International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Real Time Fake News Detection Using Machine Learning and NLP

Aman Srivastava[1]

[1]Student at Department of Electronics and Communication Engineering, JSS Academy of Technical Education Noida, Uttar Pradesh, India

---

Abstract - *News is the most vital source of information for common people about what is happening around the world. Newspapers are an authentic source of news, but nowadays social networks have become the emerging source of news. Due to easy access to these social networks, the news can be easily manipulated which gives rise to fake news. Fake news can be used for economic as well as political benefits. It can be used as a weapon to spread hate among the community which can harm society. So it is crucial to detect fake news to avoid its consequences. There is no existing platform that can verify the news and categorize it. This paper proposes a system that can be used for real-time prediction of news to be real or fake. This system is based on natural language processing to extract features from the data and then these features are used for the training of machine learning classifiers such as Naive Bayes, Support Vector Machine (SVM), Random Forest (RF), Stochastic Gradient Descent (SGD), and Logistic Regression (LR). Each of the classifier performance is evaluated on various parameters. Then the best performing classifier is deployed as a website using flask API for real-time prediction of the news*

Key Words: Fake News, Bag of Word, TFIDF, POS Tagging, Naïve Bayes, Support Vector Machine (SVM), Random Forest (RF), Stochastic Gradient Descent (SGD), and Logistic Regression (LR), Pipeline, Flask API

1. INTRODUCTION

In traditional news making procedures, very limited and authorized individuals are involved and newspapers, radio, television were the only source of news. Due to these reasons news, credibility and authenticity are preserved. But in the era of internet, social network is becoming a news source of news. Easy and free access to these social networks makes the task of fabricating fake news and manipulating news a very effortless task. There is no authorize control point of these manipulated fake news which creates a question over there credibility and authenticity. The ease of getting direct news from the platform they mostly use has attracted the user. The reason to spread fake news can be social, political, and economical. Fake news in business can affect the stocks of the company leading to a huge capital loss. During the election campaign, fake news is used as a weapon against each other in a political war to defame the opposition. The most adverse effect is seen when it is used to spread communal hates which leads to riots. The Delhi riots are the best example of the destruction caused by fake news. Fake news about the COVID-19 in India lead to an attack on the medical team in various parts of the country and thus making the fight against the virus weak. The rate at which it spread is very fast due to which controlling the spread manually is not possible.

There is no platform via which the user can check the credibility and authenticity of the news and where authorities can directly inform about the fake news prevailing. Due to which people can believe in the news which can be a trouble for them and as well for society also. In the existing system, the action is taken after the adverse impact had already hit society. The proposed platform is useful for both common people and official authorities to prevent the spread of rumours in form of news.

2. LITERATURE SURVEY

Kushal Agarwalla [1] proposed a system by comparing there ML classifier i.e. Logistic Regression, Naïve Bayes, Support Vector Machine using tokenizer as a feature. The maximum efficiency of system is achieved by using Support Vector Machine. The accuracy of model is only 81.25%.

Chaitra K Hiramath [2] has compared both Machine Learning and Deep learning in the proposal. The maximum accuracy is given by Deep Neural Network which 91% .Machine is learning gave only 89% accuracy as highest and that to by Naïve Bayes, which is very basic classifier .

A.Lakshmanarao [3] has used two NLP methods i.e. Bag of Words and TFIDF and compares the output of Machine Learning algorithms. The highest accuracy of 90.70% is given by Random forest classifier. The results are shown by plotting confusion matrix.

Abdullah-All-Tanvir [4] has compared both Machine Learning and Deep learning in the proposal. Count Vector, TFIDF, Word Embedding is used to build models. Highest efficiency of 89.34 % is given by Support Vector Model.

From the above proposals, we can conclude that for higher accuracy of the model can be obtained by using a combination of the Natural Language Processing features. Further accuracy can be increased by smart cleaning and processing of data set. Words with meaning and importance should not be removed or manipulated.

3. METHODOLOGY

Our objective is to find the best Machine Learning Algorithms and Natural Learning Process methods for the prediction of news. Then the best performing model will be saved and will be linked to a user interface by which it can predict new input data in real-time. To achieve this objective training data have

to go through various intermediate processes before giving it to Algorithms. The main parameters on which the performance of the model will be judge are F1 score and accuracy of the model.

The basic work flow diagram is shown in figure 1 which depicts various intermediate processes for create the real time platform for prediction of news.

