Fake News Detection using Machine Learning

Pavan M N[1], Pranav R Prasad[2], Tejas Gowda[3], Vibhakar TS[4], Dr. Sushila Shidnal[5]

[1-4]Student, Dept. of Computer Science Engineering, Sir MVIT, Karnataka, India
[5]Professor, Dept. of Computer Science Engineering, Sir MVIT, Karnataka, India

Abstract - *Newspapers are the primary source of news for people worldwide. However, off late, due to the significant growth and updates in technologies, there has been a stupendous rise in the popularity of social media. The number of people who use social media has increased remarkably. As a consequence, social networks such as social media, websites, blogs, etc. have emerged as relevant platforms to gather all kinds of news. People rely more on social networks than newspapers these days. With the availability of the internet, these networks can be accessed easily. This can lead to easy manipulation of the existing news, thereby causing fake news. Fake news can be used as a vital tool to project people in a wrong way. It can spread hate among people which can further harm the society. Hence, it is very necessary to prevent the spread of fake news. This survey paper describes the various methods and models used for the detection of fake news. Our project aims to use Natural Language Processing to directly detect fake news, based on the text content of news articles.*

***Key Words*: Fake news, Fake news detection, Machine learning, Natural Language Processing, Python, Corpus, Classifier, Dataset.**

1. INTRODUCTION

In the modern era, the spread of fake news has become very evident. Fake news is being used for both economic and political benefits. The need of the hour is to prevent the spread of fake news. The first thing that needs to be done to achieve this is to detect fake news.

Our project aims to develop a machine learning program to identify when a news source may be producing fake news. We use a corpus of labelled real and fake articles to build a classifier that can make decisions about information based on the content from the corpus. Our model focuses on identifying sources of fake news, based on multiple articles originating from a source. Once a source is labelled as a producer of fake news, we predict that all future articles from the same source are also a producer of fake news.

The intended application of our project is to assist in applying visibility weights in social media. Social networks can make use of the weights produced by the model to obscure stories that are highly likely to be fake news.

2. LITERATURE SURVEY

Rohit Kumar Kaliyar (2018) proposed Fake News Detection Using A Deep Neural Network. In this proposed system they have used Natural language processing, Machine learning and Deep learning techniques to implement this model and compare which model will give more accurate results. They utilized the DGX1 nvidia computer to get accurate results and divide the dataset into real and fake news. This model contains varieties of machine learning algorithms such as K nearest neighbour (KNN), Naive bayes (NB), Random forest (RF), Decision tree (DT). They have investigated deep learning models such as Shallow convolution network (SCN) and also Very deep convolutional network (VDCN) and gated networks such as Gated recurrent Unit (GRU) with the help of Convolution Network and Long short-term memory (CN-LSTM). They additionally investigated the adequacy of word embeddings and also, word2vec highlights with Deep neural networks. This model utilizes chi2 for inclusion in the Machine learning model to produce more accurate results.

Wenlin Han (2019) proposed Fake news detection in social networks using machine learning and deep learning performance evaluation. In this proposed system they have referred to some traditional machine learning approaches such as Deception modeling, clustering, Naive bayes are evaluated for accuracy detection TF-IDF and PCFG with Convolution neural network and Recurrent neural network models are assessed to think about the execution with conventional machine learning techniques. Deep Modeling speaks in high space dimensional text and computer algorithms that need extracted text to display in a reasonable way. Fake information and fake articles have a great measure of normal properties so to characterize them Naive Bayes classification is applied. The accuracy of the system is utilized using diverse methods such as bigram recurrence which is utilized by TF-IDF and PCFG with the combination of both methods. The CNN and RNN models are referred for text mining or picture recognition also RNN for time and arrangement based expectations.

Ranojoy Barua (2019) proposed An Application for Fake News Article Detection using Machine Learning Techniques. This system is developed based on machine learning approaches such as Long short term (LSTM) and Gated recurrent unit (GRU) to characterize a information into spam or original. The trial results on the dataset arranged in this work looks accurate. The model is additionally prepared and tried on other data and similar outcomes show the productivity of the proposed model. To detect the working of FNAD model. It is tried and utilized on 1,000 text information in FNAD's data to comprehend the execution of the model they have utilized estimation measurements for example, Confusion model, Score model and Accuracy model. Execution results of the proposed model using FNAD produces 80.2% accuracy. Hence it can be utilized to confirm text information from different sources of the web prior to tolerating it as a reality.

