# TECHNIQUES TO DETECT FAKE NEWS

Our main agenda is detecting fake news with maximum accuracy. In our paper we have used three modelling techniques Logistic Regression, Random Forest, Support Vector machine to understand which method gives the highest accuracy rate. Also, an appropriate Python code to get the optimized results using the data sets which is a combination of both real and fake news. So, data is collected from various news articles and journals. This data has been preprocessed to make it useful and then imported to the modeling algorithm called logistic regression. The results contain the news detection in the binary form of 0 and 1, confusion matrix, accuracy.

## A. Research Questions

During this experiment we want to explore the research questions that we got in the process.

RQ1: Can we detect Fake news early?

RQ2: What can be the best approach to detect fake news early?

RQ3: What will be the accuracy of the algorithm in detecting the fake news?

These research questions show the approach from where to where we have evolved in the research progress. Initially when we started with fake news research, we wanted to know if fake news was detected early so we can avoid the causes of fake news and cyberbullying early. In the recent time we have seen a new feature in the tweeter that anyone can get the verified tick mark on purchasing it with eight dollars. This feature is misused by creating fake accounts. A fake account was created in the name of Eli Lilly pharma company and that fake account got the verification tick mark by purchasing and then posted a tweet that Diabetic medicine insulin is free. And this tweet had gone viral, and many thought it was true. This issue was made pointedly and because of this issue company share value has gone down by 4.45%. As Eli Lilly is a big organization, they have defended what if a group of people target a common man and it would be hard from him/her to defend against the fake news spread about them. So, detecting fake news before it goes viral can stop these kinds of issues and cyberbullying.

And then we came to know that there exist two different approaches, one is manual approach and other one is algorithms to detect fake news. In our paper we have included machine learning algorithms, logistic regression, random forest and support vector machines. In our paper we wanted to present the accuracy of these algorithms.

## A. Methodology

Fig 3.1: Flow chart of proposed system

# I. Data Loading

The data which we have used in this project is available in Kaggle, which can be downloaded using the Kaggle API. Data sets consist of two folders which have both fake and real news. The file we got here is in json which was later converted to .csv using the packages glob and zipfile in python.

# II. Data Preprocessing

Data Preprocessing is an important step which helps to reduce the dimension of our data. The data must be clean, and all the null values must be replaced. Data preprocessing includes two most important steps: Tokenizing and Stemming. *Tokenizing* means converting the text into smaller units with rightful meaning. The text we have goes through several stages here. All the text is converted to lowercase because in machine learning words with Upper case are considered as another feature even if it means if we have the same word in lower case. Next step is to remove the tabs and return spaces with empty strings. Removing the punctuation in all the text like comma, full stop, semicolons and spaces is also necessary. Finally, we need to remove stop words which are generally called filler words that help in constructing a sentence, which has no to little meaning in it. They are usually prepositions and conjunctions in the English dictionary. We have used a submodule which contains all the stopwords present in the English vocabulary to eliminate them.

After the initial preprocessing, data goes to the next level of preprocessing called Stemming. Stemming is an algorithm which reduces all the words into their root form. For example, the words like play, played, and playing all have the same meaning but they are in different tenses.

*WordCloud* helps us to determine the frequency of our data. This gives the size of each word in text and the word with high size has the highest frequency which gives us some insights of our data.

# III. Data Transformation

Now the data is clean. Still, we cannot feed it to the machine learning model, as they can only understand numerical data types, so we need to convert it to the numerical format. To do that, we use two methods, Count vectorization, which is also known as the Bag of Words (BOW) representation, where we store the frequency of each word in a sentence in a vector format.TF-IDF - Term Frequency Inverse Document Frequency is an extended version of the previous one, with smoothing and normalization.

# IV. Modeling

We have used three different types of models, Logistic Regression - A simple linear model, which we have used as a baseline model, this model assumes a linear relationship between features and labels. Random Forest - An ensemble of weak models, which combines the predictions of multiple weak models. Support Vector Machines - Another non-linear model, which models a relationship between the features and labels by projecting the features into higher dimensional space, and creating a linear decision boundary between the classes..

Fig 3.2: Classification Algorithms

# V. Evaluation

We have used two metrics in evaluating the model’s performance. Accuracy Score, which is the percentage of matching between the actuals and predictions. Confusion Matrix, defines a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.

# C. ANALYSIS

The datasets consist of training and testing data, the training data is divided into training and validation partitions. The evaluation procedure is done on each of the three partitions, training, validation, and testing sets.

