Classification of Tweets as Fake, Facts and Opinions

Aditya Garg
IIIT Delhi
aditya18124@iiitd.ac.in

Tushar
IIIT Delhi
tushar18201@iiitd.ac.in

Bhavesh Khatri
IIIT Delhi
bhavesh18030@iiitd.ac.in

Mayank Joshi
IIIT Delhi
mayank18048@iiitd.ac.in

Mohd Huzaifa
IIIT Delhi
huzaifa18158@iiitd.ac.in

1. INTRODUCTION AND MOTIVATION

Twitter can be considered as one of the contemporary and popular online social networks. The microblogging platform experiences over 500 million tweets daily and around 200 billion tweets per year [10].

Community posts on Twitter often include traffic reports, weather, local news, special events, restaurant reviews and other information useful to locals and visitors alike. Along with factual information, the spread of misinformation is very prevalent on Twitter. Propaganda by fake news is not only a concern in politics nowadays, but it is affecting each and every one. Previous research has shown that misinformation spreads many folds faster than the factual information [13]. Consequently, It has become a necessity to regulate fake tweets on the platform.

It is also important to note that not all tweets can be classified into fact or misinformation, some tweets are just mere opinion of an individual. These opinions are usually subjective expressions that describe people's sentiments or feelings towards entities, events, and properties.

Now, with the rise of AI, more and more things are being automated. Tweets classification is an active research area and it has got its applications in many domains ranging from prediction of stock prices and music sales to election results. It is often observed that individuals often misunderstand opinions as facts. This tandem leads to conflicts.

There were some datasets ([5],[7],[14]) readily available for the fake tweets classification, though the lack of a database for the fact, misinformation and opinion classification make it a difficult and challenging task for researchers to tackle this problem.

This motivated us to make a database for tweet classification (fact, misinformation, opinion) and comparing the performance of various ML/DL algorithms(CNN, RNN, LSTM, etc.) for the classification task. We currently plan to classify tweets on various topics including Covid-19 and Delhi riots. In addition to this, we will try to provide a deterministic source for the fact and misinformation.

The Github repository for our project can be found **here**.

Keywords - Social Media, Tweet classification, Fake news, Opinion classification, Information credibility.

2. RELATED WORK

Figure 1: Sample Tweets

2.1 Literature Review

- Dabas et al. [3] take into account multiple features like title, content and images used in the article to classify into fake or real news. The study has used the knowledge graph to analyse the relations between various entities like location, the person used in the news articles. They combine the content embeddings(using GloVe) and knowledge graph embeddings to get the best information about the content. Finally, they combine all the three features; title vector, content vector and image vector (using activation function) to make one news vector and classify it as real or fake. The proposed model used multi-channel CNN classifier and got F1-score of 0.955 and 0.701 on publicly available

datasets- MFN and FakeNewsAMT respectively.

- Chatterjee et al. [2] proposed a deep learning approach combining BOW (Bag-of-Words) features and manually engineered features (ME) for classifying facts and opinions in Twitter messages. To generate the ME features they extracted prominent information from the data including- grammatical mistakes in the text, presence of URL, type of user, followers of the User etc. The deep learning architecture used consists of two dense layers followed by ReLU activation for ME and BOW features respectively. The output of these layers is fed to a dense merged layer followed by Sigmoid activation. The authors used OpinionFinder[9][15], SentiStrength[12] as some of their baseline models. The proposed DL model got an accuracy of 87.08% with an F1-score of 85.20. The authors also showed that this classification provides useful insights into social media analytics such as consumer complaints and business perspectives.

- A. Gupta and P. Kumaraguru in their work [6] showed that pairwise ranking algorithms can be used to assess the credibility of tweets based on various Twitter features. The NDCG (Normalized Discounted Cumulative Gain) score of 0.37 was achieved on an average for the top 25 tweets using the SVM ranking algorithm.

- Ajao et al. [1] tried to detect relevant features of the fake tweets without the previous knowledge of domain or topic of the tweet. Using a hybrid of CNN and LSTM, they achieved 82% accuracy.