Chaitra K Hiramath (2019) proposed Fake News Detection Using Deep Learning Techniques in this model they have utilized classification algorithms for dividing fake and real information depending upon news sources. The algorithms utilized by this model are Support vector machine (SVM), Random forest (RF) and Deep Neural Network (DNN). The memory examination of different calculations shows that Logistic Regression needs less amount of memory space and the time examination of algorithms shows that Deep neural networks require 400 ms which is less time. The Accuracy calculation shows that Deep neural network has 91% accuracy which is greater than other four algorithms. So by using regression, Support vector machines (SVM), Random forest, Deep neural network (DNN) classification techniques it is easy, faster and accurate to classify fake news from a huge amount of datasets.

Karishnu Poddar (2019) proposed Comparison of Various Machine Learning Models for Accurate Detection of Fake News. The main aim of this model is to handle the problem of fake information classification. With the scores of distinctive vectorizers to be specific check and Term frequency inverse- document formatting (TF-IDF) is contrasted to locate the vectorizers. English stop words are utilized for improving the score and accuracy. Different classifications are used : Regression, Support Vector Machine (SVM) and Decision tree classifiers are utilized for precise fake source identification. The reenactment data show that the Support vector Machine and TF-IDF will give the most accurate precise prediction. The famous method is to detect word count using the TF-IDF method and then the calculations of TF-IDF are joined to each word document. The Text preprocessor converts numbers to text before and utilized with TF-IDF for identification of fake text.

Sahil Gaonkar (2019) proposed Detection Of Online Fake News A Survey. This study includes a model that groups inconsistent information into genuine or fake news with registering a score based on different information received from URL. This survey utilizes Machine learning strategies like Linear Regression, Logistic Regression, Support Vector Machine (SVM) with the help of Multilayer perceptron(MP) to obtain more accurate data. All data that are acquired are then added to obtain the last score and classify the information as fake or real. From this survey it is seen that the given data preprocessed using preprocessing strategies like tokenizer and stemming. TF-IDF uses Context free grammar (CFG) for further classification of news. From analyzing the survey we can state that identifying fake information in social sites is easier than any other online platform.

Syed Ishfaq Manzoor (2019) proposed Fake News Detection. The kinds of Fake News depicted are based on Visual, User, Knowledge, Style, Stanze. They proposed different methods of Fake News Detection such as Linguistic-Basis, Clustering, Predictive, Content-Cue-Based, Non-Text-Cue-Based modelling. Algorithms considered for the proposal were Naive Bayes, Decision Trees, SVM, Neural Networks, Random Forest, XG-Boost. Many methods for the more accurate results of detection of Fake News such as, Convolutional Neural Networks, Deep Boltzmann Machine, Deep Neural Network and Deep Autoencoder Model were applied on this which were handy and yielded more accuracy in recognising whether the News is mis-leading or Real.

Anjali Jain (2019) proposed a Brilliant System for Detection of Fake News of which they reported about Stopping of Fake News by Facebook and Whatsapp too. In the proposed model they talk on how Aggregators, News Authenticators, News Suggestion or Recommendation System help in promoting the Fake News based on User interest. The proposed system was checked and evaluated by different implementation methods such as NLP, Naive Bayes, CNN, "Naive Bayes SVM NLP" and found the accuracy as 76%, 74%, 87%, 94% respectively and they talked about each and every implementations and its evaluation results and System Architectures, SVMs' etc. And as of high accuracy they finalized to use SVM, NLP with Naive Bayes Algorithm implementation which yielded the highest accuracy on their crisis.

Rohit Kumar Kaliyar (2019) proposed a Multi-class Counterfeit News Detection utilizing Ensemble Machine Learning of which they report about the use of Naive Bayes Algorithm and also, k-NN calculation which is a non-parametric strategy utilized for characterization and relapse and Decision Trees as well. They examined all aspects and tried to implement Gradient Boosting Machine which is a High-performance ML Algorithm which has significant achievement in the field of Machine Learning, which could yield around 86% accuracy and Confidence score of about 76%, And as number of iterations increase the training losses decrease iteratively. And as the number of iterations increases the time required will relatively decrease. These characteristics make the Gradient Boosting algorithm opt for the system.