Once we get the predictions from each of the partitions, we use predicted values and ground truth values in generating the confusion matrix. Confusion matrices determine the numbers of 1 and 0. Here we use two matrices, for calculating the model performance. I.e,

1. Accuracy Score
2. Confusion Matrix

Accuracy Score is the matching percentage between the actual value and predicted values. Here, the accuracy score depends on the dataset. dataset divided into 3 parts tests, train, and validate accuracy score . Training accuracy score means that data fed to the model for training. The performance of the training data determined the validation accuracy score. Based on the value of the validation accuracy score we tune the hyper-parameters of the model to make it more robust and generalize on the test dataset. The testing accuracy score is a replica of real-world data, whose distribution will be very different from the training dataset. which gives the actual performance of the model when we deploy it in the real world.

Here, we used three models that are represented in two vectors

1. TF-IDF Vectorization

2. Count Vectorization

The below fig 4.1 shows the accuracy score in 2 vectors of logistic regression, random forest, and support vector Machines.

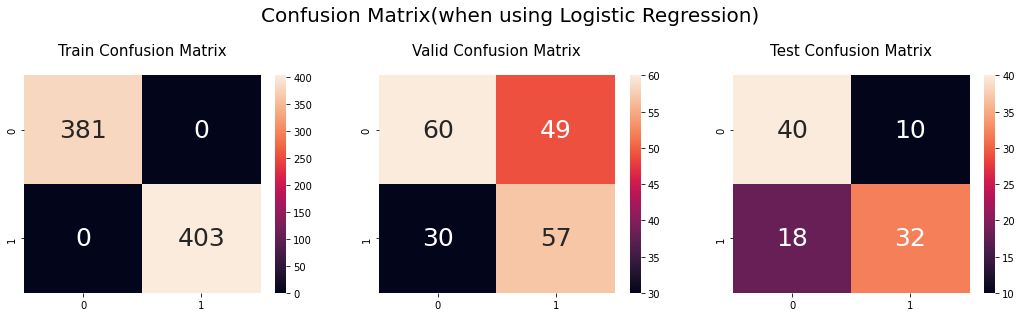
| **Model** | **Training Accuracy**  **(Count Vec)** | **Validation**  **Accuracy**  **(Count Vec)** | **Testing**  **Accuracy**  **(Count Vec)** | **Training Accuracy**  **(TFIDF Vec)** | **Validation**  **Accuracy**  **(TFIDF Vec)** | **Testing Accuracy**  **(TFIDF Vec)** |
| --- | --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **100%** | **58%** | **64%** | **95%** | **46%** | **68%** |
| **Random Forest** | **100%** | **64%** | **68%** | **100%** | **63%** | **63%** |
| **Support vector Machines** | **84%** | **56%** | **75%** | **99%** | **38%** | **67%** |

2. Confusion matrix

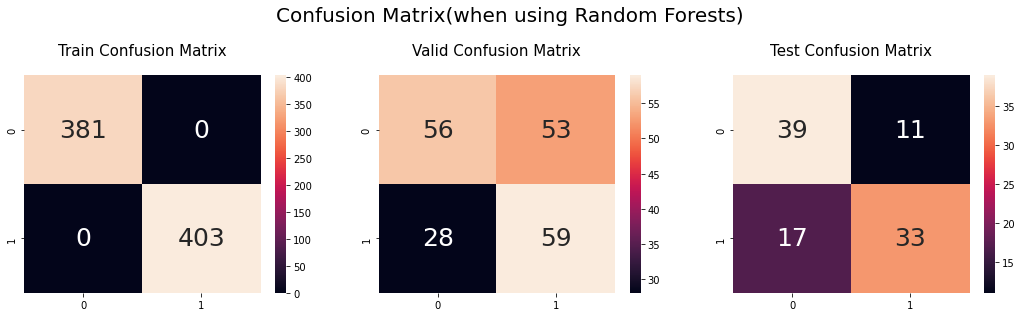
It displays the ratio of actual value to the predicted value. It calculates how many predicted values are really predicted as actual values. And how many predicted values are really predicted as not actual values.

Count Vectorization Representation:

