CSE 472: Social Media Mining
Prof. Huan Liu
Project II - **Project Type 1**

Task: **Fake News Detection**

Submitted by-

Name	E-mail
Arindam Jain	ajain243@asu.edu
Snigdha Reddy	smudired@asu.edu

(1) Abstract

Fake News refers to false, misleading, or unproven information. The amount of misinformation that is spread by word-of-mouth or through social media, memes, edited videos, and false advertisements have rapidly increased after the COVID-19 pandemic. The most contributing factor to this is digital communication. Social media and the press publish fake news to increase readership or as part of psychological warfare. In this project, we perform predictive analytics of the COVID-19 information available online in order to make sure that the public consumes valid and authentic information and doesn't fall into the trap of unethical, misleading information. This project aims at classifying the claim into one of the four categories(false, misleading, true, unproven) by taking a claim of misinformation related to COVID-19 from the given dataset. We developed various machine learning models and the best results were obtained using the BERT model. Root mean square error and accuracy are the important parameters we are focusing on. This study looks into fake news classifications based on psychology and sociological theories, as well as existing data mining techniques, evaluation criteria, and sample datasets, as well as a thorough look at spotting false news on social media. We also explore relevant research ideas, problems, and future research objectives.

(2) Introduction

Even a few words of fake news on the Internet produces large amounts of fake news or false information online for a variety of purposes, including financial, threatening, and political interests. The proliferation of fake news and false information can have serious negative consequences for individuals and society.

(I) Disrupt the news ecosystem's balance of trustworthiness.

(Ii) Intentionally encourage customers to believe biased or incorrect information.
(ii) Alter people's perceptions of and responses to actual communications and facts.

As a result, spotting fake news and misleading information on social media is crucial. The purpose of this project is to classify the labels of a specific fake news dataset using a specific measure (mean square error). In order to reduce classifier errors, it is expected to get the lowest possible score by using this metric. The aim of the project is to categorize a claim of COVID-19 misinformation into one of four categories (the number indicates the label):

•FALSE(0): a deceptive or untruthful claim.
•MISLEADING(1): an assertion that causes people to misinterpret a situation.
•TRUE(2): a valid claim that has been confirmed.
•UNPROVEN(3): a claim that cannot be supported.

Dataset:

Dataset	Columns	Length
Train	Country (mentioned)', 'Review Date', 'Claim', 'Source', 'Label','Fact-checked Article	6384
Test	Country (mentioned)', 'Review Date', 'Claim', 'Source', 'Fact-checked Article	710

It is observed that the dataset is highly imbalanced.

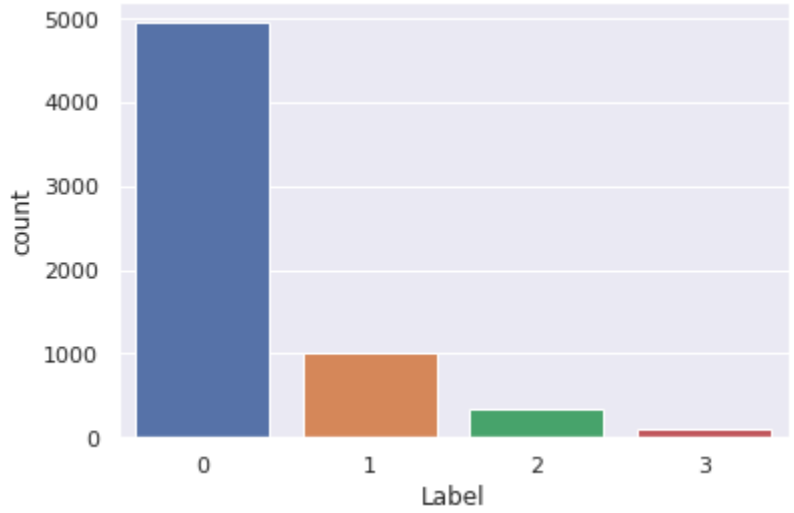

Figure showing the imbalance in the dataset

(3) Related Works

A lot of work is going on in the field of Fake News Detection considering its seriousness and the impact it has on the public. Early in February 2020, the WHO issued a warning that the COVID-19 epidemic had resulted in a large 'infodemic,' or a burst of real and fake news, which contained a lot of misinformation. And this misinformation was spread mainly through social websites, microblogging, advertisements, online news, memes, and edited videos.

Extensive research in Natural Language Processing has been helpful to detect fake news. We came across several research papers working on fake news, and the common problem that is mentioned in all of them is finding accurate automatic news detection. This is because there is no proper dataset that is available that can be used as a benchmark and this is the major problem faced. To deal with this problem, a few authors proposed a new dataset to automatically detect fake news by performing various experiments using machine learning and deep learning techniques.

