COVID-19 PANDEMIC AND FABRICATED NEWS DETECTION

1st YASH NARESHKUMAR
Master of Computer Engineering
Stevens Institute of Technology
New Jersey, USA
Yashnareshkumarpatel49@gmail.com

2nd LONGYUE GUAN
Master of Business Intelligence and Analytics
Stevens Institute of Technology
West New York, USA
lguan6@stevens.edu

Abstract—**Mortal pandemic situations like COVID-19 and it's relative impacts like how long it will probably be going to last, remedies to prevent it, it's negative impact on people across the globe etc. plays a vital role in people's life. In such a situation it's highly recommended to make people more aware of what are possible ways to fight against such a pandemic situation without being panic. And hence we are initiating an attempt to solve a problem created by some of the unusual factors that create an unnecessarily threat like situations among the people. Through this initiative we are trying to find out a solution that will make the prediction of how long this particular deadly pandemic going to last based on certain factors like the amount of people being affected by this pandemic, the recovery rate across the globe and other such similar factors. Also, through this initiative we are also trying to eradicate the consequences that occurs among the people due to impact of the fake news to a great extent that plays with people emotion especially during such situations.**

I. Introduction

- **PROBLEM STATEMENT: -**

In pandemic situation like Covid-19 it is really sturdy to know that how long this pandemic situation is going to continue and along with that the spread of fake news regarding this pandemic creates panic like situation and an environment of threat among the people.

- **OBJECTIVE: -**

Through this initiative we are trying to find out a solution that will make the prediction of how long this particular deadly pandemic going to last based on certain factors like the amount of people being affected by this pandemic, the recovery rate across the globe and other such similar factors.

Also, through this initiative we are also trying to eradicate the consequences that occurs among the people due to impact of the fake news to a great extent that plays with people emotion especially during such situations.

- **SCOPE: -**

If this algorithm is implemented accurately, it will help the people to be plucky in such a mortal situation. It will also help people not carried away by the fake news easily along with helping them detecting how genuine the news are which other people are broadcasting.

II. Related Work

This section summarizes existing solutions to the problem or similar problems. Please try to categorize these existing techniques and provide some discussion on the pros and cons of them. Don't forget to include references to any existing work you mention. The paper "A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection," written by Sourya Dipta Das, Ayan Basak, and Saikat Dutta, mentioned how to use heuristic post-processing to do fake news detection. The paper figures out the fake news identification post-process block diagram and provides the performance of each model. The paper's process is very similar to what our team is trying to do and trying to figure out. Therefore, the report could evaluate and analyze the performance of COVID-19 fake news detection.

Another paper, "Fake news and COVID-19: modeling the predictors of fake news sharing among social media users," written by Oberiri Destiny Apuke and Bahiyah Omar, provides information about understanding research models and hypotheses of fake news from different perspectives such as altruism, entertainment, socialization and so on. The paper mentioned the convergent validity and discriminant validity to measure the model and used path coefficients, the t-test value, predictive relevance, and coefficients in the structural model.

We were learning from those two papers since those provided brilliant information and gave the excellent process of fake news detection. It could help us notice what we should do in our project process and what kind of result and outcomes we expected to get. After reading those two papers, we would like to do heuristic post-processing and get the confusion matrix to get the project's desired outcomes

III. Our Solution

This section elaborates our solution to the problem.

A. Description of Dataset

- **DATASET SCRAPING: -**

First of all, we got the datasets from Codalab Covid 19 Fake News Detection in English Competition. It includes 3 datasets, which are constraint test, constraint train and constraint val.

Therefore, we concat three data frames together and get 4480 real news and 4080 fake news.

When we set up all datasets, the dataset has more than one column, which includes ID, tweets columns.

However, we only want to keep the content part as a helpful resource and as our datasets in our project. Therefore, it is necessary to save the Data Frame with only the content column.

We could get the basic data statistics of our dataset, it could notice that the concat DataFrame got the count number as 10700, the mean is 2354.5 and the standard deviation of the dataset is 1820.13.

However, we get a different basic data statistic with the dataset after the data preprocessing step. We could get the tokenized DataFrame of count as 10700 and unique as 10450.

- **DATA PREPROCESSING: -**

It is necessary to do the data preprocessing part because when we got the datasets from CodaLab competition, the content format will be very different from standard content.