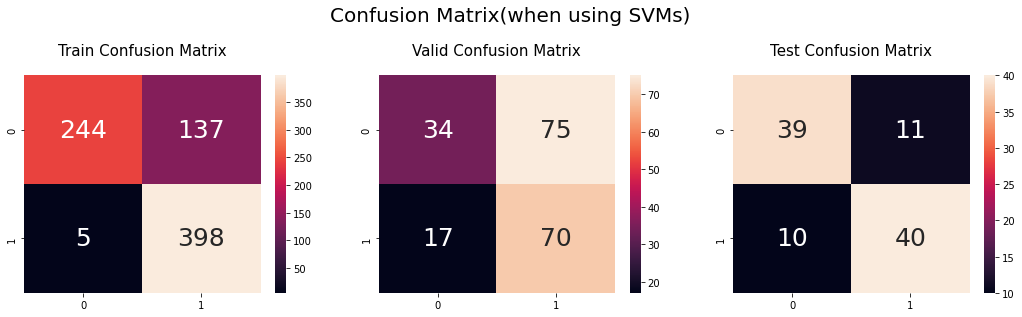
Logistic regression



Random Forest

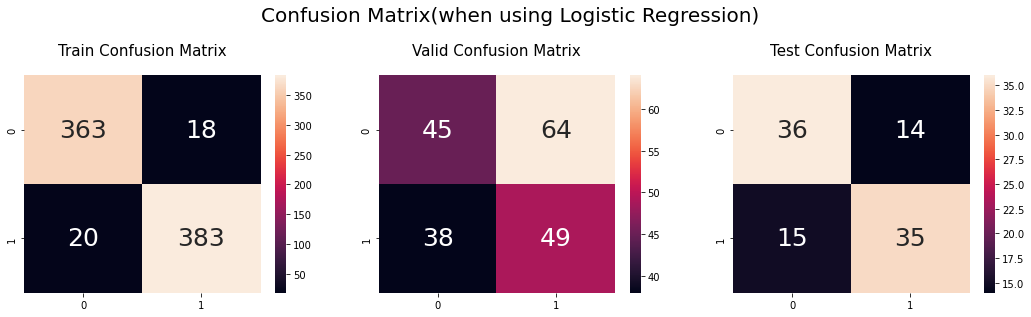


Support Vector Machines(SVM)

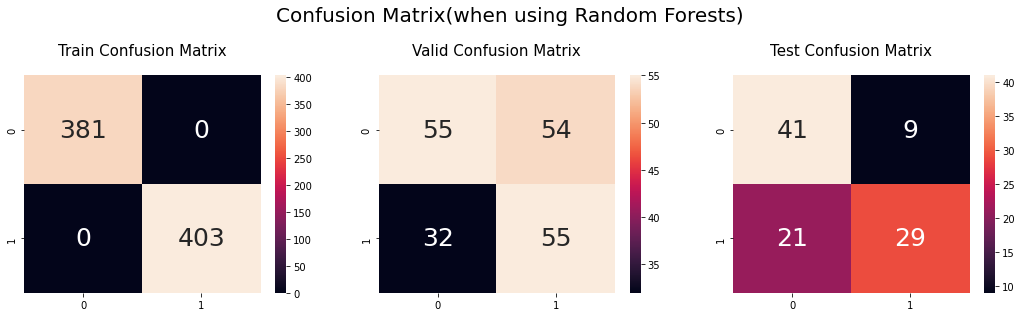


TF-IDF representation for

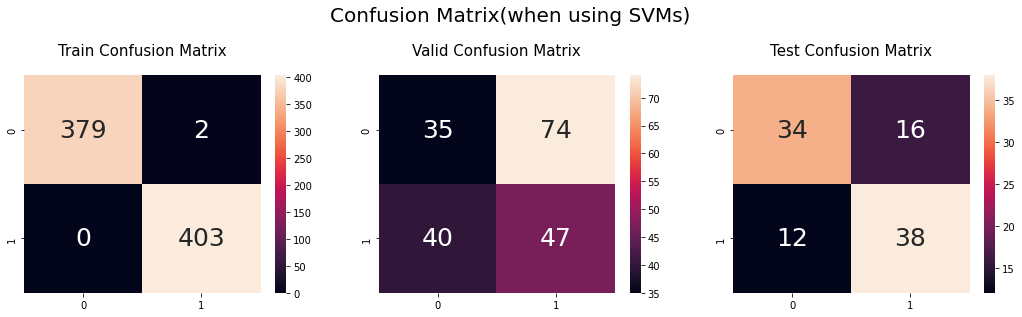
Logistic regression



Random Forest



Support Vector Machines(SVM)



The Confusion matrices for the training partition are almost the same with approximately zero False Positives and False Negatives.

# CI. THREATS TO VALIDITY

## A. Internal Threats

The internal threat to our paper would be the accuracy of the model we are using. It will be able to provide more accurate results if we provide it with more data and test it with various other datasets.

Logistic regression is a simple mode with less complexity. Random forest is a complex model with a high number of parameters. SVM is intermediate between both the model's logistic and random forest. So, random forests may cause overfitting.

Another reason for overfitting is due to less data. That means we should use large datasets.

## B. External Threats:

The external threat to our paper would be not being able to find the correct datasets to test. The wrong datasets can make the results baised and won't produce the desired output when we deploy the model in a real-world application and the results that we obtained won't be accurate.

# I. CONCLUSION

In this paper we have studied the fake news Detection in the articles , how these false news leads to Cyberbullying . The dataset has been taken, a set of features were extracted and provide these data to the Count Vector Representation and TF – IDF to input these obtained data to the Machine Learning models. Experiments have been done on a real-world dataset from Kaggle giving us the desired results with at most accuracy by using our proposed methods of detecting fake news.

Results

By performance of the model we can determine that, in one aspect, support vector machines detect the fake news with 75% accuracy. And in another aspect Logistic regression detects the fake news with 68% accuracy.

By above 2 questions we found that support vector machine can detect the fake news, and