- Singh and Lefeve [11] tried to develop a way to do a sentiment analysis of bilingual tweets. They tried to use unsupervised cross-lingual embeddings to make Hinglish words meaningful and do further analysis. They started by training a basic model on a collection of such mixed tweets and gradually improved upon the techniques to achieve a decent F1 score of 0.635. Finally, they analysed and did a comparative study on model performance with and without the use of cross-lingual embeddings.

- Patwa et al. [8] created the dataset, restricted to Covid19 having labels as either "real" or "fake". The authors used term frequency–inverse document frequency (tf-idf) for feature extraction. They trained several machine learning models, namely Logistic Regression, SVM with linear kernel, Decision Tree and Gradient Boosting. The SVM based classifier performed best among all, with the F1- score of 0.9346.

- Das et al. [4] talked about the role social media plays nowadays and its impact related to the Covid 19 pandemic. For the same reason, they decided to create a fake news detection system that can successfully detect fake news content in the tweets if any. They have used an ensemble model consisting of pre-trained models and successfully achieved the F1-Score of 0.9831. They further achieved SOTA with an F1-score of 0.9883, by incorporating a novel heuristic algorithm based on username handles and link domains in tweets. The methodologies, techniques and analysis they did guided us a lot in classifying the tweets into the three predefined categories : Fake, Real and Opinion.

2.2 Limitations

- Opinion mining is one of the active research areas in the field of social media analysis, under which many brands use sentimental analysis to predict the mindset of people towards their brand, but there are comparatively fewer studies which distinguish opinion, facts and fake tweets.

- Subjectivity detection is also one of the ways for classifying news articles as facts-opinion-fake, which uses lexicon-based methods. But it may not give satisfactory results as tweets are short and informal, containing abbreviations, emoticons and other grammatical errors.

- Although fake tweets detection has gotten attention in recent years, the datasets already available are not sufficient for this task. There is a need to manually annotate the tweets with facts-opinion-fake.

3. COLLECTION AND ANNOTATION

3.1 Guidelines for annotation

We follow a simple guideline during the data collection phase as follows:

- Content is related to the topic COVID-19/Delhi Riots/Union Budget 2020.

- The language of the tweets is English.

We took help of human annotators to obtain the ground truth regarding the categorization of tweets. We filtered the English language tweets using the python's langdetect library. For the purpose of annotation, we developed a GUI in python.
To assess the tweets, we asked the human annotators to select one of the following options for each tweets:

- Fact: If there exist articles online from trusted sources that agree with the tweet.

- Fake: If there exist articles online from trusted sources that disagree with the tweet.

- Opinion: If tweet doesn't contain any claim but consists of opinions of an individual.

- None: If the language is not english/not possible to determine the category of tweet.

Finally, only the tweets that have been labelled same by the two annotators have been included in the dataset; we discarded all tweets for which the annotators have given different labels.

3.2 Fake and Real News

We have used the dataset for CONSTRAINT COVID-19 Fake News Detection in English challenge. The dataset consists of Real news collected from Twitter using verified twitter handles. Whereas Fake claims are collected from various fact-checking websites like Politifact , NewsChecker , Boomlive , etc., and from tools like Google fact-check-explorer and IFCN chatbot.

(a) Fake Tweets

(b) Real Tweets

(c) Opinion Tweets

Figure 2: WordCloud for the Dataset

3.3 Opinions

We use the twint tool to collect potential tweets that are opinions. Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API. We use keywords like COVID19, coronavirus, CovidVaccine, etc. to narrow down our search for tweets. Each tweet is read by a human and is marked as an opinion if it contains opinion/reaction of topics related to COVID-19 (numbers, dates, vaccine progress, government policies, hotspots,etc.).

3.4 Dataset Statistics

Attribute	All	Fake	Real	Opinion
Avg Chars/post	199.210	143.210	217.124	228.426
Avg Words/post	29.084	21.652	31.967	32.516
Unique Words	32909	11941	10780	22816

Table 1: Numeric features of the Dataset

We found that the Average Characters per post and Average Words per post in case of Real and Opinion is more as compared to the Fake tweets. Furthermore, Unique Words for Opinion tweets is more as compared to those of Real and Fake tweets as shown in Table 1.
The dataset is split into train (60%), validation (20%), test (20%). The dataset is class-wise balanced as class has equal proportions of data samples i.e. 6342 or 37.21% of samples are of Opinion class, 5600 or 32.85% of samples are of Real class and 5100 or 29.92% samples are of Fake class as shown in Table 2.

