Detection of Fake News using Machine Learning Algorithms

Mhualika Dhar[1] , Reet Rai[2] , Shamanth J P[3] , Soumita Das[4], Shilpa M[5]

[1] *BE Student, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Karnataka, India*
[2] *BE Student, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Karnataka, India*
[3] *BE Student, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Karnataka, India*
[4] *BE Student, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Karnataka, India*
[5] *Associate Professor, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Karnataka, India*

ABSTRACT

The sharing of information via internet has been increasing over the years. The internet has been a source of easy information and is used more than traditional ways like newspapers or magazines. It is important to identify information from the internet as real or fake as mislead information could cause a lot of havoc in the society. Fake information can be the cause of riots, chaos and can affect a large group of society. In this paper, we talk about the methodology used to detect false news using machine learning classifiers and natural language processing to authenticate whether a news is real or not. For the generation of feature vectors, we use the TF-IDF vectorizer. To detect the news as fake or real we're comparing several classifying techniques to find out the best model that could be used to detect false news.

Keywords: - *Fake news detection, logistic regression, decision tree, random forest, naïve bayes, NLP, Machine learning.*

I. INTRODUCTION

People dwell on the internet for different reasons. As data is present in abundance over the internet, one should be cautious that the data is original or not. We share our feelings or information over the internet via audio, video or text. These days a large population is blinded by the technology because of which there are serious consequences of fake news over groups of society.

According to a survey, people living in the USA reside on getting news online than print media. It's important to conserve this data thus, this paper discusses the vulnerability of individuals and the escalation of spread in fake new and also the required mechanism to detect fake news to protect the society.

Fake news spreads faster than the real news so in the proposed system we use a dataset from the dataset obtained from Kaggle. The data is labeled into two categories- real and fake news and then combined as one dataset. This dataset is used to train the machine learning model. In this project, we're trying to build a machine learning model using four different classifiers and using Tf-idf vectorizer. The aim is to predict news which misleads the user and create chaos.

II. RELATED WORKS

There have been several approaches to detect fake news. A lot of research has been conducted by several people to bring some clarity into this field, so that the chaos due to mislead information could be avoided.

- The paper [1] used three classification algorithms and then combined all three to achieve a higher accuracy. The classifying techniques used were SVM and naïve bayes and then they were combined. It was found that the combined model yielded a higher accuracy of about 94% when compared with the other two. The news

authenticator compares the news articles with other news on the internet. If it finds an authentic source, it marks it as true else it is fake.

- [2] involves using a tool designed to detect and eliminate fake news. Websites containing mis-information is flagged as false. The tool should be installed in the system of the user. It uses several classifiers like, logistic regression, naïve bayes, random tree etc, for the purpose. According to them naïve bayes provides the highest accuracy.

- [3] proposes four different models using different classifying techniques like logistic regression, K-nearest neighbor, decision tree and random forest and choses the best amongst them all. According to their result, logistic regression yields the max accuracy of 71%.

- [4] uses three classifying algorithms SVM, naïve bayes, logistic regression and compares the accuracy of all the results. According to them the naïve bayes model with lidstone smoothening yields the maximum accuracy of 83%. It tests the model on a dataset from Kaggle having about 2000 fake articles and 1800 true news articles.

- [5] used naïve bayes model for the purpose of detecting fake news. This simple model yields an accuracy of 75%. The accuracy could be increased by using a different dataset and also by combining several classifying algorithms.

III. METHODOLOGY

The different steps involved in building the proposed system is as follows:

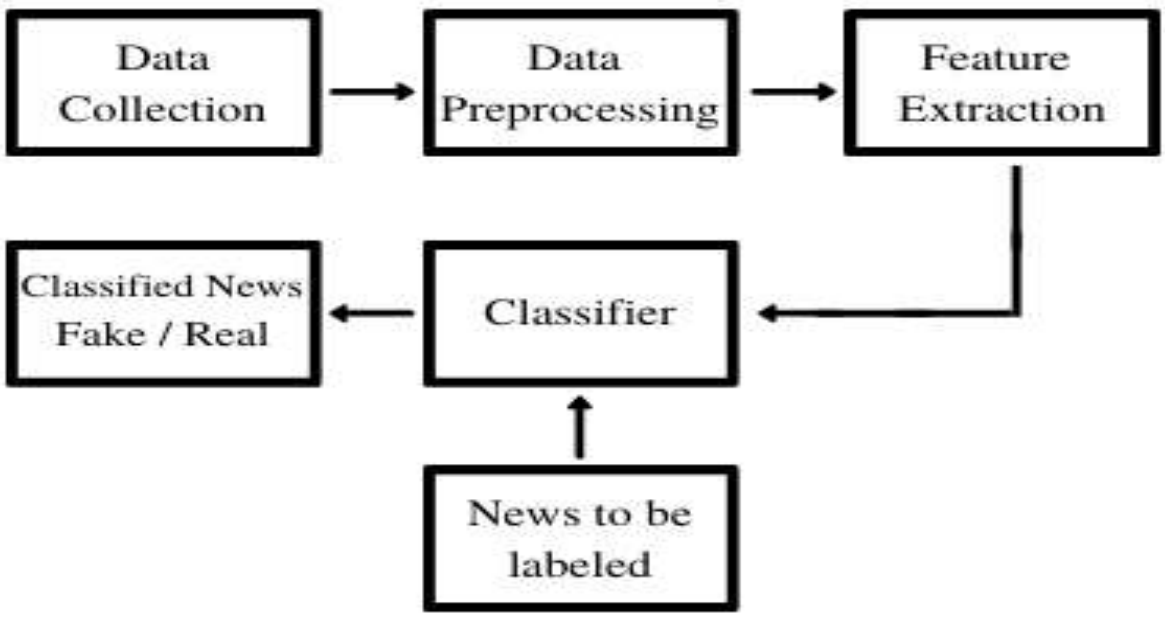

Fig -1: Flow Chart of Detection Model

1. Data Collection

The data is collected from the Kaggle database [6] which consists of labelled data. There are two datasets, one for fake news and one for true news. The dataset consists of around 20000 real news and 20000 fake news. The dataset has headlines, date and body attribute for each article. A label attribute is added to both the datasets. The two datasets are combined retaining only the headline and label attribute.

2. Data Pre-processing

The dataset has to be preprocessed so that the machine learning algorithm can detect patterns easily. All the non-informative words and special characters are removed and the text is converted to lowercase. We lemmatize the text and remove the stop words.

3. Feature Extraction

The data has to be interpreted to analyze the text. The text should be converted into integer or floating-point values and then sent to the machine learning model. The vectorization method used in the given system is Bag of Words method using the TF-IDF vectorizer.

4. Test and train splitting

Machine learning models require two sets of data to work. The data obtained is further divided into two parts for training and testing purpose. 80% of data obtained is used to train the models and the remining 20% data is used to test the accuracy of the models selected.

5. Classification Models

A total of 4 machine learning algorithms are used to predict the fake news in this proposed system. TF-IDF vectorizer is used to convert data into vectors. It is then passed into the model and the listed algorithms are applied to find the best algorithm for this proposed model.

a) Gaussian Naïve-Bayes Classifier:

A Naive Bayes classifier is a machine learning model that is used for data classification. The Gaussian naïve-bayes classifier is a variant of naïve bayes that follow Gaussian normal distribution and supports continuous data. It is a supervised machine learning algorithm and the classifier mainly is based on the Bayes theorem.

The assumption is that the occurrence of one certain element is independent of the occurrence of another element. So, it is called Naïve. Such as if the fish is identified on the basis of size, colour and type of water then medium, glittery and sea water is recognized as a snapper. Hence each feature is considered independent to identify that it is a snapper without considering the other features.

it is a snapper without considering the other features.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

b) Logistic regression:

Logistic regression is a widely used machine learning algorithm extensively used for predicting the probably of a variable. It is a supervised learning classification algorithm.

Here the target variable can have two classes, i.e. it is binary in nature.

It can be used for multiple classification problems like fake news detection, email spam detection, heart disease etc.

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

c) Decision Trees Classification:

Among the other classification algorithms, one of the most commonly used classifiers is Decision Tree classifier. Decision tree classifier is a supervised learning algorithm and also a very powerful classifier. Decision tree classifier can perform both classification and regression like the support vector machines. All the possible solutions to a decision are graphically represented.

It is easy to understand as it uses tree analysis to classify the data. The data is broken into smaller parts and the decision tree is built. Decision trees support both categorical data and numeric data.

d) Random Forest Classification:

Random forest classification is a group of decision trees from a subset of randomly selected training data set. It combines the weight from each decision tree to find the final test object.

