: https://t.co/6CcqIYgbXL @MoHFW_INDIA @drharshvardhan @DrHVoffice @MantralayaRoom #COVID__19 #Covid_19 #COVID19 #CoronaUpdatesInIndia #Coronavirus #CoronaVirusUpdates #CoronavirusPandemic https://t.co/k6ICon5kZ6"*

The first first URL in the tweet redirects to https://arogya.maharashtra.gov.in/1175/Novel--Corona-Virus.

(b) *"As of today there are 120 active #COVID19 cases in #Yavatmal District #Maharastra: https://t.co/6CcqIYgbXL @MoHFW_INDIA @drharshvardhan @DrHVoffice @CMOMaharashtra #COVID19 #Covid_19 #COVID19 #CoronaUpdates #CoronavirusIndia #CoronaVirusUpdates #coronavirus #COVIDUpdates https://t.co/CWRfeWweAr"*

The first URL in this tweet redirects to https://arogya.maharashtra.gov.in/1175/Novel--Corona-Virus.

There are many more such examples. Some websites often post biased and fake news and some mostly publish accurate and real news. This understanding makes this feature relevant to the fake news detection task.

To construct these features we did the following steps **on the pre-processed dataset part 2**:

- For every shortened URL present in the tweets, we converted it back to the original form, i.e. the URL to which a shortened URL will redirect to. *Table 3* shows two examples.

Tweet ID	Shortened URL	Original URL
195	https://t.co/6Yx9XzSUZT	https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1653254
196	https://t.co/t6s1HdB0MH	https://www.medscape.com/viewarticle/936896?src=soc_tw_200906_mscpedt_news_mdscp_twindemic&faf=1

Table 3. Some conversions from the training set

- We stored all the unique converted URLs from the entire dataset in a single set.

- Most of the URLs in this set were different web pages of a single website. We found 240 unique website whose web pages were mentioned in the tweets. *Table 4* shows some examples of the unique websites.

Some of the unique websites
https://www.axios.com
https://www.azdhs.gov
https://arogya.maharashtra.gov.in
https://www.thespoof.com/

Table 4. Examples of unique websites in the tweets

- We constructed 240 features for these 240 websites. If a URL in a particular tweet redirects to a webpage of a particular website domain, we increase the count of occurrence for that website(feature).

- After doing the above step we applied a TF-IDF transformer on this, to convert the counts into tf-idf form. This is similar to the step we have done for word n-grams.

– **Word Embedding:** The text in the tweets **(preprocessed data part 1)** was used for generating word embedding using the word2vec model provided by gensim's python library. The vector size for each word was kept 200.

3.4 Classification Models Used

– **Decision Tree (DT)** is a tree like structure containing decision rules for classification which have a strong advantage over other algorithms due to its easy interpretability.

– **Logistic Regression (LR)** uses sigmoid function to output a continuous probability of class in binary classification, to remove overfitting inverse regularization parameter was set to 100, sci-kit learn python library was used for implementation.

– **Support Vector Machine Classifier (SVM)** maximises the distance between boundary separating the classes, a kernel trick is used in SVMs to transform data into higher dimensions making it linearly separable. We ue

– The **Multi-layer Perceptron (MLP)** model consists of 2 hidden layers of sizes 32 and 10 with ReLU activation function followed by a batch normalization layer which helps in speeding up the gradient descent algorithm for optimization.

– **Convolutional neural network (CNN)** is very effective in extracting features from series data as it utilise convolution operation with some small-sized filters that are applied to the complete series (sentence) to learn the underlying structure of the sentences. We used a CNN model with word2vec word embedding for the classification.

4 Results

Model / Metric	Accuracy	Precision	Recall	F1-score
LR	0.936	0.936	0.935	0.933
SVM	0.966	**0.966**	0.966	0.966
DT	0.903	0.903	0.903	0.903
MLP	**0.971**	0.965	**0.976**	**0.971**
CNN (word2vec)	0.9196	0.951	0.891	0.920

Table 5. Results obtained on validation set

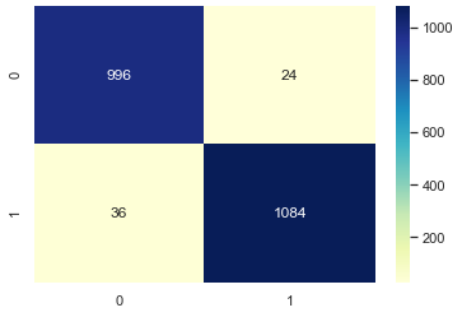

Fig. 1. Heat map for MLP predictions on validation set. 0 is fake and 1 is real.

We used n-gram and the websites features with LR, SVM, DT and MLP. We used word embedding with the CNN model. We observed that for the validation set MLP classifier outperformed every other model using the TF-IDF n-grams and the website features. These two types of features increased the F1-score for all the models by 2-3 % as compared to only TF-IDF n-grams. We also observed that the CNN model which uses word embedding performed the poorest of all. Since the MLP was the best performing model, we used that for predicting the test Labels. To predict the test labels we trained our model by combining the training and the validation set. This helped us achieve an F1-score of 97.8% on the test dataset.

5 Conclusion

We proposed a novel feature for classification of tweets in context of fake news and tested our hypothesis using four different models. We achieved the best performance, an F1-score of 97.8% on the test set, using the MLP classifier. The website features which were extracted using the URLs gave us promising results, thus this feature should be explored further. The word embedding perform very poor as compared to the conventional n-grams approach. This could be because the word embedding captures sequential context which may not be relevant for fake news detection. In future, the content of the quoted URL (web page) and the text of the tweet could be compared.

References

1. Stahl, K., 2018. Fake news detection in social media. California State University Stanislaus, 6.
2. Kar, D., Bhardwaj, M., Samanta, S. and Azad, A.P., 2020. No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection. arXiv preprint arXiv:2010.06906.
3. Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1), pp.22-36.
4. Patwa, P., Sharma, S., PYKL, S., Guptha, V., Kumari, G., Akhtar, M.S., Ekbal, A., Das, A. and Chakraborty, T., 2020. Fighting an Infodemic: COVID-19 Fake News Dataset. arXiv preprint arXiv:2011.03327.