Split	Opinion	Real	Fake	Total
Training	3805	3360	3060	10225
Validation	1269	1120	1020	3409
Testing	1268	1120	1020	3408
Total	6342	5600	5100	17042

Table 2: Data Distribution

To get a better understanding of the frequency of words in Fake, Real and Opinion data. The WordCloud for each of them was plotted using the wordcloud library and is shown in Figure 2.

4. BASELINE MODELS AND RESULTS

Patwa et al.'s [8] paper titled "Fighting an Infodemic: COVID-19 Fake News Dataset" was used for implementing the baseline model of this project.

4.1 Pre-Processing

All the tweets were converted to lower case. The links, Non-Alphanumeric characters, special characters and English Stop words present in the tweets were removed. The English stopwords were taken from the nltk library.

4.2 Feature Extraction

As a part of feature extraction, we used Term frequency-inverse document frequency (tf-idf) and Matrix having counts of string token. Term Frequency Inverse Document Frequency (tf-idf) addresses the significance or meaning of a word to a document in collection of various such documents. Words present in the vast majority of the documents in the collection have lower tf-idf score. This is on the grounds that such words follow back to numerous documents and subsequently, don't help much in getting a best match. Also, the words which are available on different occasions in a similar document will have higher tf-idf score. TfidfTransformer, CountVectorizer feature extractor from the sklearn package are used to calculate this score.

4.3 Machine Learning Algorithms

For this project, we applied several classification models (from the sklearn package) on our dataset which include:-
1. Support Vector Machine (SVM) with Linear Kernel
2. Logistic Regression (LR)
3. Gradient Boosting (GB)
4. Decision Tree (DT)
5. Multi-Layer Perceptron (MLP)
6. Ada Boost
7. Ridge
8. Random Forest (RF)
9. Bagging Classifier
10. K-Neighbours

11. Stochastic Gradient Descent (SGD)

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.850	0.851	0.850	0.849
Gradient Boosting	0.767	0.772	0.767	0.767
Decision Tree	0.739	0.744	0.739	0.740
SVM	0.867	0.866	0.867	0.866
MLP	0.855	0.856	0.855	0.856
Ada Boost	0.713	0.715	0.713	0.714
Ridge	**0.868**	**0.868**	**0.868**	**0.868**
Random Forest	0.789	0.791	0.789	0.787
Bagging	0.772	0.774	0.772	0.771
K-Neighbors	0.753	0.793	0.753	0.756
SGD	0.867	0.866	0.867	0.866

Table 3: ML Models Results on Test Data

Table 3 and Figure 3 shows the results of ML models on the test data. The best F1 score of 86.8% was acheived by the **Ridge** Classifier followed by the SVM and SGD classifiers with 86.7% and 86.7% F1 score respectively.
We observe from Table 3 and Figure 3 that linear models such as Ridge, SVM, SGD etc. and MLP (Neural Network Based) performs much better as compared to the ensemble and tree based models such as Gradient Boosting, Random Forest etc.
The confusion matrix of the Rigde model on the Test data is shown in Figure 4.