In the paper [2], a COVID-19 dataset of 5182 fact-checked news articles is presented. This was prepared by collecting the data from ninety-two websites. A BERT-based model to identify COVID-19 fake tweets was proposed by [3] by considering additional features extracted from Twitter. An automated COVID-19 fake news detection pipeline is also developed for checking the various algorithms and to evaluate the best algorithm.

In [4], they have mentioned how sentiment analysis is extremely useful in identifying fake news whether by taking it as a main or complementary feature. They have also discussed the most important requirements while developing the model along with its failures.

(4) Model Description

The only source of detecting whether a news is genuine or fake in a traditional news media is through the news content. Whereas for online news or social media news, a lot of features contribute to detecting the originality of the news. In this project, we extract the most useful features by first scraping the text from the urls in the dataset and then performing exploratory data analysis. The features that we have considered while developing the algorithm are:

- Claim: A given news statement whose validity has to be checked from the news fact checker websites
- Source: The source of the information like Facebook, Twitter, single person, multiple people, Obama, Joe Biden, Kamala Harris,....
- Text: This is the main text that is generated from the url. It contains an explanation for the claim and gives more details about the given story. The relevant information from the website data is taken. If the data is in a language other than English, it is translated into English.

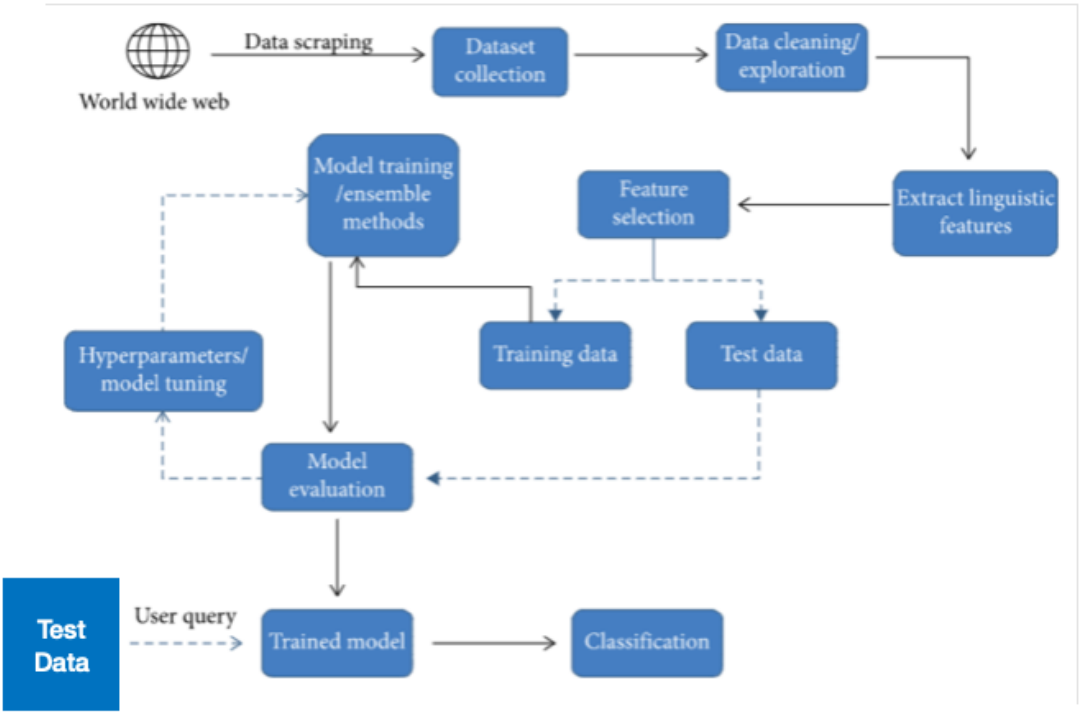

Figure showing the overall flow of the project

We have to preprocess the corpus obtained from the Internet because it has unwanted information like the authors name, URL info, advertisement information, the date of the news publication, copyrights and many more. Then this preprocessed information is passed as an input while training the model. Also, articles that have very little information are discarded as they do not add any value to the claim. All the multicolumn articles are further converted to single-column articles for consistency of format and structure. These operations are performed on both train and test datasets in order to ensure consistency in format and structure.

The next stage is to extract the linguistic characteristics after the necessary qualities have been determined during the data cleaning and exploration phase. The numerical conversion of textual properties for use as input in training models is known as linguistic features. These qualities include the ratio of words that imply good or negative emotions, stop words, punctuation, function words, informal language, and the percentage of adjectives, prepositions, and verbs used in sentences.The raw data is converted into a matrix of TF-IDF features using the sklearn TfidfVectorizer.