It might include other things that will cause the result of the detection of fake news.

Therefore, it is necessary to do the data preprocessing and do the clean data part.

It needs to go through some steps to finish the preprocessing process.

1. Remove emoji: -

It is necessary to remove emoji is because many Twitter users like to use emojis to show their emotions and reactions to specific tweets.

However, it does not have any meaning for our project; therefore, it should be removed from the content datasets.

2. Remove URL: -

When you enter Twitter and take a look at any tweets, it could find out that every tweet will contain a URL link at the end of the tweets.

Mainly, when we scrapped the datasets from Twitter, the URL link will follow at the end of each content. It will also not help detect, but it will also affect our test score and the detection result.

Therefore, it is necessary to remove all URL links of the datasets.

3. Customize stop-words: -

Customize stop-words are words that do not have any special meanings in English, such as hi, like, want, it, and others. Since those kinds of words do not have any particular purpose, they should be removed from the content of the datasets. It will avoid any influences of further steps to do the detection.

4. Lemmatization: -

Lemmatization is a process that could transfer all words into their dictionary format or lemma format.

And the main idea of the lemmatization process is to try to reduce all words into their dictionary format.

That step could avoid all possible issues or influences for further analysis and other detection.

5. Remove whitespace: -

Removing whitespace is because those whitespaces might lead to different answers.

when it is necessary to do the tokenization process. Because the whitespace will count as one space when it is needed to do the tokenization or bigrams.

That will lead to the result very different as expected. Therefore, it needs to remove whitespace during the data preprocessing process.

6. Make text all lowercase: -

Take text to lowercases is because if the first words of a sentence are the same as any words in the sentence. It will count twice and will treat as a different word from the system. That's why it is necessary to make text all lowercase, and it could avoid this kind of situation happens during the process of detection.

7. Remove punctuations: -

And the last step for the data preprocessing process is to remove punctuations, and it will lead same situations as the whitespaces. So, it is also necessary to remove all of them. They do not have any special meanings and will confuse when to do the tokenization and other steps. Therefore, it should be removed during the preprocessing steps.

• TOKENIZATION: -

Tokenization is a process that could break each sentence or content of the datasets into different parts before doing any model or test. The tokenization process could also split one sentence into bigrams or trigrams. And it depends on what we want to use for the further steps. Also, we could use the tokenization process to find out what kinds of words frequently show up in the tweets that we scrapped. Therefore, tokenization is a necessary method that we need to use to find out a lot of information, and it is a step that we need to use to ready to do the further steps.

B. Machine Learning Algorithms

1. NAVIE BAYES CLASSIFIER: -

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Here we are using this algorithm to calculate the probability of the death rate of the people.

It handles both continuous and discrete data. It is highly scalable with the number of predictors and data points.

It is fast and can be used to make real-time predictions.

2. CONFUSION MATRIX: -

A confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier.

It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score.

3. SVM CLASSIFIER: -

The SVM is used for classification and regression problems it uses a kernel trick technique to transform our data and then based on this transformation it finds an optimal boundary between the possible outputs.

4. PASSIVE AGGRESSIVE CLASSIFIER: -

The passive aggressive classifier is used for large-scale learning.

This algorithm remains passive if correct result is obtained after classification and turns aggressive if there is any miscalculation or incorrect result.

Here we will use it in the fake news detection part.

5. LOGISTIC REGRESSION: -

Here we use Logistic regression as it is appropriate to conduct when the dependent variable is binary.

It is used to describe the data and to explain the relation between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic Regression is a Machine Learning algorithm which is also used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1.

Although logistic regression is best suited for instances of binary classification, it can be applied to multiclass classification problems, classification tasks with three or more classes. You accomplish this by applying a “one vs. all” strategy.

It is better because good accuracy for many simple data sets and it performs well when the dataset is linearly separable. Logistic Regression requires average or no multicollinearity between independent variables. ... Logistic regression is less inclined to over-fitting, but it can be overfit in high dimensional datasets.

6. DECISION TREE CLASSIFIER: -

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree.

7. RANDOM FOREST CLASSIFIER: -

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity.

8. GRADIENT BOOSTING CLASSIFIER: -

Gradient boosting is a greedy algorithm and can overfit a training dataset quickly.