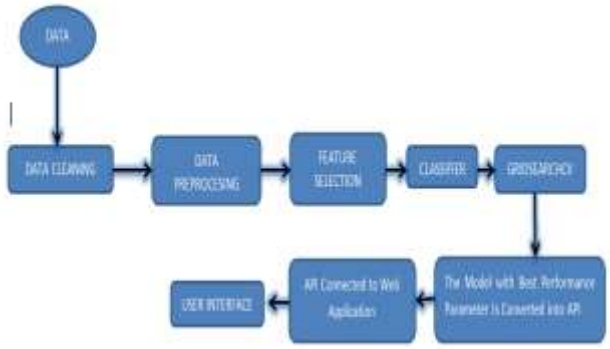

Fig -1: Flow Diagram of Proposed System

3.1 Data

Data is the prime ingredient of this project, as these data features are extracted using Natural Language Processing. By using these features of the data, Machine Learning Algorithms are trained and models are created. In this proposal, we have taken 6630 news with equal proportionality of fake and real. Data is saved in Comma Separated Value format. This data set is divided in the ratio of 80:20 for training and testing of algorithms.

3.2 Data Cleaning

Data set is a chunk of data that is in raw form. It may contain certain symbols like digits, special characters, blank lines, and data without any label. These symbols should be removed as it is of no importance and it can also affect the performance of the model in an adverse manner.

3.3 Data Preprocessing

After data cleaning, data is now free of unwanted symbols. This data should be converted into the form which can be used for extracting the features easily. It includes the following process:

•Converting each of the word in lowercase to avoid ambiguity between cases.
•Removal of the words which contain just one letter.
•Removal of the words that contain digits.
•Tokenize the data and removal of punctuations
•Removal of empty tokens.
•**Stop Words**: Stop words are useless words for NLP like "the", "a", "an", "in". This should be removed.
•**Lemmatization:** It is the process of converting the words to their root forms such as studies and studying is converted into a study. But this can change the meaning of words. So to preserve it Part of Speech tag is attached to individual token to preserve its meaning. This is most important as the proposal studied in the literature survey does not have high accuracy due to the absence of POS which affect the performance of the model in an adverse manner

3.4 Feature Selection

Feature Selection is the process where we select those features which contribute most to your prediction variable or output. Certain features are present in fake news, we have to extract them and accordingly, our classier is trained to predict the news. Here feature are the important words which appear in news. Natural language processing methods which are used to extract features are:

1) **Bag of Words**: In this, each document is converted into a vector. First, all the words are taken out of the dataset forming bag of words. The vector of each document is the frequency of words present in it out of bag of word.

2) **TF-IDF**: Bag of word depends only on the frequency of the words only. Certain words have very high frequency but are of no importance. To avoid this inverse document frequency is added which downscales words importance that appears a lot across documents. TF-IDF is word frequency scores that try to highlight more interesting words.

3)**POS-Tag**: The process of classifying words into their parts of speech like nouns, verbs, adjectives, and adverbs and labelling them accordingly is known as part-of-speech tagging. POS tagging looks for relationships within the sentence and assigns a corresponding tag to the word.

4) **N Gram**: In this method instead of the individual word, N number of words is taken into consideration as a single unit at a time. N number of words is considered to a feature.

These feature selection methods will be used as individually and in combination with each other to increase the efficiency, and the more important feature will be extracted from the data..

3.5 Classifiers

Classifiers are the algorithms that are capable of classifying the input based on the features present in the input. The first classifier should be trained on features which will be present in a different class.

1) **Support Vector Machine**

SVM is a supervised machine learning algorithm that can be used for classifications. Features of both the classes are mapped to the graph and an optimum plane known as hyperplane is drawn between them to segregate the classes. This plane is drawn the basis of two support vectors location each of which is nearest to features point of the each class.

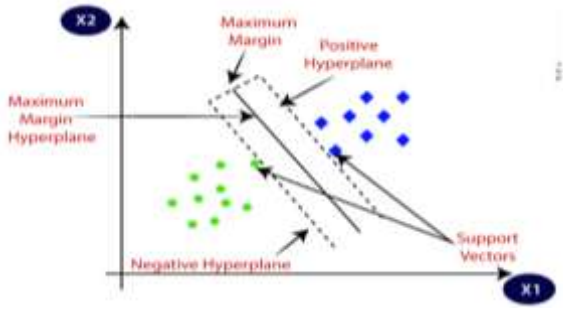

Fig -2: Hyperplane of SVM Classifier

2) Naïve Bayes

Naïve Bayes classifier is a probabilistic classifier based on Bayes Theorem with the assumption of independence among the features. Multinomial Naive Bayes uses the frequency of the words as a feature to classify the data in various classes.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where A and B are independent events/features.

