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

Analysis of Various Machine Learning Models for Detecting Fake News

Tini Thomas[1], Praseetha S Nair[2]

[1]*Student, Dept. of CSE, Musaliar College of Engineering and Technology, Pathanamthitta, Kerala*
[2]*Assistant Professor, Dept. of CSE Musaliar College of Engineering and Technology, Pathanamthitta, Kerala*

Abstract - *Recently, there has been a sudden increase in fake news due to the immense use of social media and online news media. It has become much easier to spread fake news than earlier. This type of fake news when spread may cause a severe impact. Hence it is very essential to take certain measures to reduce or differentiate between real and fake news. In this paper, we presents an analysis of fake news detection using various classifiers like Naive Bayes, Support Vector Machine (SVM), Logistic regression and decision tree classifier.*

***Key Words*: fake news, machine learning, naïve Bayes, support vector machine.**

1. INTRODUCTION

Traditionally people got news from various trusted sources, but nowadays people get news from online sources and social media which now has become a trend, and most of the time it becomes difficult to decide whether stories are authentic or not. Social media sites can have a major influence in expanding fake news.

Fake news is a news created to intentionally misguide or mislead readers. Ever since the birth of social media and online news media, the spread of fake news has increased drastically. Social media sites such as Facebook, Whatsapp, twitter etc. are one of the biggest sources of spreading fake news. This paper provides a survey on recent past research papers done on this particular region and also proposes a model that can predict whether the news is real or fake using NLP techniques and Machine Learning. The model will be using the URL as an input that not only validates the headline but also validates the behavior of the site and other related parameters that can provide an improvement over past research.

For detection whether the news is fake or not we are implementing a fake news detection system and comparing the accuracy of the system using various machine learning models. This system consists of admin, moderator, NLP actions and web scraping. Then machine learning model creation and prediction is done.

2. METHODOLOGY

2.1 Data collection

Here dataset creation was done by the moderator. The moderator adds news and label them manually. The moderators are actually the employees set by the admin or the company. Moderator adds the labels into csv files and store them.

2.2 Model creation and Feature Extraction

2.2.1 Text Analysis Operations using NLTK

NLTK may be a powerful Python package that gives a group of diverse natural languages algorithms. It is free, open source, easy to use, large community, and well documented. NLTK consists of the foremost common algorithms like tokenizing, part-of-speech tagging, stemming, sentiment analysis, topic segmentation, and named entity recognition. NLTK helps the system to analysis, pre-process, and understand the transcription.

2.2.2 Tokenization

Tokenization is the first step in text analytics. The process of breaking down a text paragraph into smaller chunks like words or sentence is named Tokenization. Token may be a single entity that's building blocks for sentence or paragraph.

- Sentence Tokenization-Sentence tokenizer breaks text paragraph into sentences.
- Word Tokenization-Word tokenizer breaks text paragraph into words.

2.2.3 Lexicon Normalization

Lexicon normalization considers another type of noise in the text. For example, connection, connected, connecting word reduced to a standard word "connect". It reduces

derivationally related forms of a word to a common root word.

2.2.4 Stemming

Stemming is a process of linguistic normalization, which reduces words to their word root word or chops off the derivational affixes. For example, connection, connected, connecting word reduce to a common word "connect".

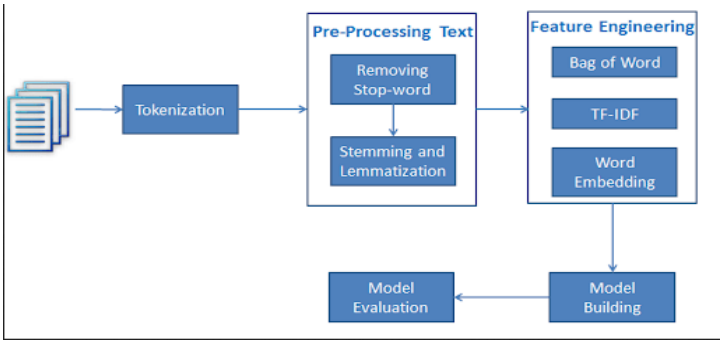
2.2.5 Lemmatization

Lemmatization reduces words to their base word, which is linguistically correct lemmas. It transforms root word with the utilization of vocabulary and morphological analysis. Lemmatization is usually more sophisticated than stemming. Stemmer works on an individual word without knowledge of the context. For example, The word "better" has "good" as its lemma. This thing will miss by stemming because it requires a dictionary look-up.

Text needs to be converted into numbers before it is used with a machine learning algorithm. In this dataset we have done feature extraction and selection methods from scikit and python. To perform this feature selection, we use TF – IDF method. We have also used word to vector to extract the features, also pipelining has been used to ease the code. TF - IDF score is then attached to each document.

- Term Frequency: A summary for the frequency of each word in the document.
- Inverse Document Frequency: Reduces the importance of words that are repeated frequently across documents.