Smitha.N (2020) proposed ML classifiers for Fake news recognition and its accuracy. In this proposed system they have referred to Count-Vectorizer, TF-IDF Vectorizer, Word Embedding for the calculation of best accuracy and best performance. By utilizing classification algorithms the accuracy obtained with SVM linear classification algorithm with TF-IDF Vectorizer feature extraction is of 94% accuracy. They concluded that use of Neural networks has similar accurate results than SVM linear classification algorithms. As using Neural networks with linear classification algorithms, would make the classification more and more complex, Hence they opted the method of SVM over Neural Networks which is less complex. The first stage in the proposed system is text collection referred to datasets. Second step is text pre-processing in which conversion steps are included. Third is Feature extraction which includes encoding of words. Fourth is Classifiers of News which includes SVM, Logistic Regression, Decision Trees etc .Classification with TF-IDF Vectorizer yield more accurate results than Count-Vectorizer and Word Embedding process, Therefore they opted for TF-IDF Vectorizer for Classification.

Rahul R Mandical (2020) proposed implementation of Detection of Fake news using Machine Learning by use of Naive Bayes and Passive Aggressive classifiers. These classifiers are used to feed to models based on TF-IDF Vectorizer where TF-IDF stands for Term Frequency-Inverse Document Frequency of which value increases with increment in number of times the word shows up in the document, but there is a loss in semantic meaning of words. The Passive Aggressive Classifier is a Linear based model, In this they had chosen to utilize Tokenizers for inputting into Deep Neural Network Models (DNN). They thought about Keras which vectorizes and converts text corpus into vectors or sequence of integers which is passed to the mac pool layer to calculate a single maximum value for each of input channels. After the examination of classifiers it was seen that Naive Bayes and Passive-Aggressive Classifiers with DNN have outperformed in most of the Datasets. Therefore they opted for the DNN with these mentioned classifiers.

Vanya Tiwary (2020) proposed classification of News headlines using Machine Learning, Clickbait, Propaganda, Comment or Opinion and Satire or Humor are the classified types of Fake News. In this they have referred to Count Vectorizer, TF-IDF Vectorizer, Hashing Vectorizer. They proposed the implementation of Logical Regression, Decision Tree, Random Forest with different Vectorizers which gave the result that use of Logistic Regression algorithm with TF-IDF Vectorizer has more accurate results when compared to other algorithms with TF-IDF Vectorizer. Over 73% of headline Fake News was accurately detected and classified, whereas Decision Tree was accurate upto 62% and Random Forest upto 66% respectively. The research done by them was successful in classifying the HeadLines of News. By this proposal they figured out that a Logical Regression algorithm with TF-IDF Vectorizer would yield the best results for the classification of Headline in the News.

3. METHODOLOGY

The performance comparison of Machine learning classifiers involved different stages in the required system such as Text collection, Text preprocessing, Feature extraction and the News to be classified. Further,the feature extraction included several constraints like Bag of words, Countvectorizer, TF-IDF and Word embedding using Spacy. The Classifiers further included the constraints like Support vector machine, Logistic regression, Decision trees, Random forest, Gradient boosting and XG-Boost.In Identification of Fake news using Machine learning, we have used ML to develop our devices. Naive bayes model assumes that the features are independent of one another statistically. The Naive Bayes model is famous for its multi-class prediction, it was selected for its case and dynamic nature of predicting the nature of the text. We opted to use tokenizers for feeding into deep neural network models. For datasets with just one attribute, we used sequential DNN and complex networks were developed with the help of functional DNN.

Support vector machine (SVM) is a good device to derive the binary class based on the information given to the model. Its function is to characterise the article into two categories (true or false). SVM is an advanced machine learning algorithm that can be used for different purposes like regression and clarification. It is based on the principle of searching the hyperplanes that best classify the dataset into two classes. SVM has the potential to deal with high dimensional spaces and will be memory proficient.

The performance evaluation of trained models are computed based on three feature sets i.e., a)Normalized frequency of parsed syntactic dependencies, b)Bi-gram term frequency - Inverse document frequency and union of a) and b). Spacy is executed in Cython , which is another set of Python language that allows C code to be used in Python by using Python/C APL. Spacy gives better yield rate for performance especially on tasks like entity recognition, parsing and so on.

A neural network based model is proposed which is trained under a very huge number of training samples. We call this a Fake News Article Detector (F-NAD) model, which acts as a binary classifier between fake and real news segments. A total of approximately 45k training samples is generated based on approximately 7.5k news articles downloaded from the Internet. Once the model is trained correctly, the test news article is fed to the device to determine whether it is fake or real.