It can benefit from regularization methods that penalize various parts of the algorithm and generally improve the performance of the algorithm by reducing overfitting.

C. Implementation Details

• CALCULATING ACCURACIES WITH DIFFERENT MODELS

Here we have calculated accuracies using different models using parameters like TF-IDF. TF*IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF).

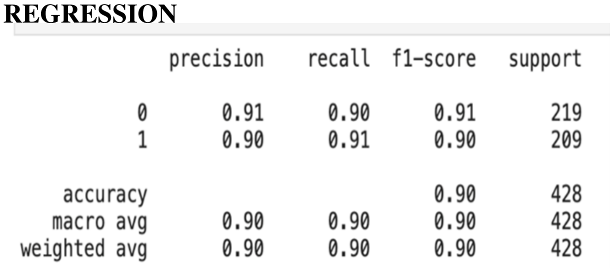
Each word or term that occurs in the text has its respective TF and IDF score. TF*IDF is used by search engines to better understand the content that is undervalued.

It is used when it has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP).

Thus, the term frequency is often divided by the document length as a way of normalization: TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency, which measures how important a term is.

• ACCURACY AND ROC CURVE OF LOGISTIC REGRESSION

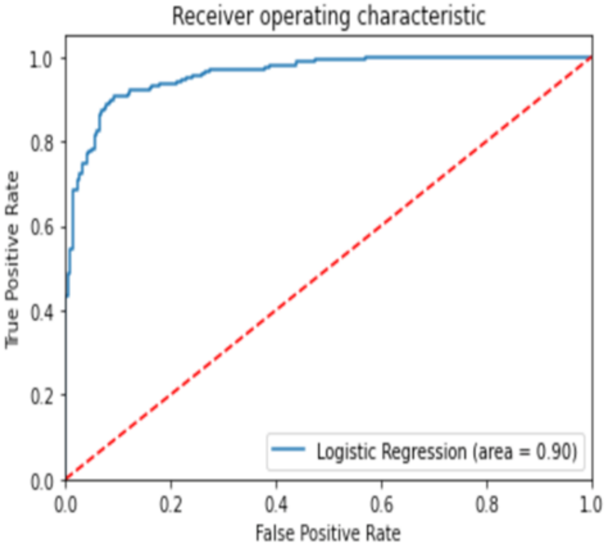


```
              precision    recall  f1-score   support

           0       0.91      0.90      0.91       219
           1       0.90      0.91      0.90       209

    accuracy                           0.90       428
   macro avg       0.90      0.90      0.90       428
weighted avg       0.90      0.90      0.90       428
```

Logistic Regression is the best model that has the highest accuracy, which is 0.9.

ROC determines the accuracy of a classification model at a user defined threshold value. It determines the model's accuracy using Area Under Curve (AUC). The area under the curve (AUC), also referred to as index of accuracy (A) or concordant index, represents the performance of the ROC curve.

• ROC CURVE OF LOGISTIC REGRESSION



An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.

The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. ... The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

ROC curves are frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests.

In addition, the area under the ROC curve gives an idea about the benefit of using the test(s) in question.

It is a horizontal line with the value of the ratio of positive cases in the dataset. For a balanced dataset, this is 0.5. While the baseline is fixed with ROC, the baseline of [precision-recall curve] is determined by the ratio of positives (P) and negatives (N) as y = P / (P + N).

• ELAPSED TIME PERFORMANCE WITH DIFFERENT MODELS

The time functions detect events that happen at unusual times, either of the day or of the week. These functions can be used to find unusual patterns of behavior, typically associated with suspicious user activity.

The machine learning features include the following time functions: -

1) Time-of-day:-.

The time-of-day function detects when events occur that are outside normal usage patterns.

The function expects daily behavior to be similar. If you expect the behavior of your data to differ on Saturdays compared to Wednesdays, the time-of-week function is more appropriate.

2) Time-of-week:-

The time-of-week function detects when events occur that are outside normal usage patterns.