The random forest classification is based on ensemble learning. It is a type of learning where you join different types of algorithms multiple times to form a more powerful algorithm which can give higher accuracy. It

combines multiple trees to form a forest hence the name random forest to give higher accuracy to the proposed model.

IV. RESULTS

After testing of all the algorithms, i.e., Gaussian naive bayes, logistic regression, random forest and decision tree we obtained the following accuracy.

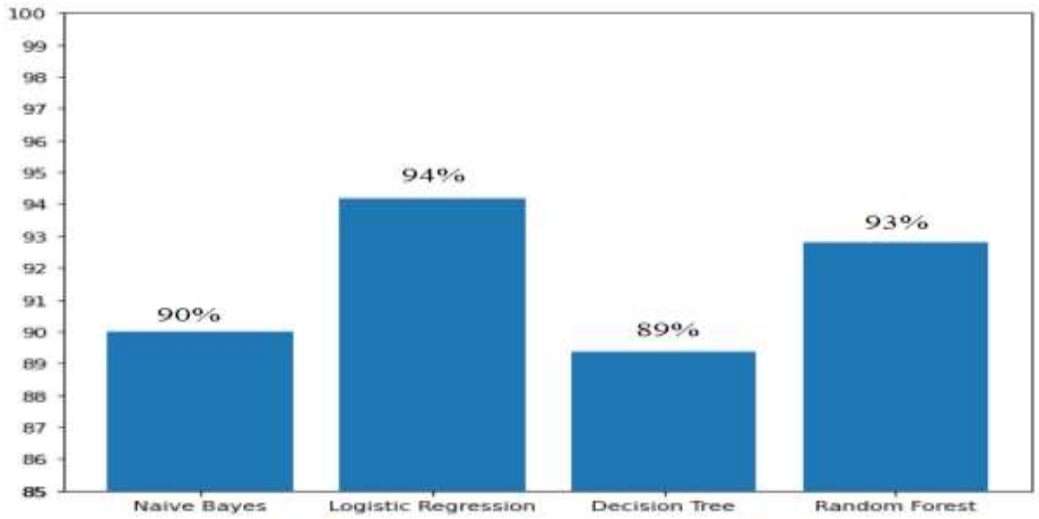

Fig 2: Accuracy comparison between the four classifying models

Amongst the four approaches or classifying algorithms used, Logistic regression gives the maximum accuracy of 94%.

Decision tree model gives an accuracy of 89% which is lowest amongst the four. Random forest and Gaussian naïve bayes give an accuracy of 93% and 90% respectively.

V. CONCLUSION

To tackle the increasing false information on the internet, the machine learning model created distinguishes an input as real news or fake news. A lot of social media sites like WhatsApp orFacebook are trying to implement such systems into their system to prevent the spread of fake news.

Amongst the four approaches or classifying algorithms used, logistic regression gives the best accuracy.

All the models can predict the accuracy of the news to a good extent.

VI. REFERENCES

[1] Anjali Jain, Avinash Shakya, Harsh Khatter, Amit Kumar, "A Smart System For Fake News Detection Using Machine Learning"

https://ieeexplore.ieee.org/document/8977659

[2] Monther Aldwairi, Ali Alwahedi, "Detecting Fake News in Social Media Networks"

https://www.sciencedirect.com/science/article/pii/S1877050918318210

[3] Vanya Tiwari, Ruth G. Lennon, Thomas Dowling, "Not Everything You Read Is True! Fake News Detection using Machine learning Algorithms"

https://ieeexplore.ieee.org/document/9180206

[4] Kushal Agarwalla, Shubham Nandan, "Fake News Detection using Machine Learning and Natural Language Processing"

https://www.ijrte.org/wp-content/uploads/papers/v7i6/F2457037619.pdf

[5] Mykhailo Granik; Volodymyr Mesyura, "Fake news detection using naive Bayes classifier"

https://ieeexplore.ieee.org/document/8100379

[6] Clément Bisaillon, "Fake and real news dataset"

https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

[7] A.Lakshmanarao, Y.Swathi, T. Srinivasa Ravi Kiran, "An Effecient Fake News Detection System Using Machine Learning"

https://www.ijitee.org/wp-content/uploads/papers/v8i10/J94530881019.pdf

[8] N. Smitha, R. Bharath, "Performance Comparison of Machine Learning Classifiers for Fake News Detection"

https://ieeexplore.ieee.org/document/9183072