Content information requires processing to implement AI on them. Stemming technique is used to remove suffixes or prefixes from a word. There is a huge amount of data stored in file but it can't be accessed by computer assisted analysis. NLP allows the analysts to find some crucial data. The Data should be in the form of preprocessing. It operates with expulsion of punctuations, URL's images, stemming and stop words. We then classify that information utilizing classifiers (for eg. LR, SVM, NB, RF and DNN).

The dataset was fetched from Kaggle and has been said as the initial step towards the classification of fake news. The content and metadata has been taken from 244 websites that are said to be associated with fake news by the BS Detector Chrome Extension by Daniel Sieradski. It consists of about 13000 posts over a period of 30 days. It consists of about 150000 articles that were mostly published between the period of 2016 and 2017 by RSS feeds and website homepages.

Probability Context-Free Grammars (PCFG) is used in deep syntax analysis. The rhetoric relation in linguistic elements can be found using the Rhetorical Structure Theory (RST) frame. Data analysis consists of article credibility analysis with textual content, creator-article publishing historical records, subject and creator credibility analysis. Social spam such as ads are often used by the spammers to make a profit by drawing the users to visit the site.

The deep neural networks include several platforms and technologies like Jupyter notebook, Nvidia DGX, Dataset collection. It also consists of important functions in CNN as Rectified linear unit (ReLU), Pooling and Padding. It also has features like Term frequency, Inverse document frequency, Confusion matrix, Hashing-vectorizer, Precision, Accuracy, Recall and FI-SCORE. The Naive bayes device has many devices like Decision trees, Random forest, CNN and LSTM.

Table -1: Comparison of Accuracy Levels of Different Models

Different Models	Accuracy Levels
Neural Network with TF	81%
Neural Network with Keras	93%
Naive-Bayes	73%
SVM	88%
LSTM	95%

4. CONCLUSION

We have discussed the various ways, methods and models of fake news detection. We have addressed the benefits and shortcomings of each model. We have tabulated the comparison of accuracy levels of all the models. We plan to overcome the shortcomings of these models through our project. We will make use of two datasets, a full training dataset and a testing training data set, associated with news articles. These datasets are obtained from Kaggle. We will use Natural Language Processing to detect fake news directly, on the basis of the text content of the articles.

REFERENCES

[1] Rohit Kumar Kaliyar, "Fake News Detection Using A Deep Neural Network", IEEE 2018.

[2] Wenlin Han and Varshil Mehta, "Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation", IEEE 2019.

[3] Ranojoy Barua, Rajdeep Maity, Dipankar Minj, Taranag Barua and Ashish Kumar Layek, "F-NAD: An Application for Fake News Article Detection using Machine Learning Techniques", IEEE 2019.

[4] Chaithra K Hiramath and Prof. G.C Deshpande, "Fake News Detection Using Deep Learning Techniques", IEEE 2019.

[5] Karishnu Poddar, Geraldine Bessie Amali D and Umadevi K S, "Comparison of Various Machine Learning Models for Accurate Detection of Fake News", IEEE 2019.

[6] Sahil Gaonkar, Sachin Itagi and Rhetiqe Chalippatt, "Detection Of Online Fake News: A Survey", IEEE 2019.

[7] Syed Ishfaq Manzoor and Dr Jimmy Singla, Nikita, "Fake News Detection Using Machine Learning approaches: A systematic Review", IEEE 2019.

[8] Anjali Jain, Avinash Shakya, Harsh Khatter and Amit Kumar Gupta, "A Smart System For Fake News Detection Using Machine Learning", IEEE 2019.

[9] Rohit Kumar Kaliyar, Anurag Goswami and Pratik Narang, "Multiclass Fake News Detection using Ensemble Machine Learning", IEEE 2019.

[10] Smitha. N and Bharath. R, "Performance Comparison of Machine Learning Classifiers for Fake News Detection", IEEE 2020.

[11] Rahul R Mandical, Mamatha N, Shivakumar N, Monica R and Krishna AN, "Identification of Fake News Using Machine Learning", IEEE 2020.

[12] Vanya Tiwari, Ruth G.Lennon and Thomas Dowling, "Fake News Detection using Machine learning Algorithms", IEEE 2020.

[13] Kai Shu, Amy Silva, Suhang Wang, Jiliang Tang and Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective".

[14] Sohan Mone, Devyani Choudhary and Ayush Singhania, "FAKE NEWS IDENTIFICATION CS 229: MACHINE LEARNING : GROUP 621".