1) Random Forest Classifier

It is the collection of many decision trees, uses bagging, and feature randomness when building each tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. Prediction of each tree is taken as a vote to give the final output.

2) Logistic Regression

It uses the sigma curve instead of the regression line for predicting the categorical dependent variable using a given set of independent variables. In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1.

$$Sig(t) = \frac{1}{1 + e^{-t}}$$

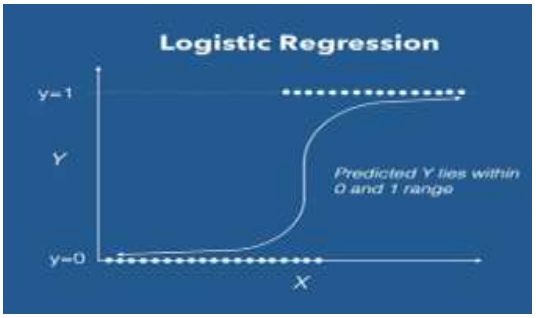

Fig -3: Sigma Function

3) Stochastic Gradient Descent

Stochastic gradient descent considers only one random point while changing weights, unlike gradient descent which considers the whole training data. Logistic Regression by default uses Gradient Descent and as such, it would be better to use SGD Classifier on larger data sets to reduce processing time. By default, the SGD Classifier does not perform as well as the Logistic Regression. It requires some hyperparameter tuning to be done.

3.6 Hyperparameter Tuning

Parameters that define the model architecture are referred to as hyperparameters and thus the process of searching for the ideal parameter is referred to as hyperparameter tuning. We have used GridSearchCV to tune the parameter of each algorithm. The grid of values of each parameter is given as input and GridSearchCV will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid. The model with the best parameter value is given as output.

4. RESULT COMPARISION

Best Model i.e. combination of Machine learning algorithms with NLP method is selected after comparing the results of the twenty models. These twenty models are made by taking all ML algorithms with a combination of NLP methods.

The accuracy of prediction and F1 Score of each model is shown in Table 1 and Table 2 respectively.

Table 1: Accuracy of Models

Models	BOW	BOW+ TFIDF	TFIDF+ POS TAG	TFIDF+POS TAG+N GRAM
SVM	93%	93%	95%	94%
NB	89%	89%	81%	81%
LR	93%	92%	94%	94%
RF	91%	92%	92%	91%
SGD	93%	92%	93%	94%

Table 2: F1 Score of Models

Models	BOW	BOW+ TFIDF	TFIDF+ POS TAG	TFIDF+POS TAG+N GRAM
SVM	0.93	0.92	0.95	0.95
NB	0.90	0.90	0.83	0.83
LR	0.93	0.92	0.94	0.94
RF	0.91	0.90	0.92	0.92
SGD	0.93	0.92	0.93	0.94

From the above comparison of the parameters, we can conclude that the Support Vector Machine classifier with TFIDF and POS Tag method is the best performer in terms of both accuracy and F1 score i.e. 95% accuracy and 0.95 F1 scores. The model with the best performance is obtained by tuning the parameter of SVM architecture.

```
Best Estimator Params
SVC(C=10, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='linear',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
```

Fig 4.Best Parameter of SVM using GridSearchCV

To dig deep into the performance we will analyse the confusion matrix parameter of each label i.e. fake and real.