• FIT METHOD

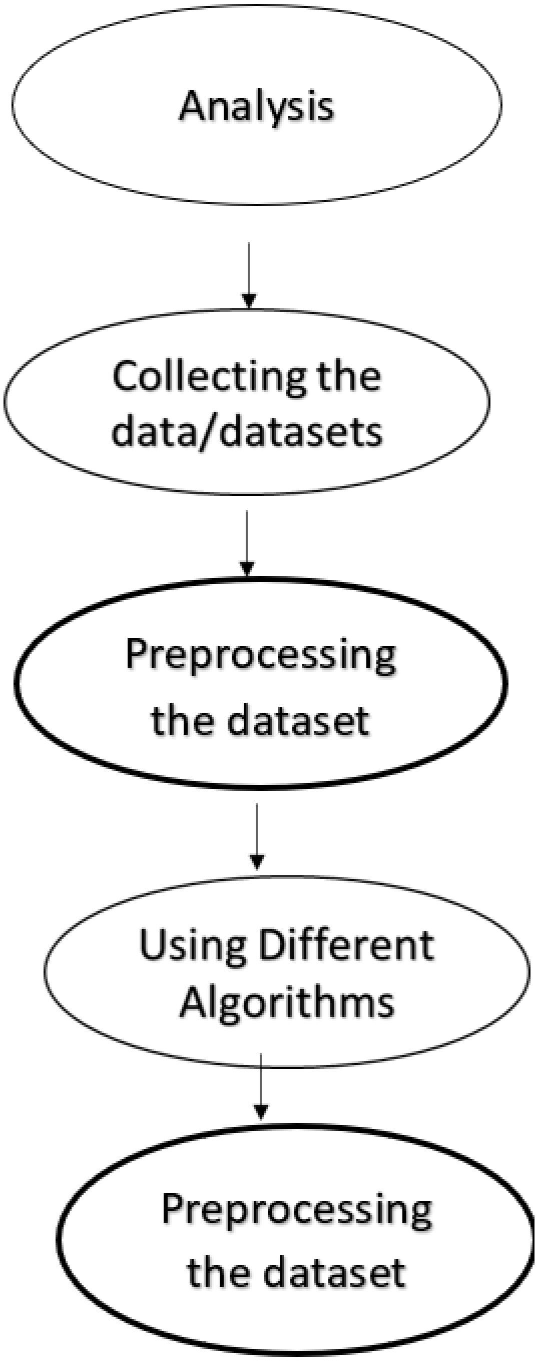
In contrast to machine learning, fitting means training. There is a fit function in ML, that is used for training of model using data examples. Fit function adjusts weights according to data values so that better accuracy can be achieved. After training, the model can be used for predictions.

The good fit is Ideally, the case when the model makes the predictions with 0 error, is said to have a good fit on the data. This situation is achievable at a spot between overfitting and underfitting.

fit () is implemented by every estimator and it accepts an input for the sample data (X) and for supervised models it also accepts an argument for labels (i.e., target data y). Optionally, it can also accept additional sample properties such as weights etc. fit methods are usually responsible for numerous operations..

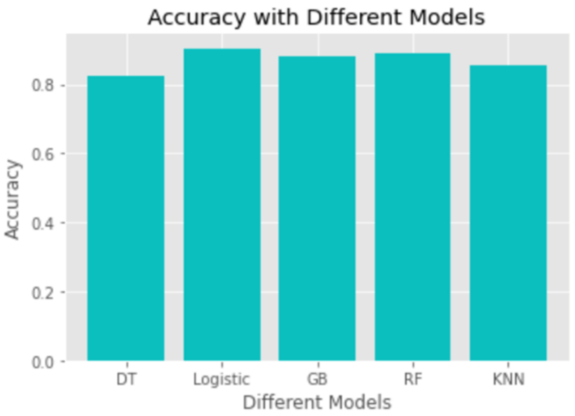
The difference between fit and fit transform is fit performs the training, transform changes the data in the pipeline in order to pass it on to the next stage in the pipeline, and fit transform does both the fitting and the transforming in one possibly optimized step.