After the data visualization, data cleaning, data extraction, and feature extraction steps were successfully completed, we developed our models. After the possible features are extracted, the dataset is divided into 80%, 20%.for training and evaluation respectively. After the classification model is trained successfully, it is tested on the test dataset. Confusion matrix is generated for testing the precision of the test dataset. Next, we evaluated accuracy, MSE, precision, recall, and F1-score of all the models to come up with the best possible model.

First we started off with the basic Logistic Regression algorithm. Then in order to get better results, we have also trained and tested with other machine learning models like Decision Tree, Random Forest, CNN (Convolutional Neural Network), BERT (Bidirectional Encoder Representations from Transformers).

(5) Experiment and Results:

1. The first step that we perform is data visualization. We check for inconsistent data like duplicates, missing values, and incorrect urls.
2. From the urls given in the dataset, we extract the text contained in it by using a web scraping library named “Beautiful Soup”.
3. A few websites contained text in languages other than English(German, Chinese, Portuguese,..). So we translated such data into English.
4. The text extracted from the urls is helpful for generating better features in order to train the model effectively.
5. We then eliminated the stop words like “a”, “the”, “is”, “are” from the claim and extracted text. This was done because the stop words appear multiple times in the data and carry very little information or are of no value.
6. We further cleaned the noise in the data like commas, dots, ids, deleting the suffixes by stemming terms using NLP NLTK libraries.
7. Generated the website name from the entire url in order to analyse which websites provide real and fake information.

	Country (mentioned)	Source	short_description	Label
0	Germany	person	germanya video circulating internet claiming m...	0
1	United States	website	united statesvice president kamala harris admi...	0
2	United States	multiple people	united statesus withhold benefits unvaccinated...	0
3	United States	No data	united statesu s president joe biden said covi...	0
4	Canada	facebook	canadacustomers required provide proof covid -...	0
...	...	...	...	...
6379	Kazakhstan	whatsapp	kazakhstanin kazakhstan rasprostranyayushih in...	2
6380	United States	multiple sources	united statesthe 2020 american presidential el...	1
6381	United States	multiple sources	united statesthe united states refused covid -...	1
6382	Brazil	jair bolsonaro	brazilbolsonaro says coronavirus hysteria bene...	0
6383	China	facebook	chinachina celebrates elimination coronathe vi...	0

Figure showing the data after pre-processing

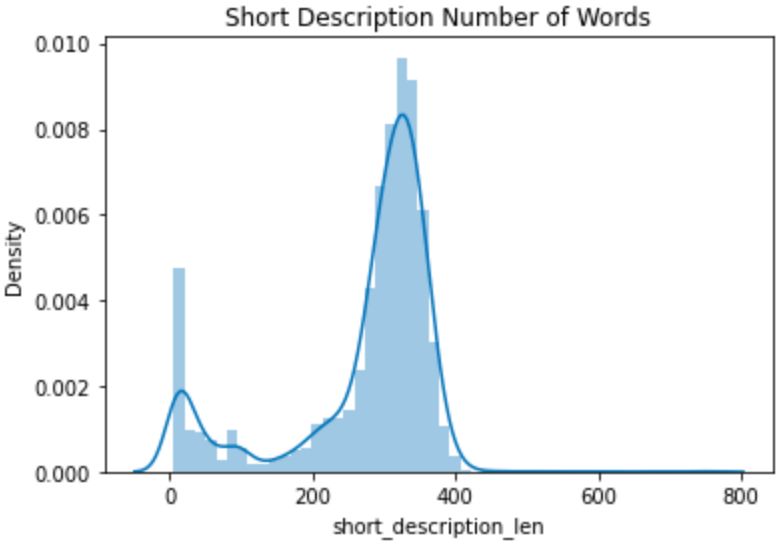

Figure showing the density vs length of the processed text

1. Feature extraction was done after a careful visualisation of lexical features, like word count, number count, and average length of the news article.
2. The raw data is converted into a matrix of TF-IDF features using the sklearn TfidfVectorizer.
3. After the possible features are extracted, the dataset is divided into 80%, 20% for training and evaluation respectively.
4. After the model is trained successfully, it is tested on the test dataset.
5. Various classification models like Logistic Regression, Decision Tree, Random Forest, CNN (Convolutional Neural Network), BERT (Bidirectional Encoder Representations from Transformers), are developed and the results are shown below.
6. Confusion matrix is generated for testing the precision of the test dataset.
7. Evaluated accuracy, MSE, precision, recall, F1-score of all the models to come up with the best possible model.
8. Of all the models we trained, the BERT model gave the highest accuracy and lowest mean square error.
9. The accuracy achieved through the BERT model is 0.91 with a Mean Square Error of 0.48.