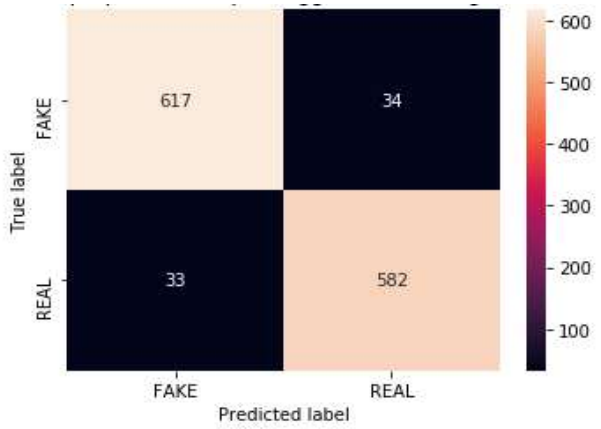

Fig 5. Confusion matrix of SVM

Table 3: Performance parameter of individual label

	Precision	Recall	F1-score
False	0.95	0.95	0.95
True	0.94	0.95	0.95
Accuracy	-	-	0.95
Macro average	0.95	0.95	0.95
Weighted average	0.95	0.95	0.95

From Table 3 we can conclude that the SVM model can predict both true and false input with equal precision and recall. This reduces the error in the prediction by the model. Macro average and weighted average have equal value of .95, which means that the system performs well on overall data set.

5. DEPLOYMENT

As SVM with TFIDF and POS Tag came out to be the best method for the prediction, we need to deploy this model by making it platform independent and creating a user interface. Development of user interface and deployment of the model involves following steps:

5.1 Pickling

Pickle is used for serializing and de-serializing Python object structures. Serialization refers to the process of converting an object in memory to a byte stream that can be stored on disk. The trained model is saved to disk using pickling by which we need not train the model every time we need it. We just need to deserialize it for predicting news input.

5.2 Pipeline

The data entered by the user will be in raw form. It should go through various intermediate processes like cleaning, processing, etc. before it can be used for prediction. Pipeline decides the flow of the data i.e. in what sequence data will go through various processes.

5.3 User Interface

An interface is created user to access the model without going into background detail of processing. It also removes the platform dependence i.e. it will only run on software on which it is modeled. In this project, a website constructed by HTML and CSS is used as a user interface.

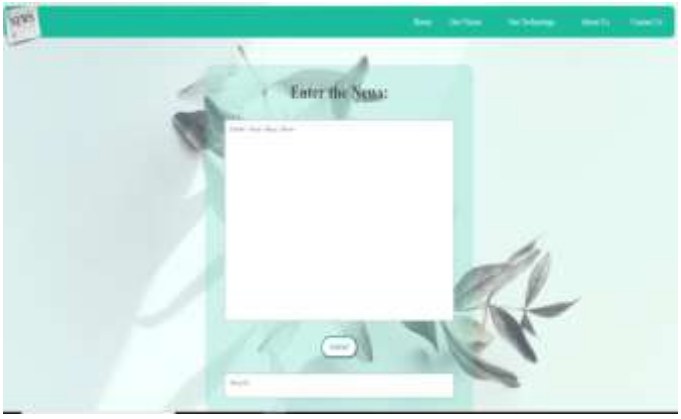

Fig 6.User Interface

5.4 API

API is the tool that is used to create an interface between the two applications. As our prediction model and user interface i.e. website are two applications that should be linked with each other to predict the input news and display it on the webpage. FLASK which is a micro web framework is used to develop a REST API.

6. CONCLUSION

A platform i.e. website is created which is linked to a trained Machine Learning model. SVM with TFIDF and POS tag NLP method is trained and use for the prediction of new news input by the user. This trained model link to the platform is capable of predicting news to be fake or real with an accuracy of 95%. In this platform, users can enter the news and it will predict whether the news is real or fake. It will show output as real or fake with the probability of news to be real.

ACKNOWLEDGEMENT

The first author likes to acknowledge the guidance and support provided by Prof. Ganesha H.S, Assistant Professor of ECE, JSSATEN, Noida, Uttar Pradesh, India. The first author would like to acknowledge the support and encouragement

provided by the management of JSS Academy of Technical Education College (affiliated to AKTU), Noida during this study.

REFERENCES

[1] Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema, "Fake News Detection using Machine Learning and Natural Language Processing," International Journal of Recent Technology and Engineering (IJRTE), Volume-7, Issue-6, March 2019.

[2] Chaitra K Hiramath, Prof. G.C Deshpande, Fake News Detection Using Deep Learning Techniques, 1st International Conference on Advances in Information Technology,2019.

[3] A.Lakshmanarao, Y.Swathi, T. Srinivasa Ravi Kiran, "An Effecient Fake News Detection System Using Machine Learning," International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8, Issue-10, August- 2019.

[4] Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq, "Detecting Fake News using Machine Learning and Deep Learning Algorithms," 7th International Conference on Smart Computing & Communications (ICSCC),2019.