WORK FLOW: -

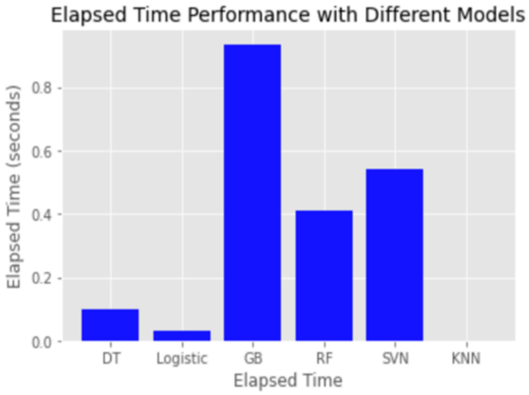

Analysis
Collecting the data/datasets
Preprocessing the dataset
Using Different Algorithms
Preprocessing the dataset

IV. Comparison

Here we are using Naive Bayes' algorithm instead of decision tree because navie bayes' algorithm will outperform decision trees when it comes to rare occurrences.



For example, that may occur a probability that some deaths occurred due to other diseases and not because of covid which will be rare in our case. A decision tree will almost certainly prune those important classes out of your model. KNN ¿ Logistic Regression ¿ Decision Tree ¿ Random Forest ¿ SVM ¿ Gradient Boosting



V. Future Directions

If we are awarded with some more time like 3-6 months extra, we would definitely like to develop an application for the same model which may have more accuracy than this and use would definitely like to do certain modifications based on the public demand in future. We could consider about not only use TF-IDF parameters, but also could use Count Vectorizer parameters. It could also do the comparison by using same models.

We might also could use one cloud platform to do the work and see how the difference between the elapsed time performance.

We could think about collecting more dataset to get a better result.

VI. Conclusion

Here we described that how we built a model by training and testing the dataset. Moreover, in this project we try to build a model using various important techniques like Logistic Regression, Decision Tree Random Forest etc. Last but not the least we also calculated the accuracy of our model. Moreover, Logistic Regression is the best model that has the highest accuracy, which is 0.9

KNN and Logistic Regression are the model has the better elapsed time performance, which are 0.001 and 0.03.

VII. Bibiliography

1. A. Douglas, “News consumption and the new electronic media,” The International Journal of Press/Politics, vol. 11, no. 1, pp. 29–52, 2006.View at: Publisher Site — Google Scholar

2. A. D. Holan, 2016 Lie of the Year: Fake News, PolitiFact, Washington, DC, USA, 2016.

3. A. Robb, “Anatomy of a fake news scandal,” Rolling Stone, vol. 1301, pp. 28–33, 2017.View at: Google Scholar

4. Brownlee, Jason. “Your First Machine Learning Project in Python Step-By-Step.” Machine Learning Mastery, 19 Aug. 2020, machinelearningmastery.com/machine-learning-in-python-step-by-step/.

5. Das, Sourya Dipta, et al. “A Heuristic-Driven Ensemble Framework for COVID-19 Fake News Detection.” ARXIV, 2021, arxiv.org/pdf/2101.03545.pdf.

6. D. M. J. Lazer, M. A. Baum, Y. Benkler et al., “The science of fake news,” Science, vol. 359, no. 6380, pp. 1094–1096, 2018.View at: Publisher Site — Google Scholar

7. J. Soll, “The long and brutal history of fake news,” Politico Magazine, vol. 18, no. 12, 2016.View at: Google Scholar

8. J. Wong, “Almost all the traffic to fake news sites is from facebook, new data show,” 2016.View at: Google Scholar

9. “Learn.” Scikit, scikit-learn.org/stable/.

10. S. A. García, G. G. García, M. S. Prieto, A. J. M. Guerrero, and C. R. Jiménez, “The impact of term fake news on the scientific community scientific performance and mapping in web of science,” Social Sciences, vol. 9, no. 5, 2020.View at: Google Scholar

11. Sah, Raman. “Simple Machine Learning Model in

Python in 5 Lines of Code." Medium, Towards Data Science, 11 Sept. 2018, towardsdatascience.com/simple-machine-learning-model-in-python-in-5-lines-of-code-fe03d72e78c6.

12. S. Kogan, T. J. Moskowitz, and M. Niessner, "Fake News: Evidence from Financial Markets," 2019, https://ssrn.com/abstract=3237763.View at: Google Scholar

13. Tomlins, Scott. "How to Scrape Tweets by Location in Python Using Snscrape." Medium, The Startup, 26 Dec. 2020, medium.com/swlh/how-to-scrape-tweets-by-location-in-python-using-snscrape-8c870fa6ec25.