Method	Accuracy	MSE (Validation)
Logistic Regression + TF-idf	0.78	0.494
Decision Tree + TF-idf	0.72	0.593
Random Forest + TF-idf	0.77	0.486
CNN + glove 6b 50 (GPU)	0.83	0.495
Bert + Tf-idf (TPU)	0.91	0.48

Observed accuracy and MSE values for various models

Logistic Regression

	precision	recall	f1-score	support
0	0.78	0.99	0.87	1238
1	0.44	0.03	0.06	250
2	0.25	0.01	0.02	78
3	0.00	0.00	0.00	30
accuracy			0.78	1596
macro avg	0.37	0.26	0.24	1596
weighted avg	0.69	0.78	0.69	1596

Performance metrics for Logistic Regression

Decision Tree Classification

	precision	recall	f1-score	support
0	0.82	0.85	0.84	1238
1	0.32	0.28	0.30	250
2	0.22	0.19	0.20	78
3	0.18	0.10	0.13	30
accuracy			0.72	1596
macro avg	0.38	0.36	0.37	1596
weighted avg	0.70	0.72	0.71	1596

Performance metrics for Decision Tree Classification

Random Forest Classifier

	precision	recall	f1-score	support
0	0.78	0.98	0.87	1238
1	0.38	0.04	0.08	250
2	0.38	0.08	0.13	78
3	0.67	0.07	0.12	30
accuracy			0.77	1596
macro avg	0.55	0.29	0.30	1596
weighted avg	0.70	0.77	0.70	1596

Performance metrics for Random Forest Classifier

	precision	recall	f1-score	support
0	0.96	0.90	0.93	4954
1	0.49	0.83	0.61	1006
2	0.00	0.00	0.00	332
3	0.00	0.00	0.00	92
accuracy			0.83	6384
macro avg	0.36	0.43	0.39	6384
weighted avg	0.82	0.83	0.82	6384

Performance metrics for Convolutional Neural Network

Bert classifier

	precision	recall	f1-score	support
0	0.95	0.95	0.95	1003
1	0.35	0.34	0.34	195
2	0.40	0.26	0.31	62
3	0.12	0.06	0.08	17
accuracy			0.91	1277
macro avg	0.43	0.38	0.40	1277
weighted avg	0.74	0.75	0.75	1277

Performance metrics for BERT

(6) Conclusion

The main goal of the project is to distinguish false news from legitimate news by finding textual patterns. In this project, major time was involved in cleaning the data, extracting the necessary text and generating the features. We have also taken into account the website details to analyse which websites generate the most inaccurate news. Then we detected false news using various machine learning models. The learning classification models are trained and parameter-tuned to generate optimum accuracy. The results of each algorithm are compared by taking the performance metrics into account. We were able to achieve the maximum accuracy and least mean square error through the BERT algorithm. The results are attached and discussed in the Results section.

(7) Future works

As future work, we will focus on collecting more data and removing the imbalances of the labelled data to a greater extent. With the new AI systems capable of generating human-like data, there will be an even more increase in the amount of misinformation being spread. So we have to come up with smarter and robust systems to detect the false news. So, in the future, we will build systems with more discriminative features by investigating the source of information in more detail(whether it's coming from a dangerous source) and with additional performance metrics. By this we can design effective classification systems and see an improvement in the accuracy.

(7) References

1. Kai Shu, Amy Sliva , Suhang Wang , Jiliang Tang, Huan Liu. “Fake News Detection on Social Media: A Data Mining Perspective”
2. T.O. Olaleye, O.T. Arogundade, A. Abayomi-Alli, A.K. Adesemowo. “An ensemble predictive analytics of COVID-19 infodemic tweets using bag of words”
3. V. Mazzeo, Andrea R., Giovanni Giuffrida. “Detection of Fake News on COVID-19 on Web Search Engines”
4. Miguel A. Alonso, David Vilares, Carlos Gomez Rodriguez, Jesus Vilares. “Sentiment Analysis for Fake News Detection”
5. Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, Muhammad Ovais Ahmad. “Fake News Detection Using Machine Learning Ensemble Methods”
6. Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, Rob Procter.” Detection and resolution of rumors in social media:”
7. Z Khanam et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. “Fake News Detection Using Machine Learning Approaches”
8. Reference for web scraping - https://www.datacamp.com/community/tutorials/web-scraping-using-python