Fig 1- Text analysis using NLTK



2.3 Classification

Classification could be achieved using machine learning algorithms like Decision Trees, Naive Bayes, Logistic Regression, and Support Vector Machines.

3. CLASSIFIERS

3.1 Naive Bayes Classifier

Naive Bayes classifiers are a set of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms that share a common standard principle, i.e. every pair of features being classified is independent of each other. Naive Bayes classifiers assume strong, or naive, independence between attributes of knowledge points. Naive Bayes assumes the independence of features which is not true for news articles as words usually come together.

Equation:

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

Using Bayes theorem, we will find the probability of an event as longs as **B** has occurred. Here, **B** is that evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of single particular feature does not affect the other. Hence it is called naive.

3.1.1 Types of Naive Bayes Classifier

- Multinomial Naive Bayes:

This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors employed by the classifier are the frequency of the words present within the document.

- Bernoulli Naive Bayes:

This is almost like the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the category variable take up only values yes or no, for instance if a word occurs within the text or not.

- Gaussian Naive Bayes:

When the predictors take up endless value and aren't discrete, we assume that these values are sampled from a gaussian distribution.

3.2 Logistic Regression

Logistic regression is a type of statistical analysis that is used to predict the outcome of a dependent variable based on prior observations. It is mainly used for binary classification. A logistic regression algorithm looks at the connection between a dependent variable and one or more dependent variables. Logistic regression needs a large data set for higher accuracy while Naive Bayes can work with small datasets too.

3.3 Decision Trees

Decision Tree solves the problem of machine learning by transforming the information into tree representation. Each internal node of the tree representation denotes an attribute and every leaf node denotes as class label. Decision tree algorithm are often used for solving both regression and classification problems. A decision tree would overfit when there are a large number of sparse features, and therefore perform poorly on the testing data.

3.4 Support Vector Machines

SVM is also a supervised machine learning algorithm that is oftently used for both classification and regression purposes. The support vector machine performs well when there is an extremely high number of features in a text classification problem but generally requires a lot of tuning and is memory intensive. Once we have labelled training data (supervised learning), the algorithm generates the best possible hyperplane which categorizes new data automatically.

4. EVALUATION METRICS

To evaluate the performance of algorithms for fake news detection problem, various evaluation metrics are used. Here, we review the most widely used metrics for fake news detection. Most of the existing models consider the fake news problem as a classification problem that predicts whether a news is fake or not:

- True Positive (TP): when predicted fake news article or report are actually annotated as fake news.
- True Negative (TN): when predicted true news article or report are actually annotated as true news.
- False Negative (FN): when predicted true news article or report are actually annotated as fake news.
- False Positive (FP): when predicted fake news article or report are actually annotated as true news.

Fig 2- confusion matrix

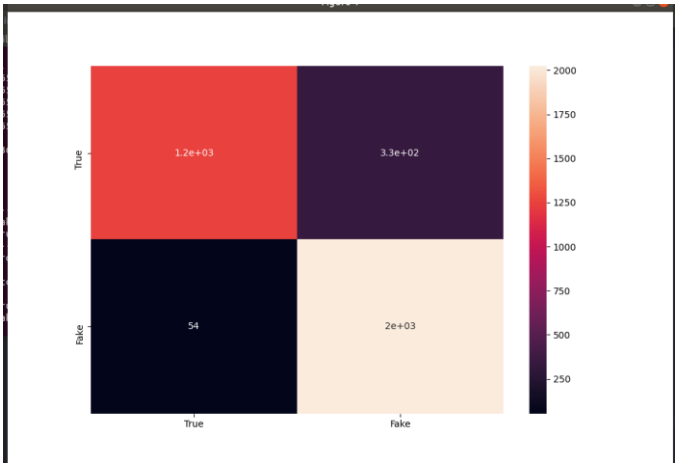


Table -1: Comparison of the accuracy of fake news classification using different classifiers

Algorithm	Naïve Bayes	Logistic Regression	Decision Tree	SVM
Accuracy	90%	85%	82%	89%

5. CONCLUSIONS AND RESULTS

Using the above algorithms, we get an accuracy value of 90% for naïve bayes,85% accurate value while using logistic regression algorithm,82%and 89% value when applying both decision tree and SVM algorithm respectively ie, shown in the table-1.

From this, we can conclude that naïve bayes algorithm is better for this fake news detection model, because it gives

90% accurate results .Other algorithms like logistic regreesion, decision tree and svm gives less accurate values. While comparing both SVM and naïve bayes ,there are slight variations in their values.

REFERENCES

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

[2] 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) ISSN: 2277-3878, Volume-7, Issue-6, March 2019

[3] Sahil Gaonkar, Sachin Itagi, Rhethiqe Chalippatt," Detection Of Online Fake News : A Survey "2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)

[4] Karishnu Poddar, Geraldine Bessie Amali D, Umadevi K S," Comparison of Various Machine Learning Models for Accurate Detection of Fake News", 2019 Innovations in Power and Advanced Computing Technology (i-PACT)