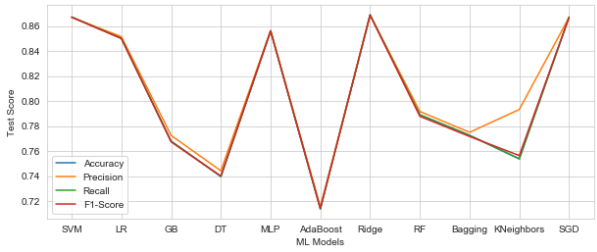

Figure 3: Graph based Visualization of ML Models Results on Test Data

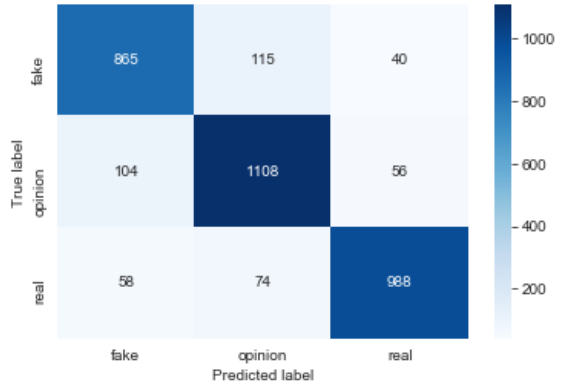

Figure 4: Confusion Matrix of Ridge Model on Test Data

5. PROPOSED METHODS AND FUTURE WORK

1. **Dataset Expansion** : We will try to expand our dataset to various fields including politics, Business and Entertainment.

2. **Deep Learning Model Application** : From Table 3, we observed that Neural network based model MLP Classifier performs quite well in our case. So, based on this inference we will also try to implement deep learning algorithms including CNN, RNN, LSTM etc. We will also try to make use of BERT architecture for better transfer learning in NLP, and its other variations.

3. **Real-Time Fact Checking** : We plan to provide reliable source for facts and misinformation.

6. EXPECTED TIMELINE

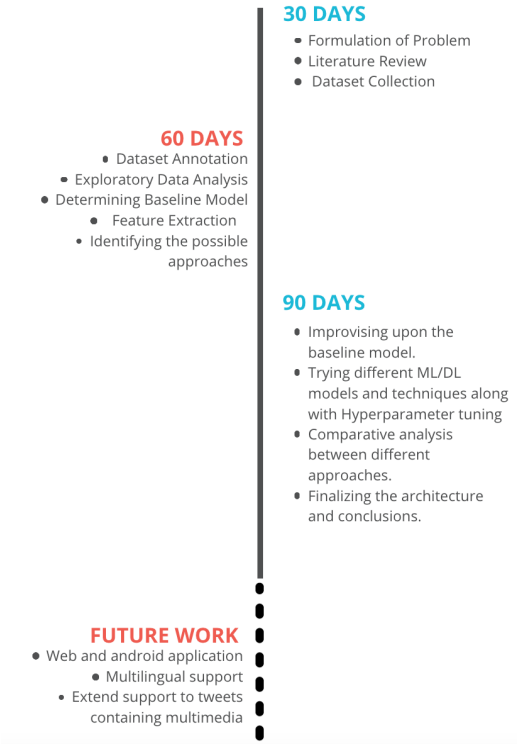

Figure 5: Timeline For the Project

7. REFERENCES

[1] O. Ajao, D. Bhowmik, and S. Zargari. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th International Conference on Social Media and Society*, SMSociety '18, page 226–230, New York, NY, USA, 2018. Association for Computing Machinery.

[2] S. Chatterjee, S. Deng, J. Liu, R. Shan, and W. Jiao. Classifying facts and opinions in twitter messages: a deep learning-based approach. *Journal of Business Analytics*, 1(1):29–39, 2018.

[3] K. Dabas, P. Kumaraguru, T. Chakraborty, and R. R. Shah. Multimodal fake news detection on online social. 2018.

[4] S. D. Das, A. Basak, and S. Dutta. A heuristic-driven ensemble framework for covid-19 fake news detection, 2021.

[5] Edward and Craig. Fake news challenge. 2017.

[6] A. Gupta, P. Kumaraguru, and Indraprastha. Credibility ranking of tweets on events breakingnews. 2011.

[7] T. Mitra and E. Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, 2015.

[8] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty. Fighting an infodemic: Covid-19 fake news dataset, 2021.

[9] E. Riloff, J. Wiebe, and T. Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 25–32, 2003.

[10] D. Sayce. The number of tweets per day in 2020, Dec 2020.

[11] P. Singh and E. Lefever. Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 45–51, 2020.

[12] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[13] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[14] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection, 2017.

[15] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *International conference on intelligent text processing and computational linguistics*, pages 486–497. Springer, 2005.