Survey on Fake News Detection using Machine learning Algorithms

Dr. S. Rama Krishna
CSE Dept.
Bapatla Engineering College,
Bapatla, India

Dr. S. V. Vasantha
IT Dept.
M. V. S. R. Engineering College
Hyderabad, India

K. Mani Deep
CSE Dept.
Bapatla Engineering College,
Bapatla, India

Abstract— Due to easy access, rapid growth, and proliferation of the information available through regular news mediums or social media, it is becoming easy for people to look for news and consume it. These days a lot of information is being shared over social media and we are not able to differentiate between which information is Fake and which is legitimacy. For publishing a news in social media the cost is low, easy access. The extension spread of fake news has the potential for extremely negative impact on individuals and society. The goal of this project is to create an efficient machine learning algorithm for identifying the fake news..

Keywords:- Fake news detection; Machine Learning; Deep Learning; Artificial Intelligence

I. INTRODUCTION

In these days seeking new information or news has been directed to online social media services instead of the traditional information sources like newspapers and radio. The reasons that account for these changes are simple to define, first compared to the traditional mediums social media is less time consuming and less expensive and second, it has made sharing, putting forward individual's opinion straight-forward, just a single click. Indeed, social networks are used for multiple information and business purposes, apart from acting as an interacting tool. They are the major competitors for other media outlets such as newspapers, radio and television.
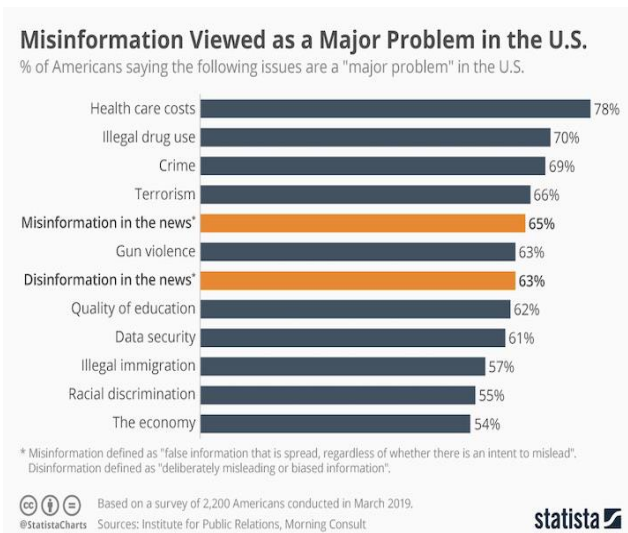

Fig 1: Survey report on misinformation impact in U.S. [21]

Figure1 explains the impact of the misinformation and disinformation in the United States of America. From the figure we can observe that fake news is the one of the top 10 significant impacted issues in the mentioned country.

Social media systems have been dramatically changing the way news is produced, disseminated, and consumed, opening unforeseen opportunities, but also creating complex challenges. A key problem today is that social media has become a place for campaigns of misinformation that affect the credibility of the entire news ecosystem. A unique characteristic of news on social media is that anyone can register as a news publisher without any cost. Corporations are increasingly migrating to social media. Along with this transition, not surprisingly, there are growing concerns about fake news publishers posting "fake" news stories, and often using "fake" followers. As the extensive spread of fake news can have a serious negative impact on individuals and society.

II. LITERATURE SURVEY

There are several algorithms for detecting the fake news. For that we analyse through different classifiers in different research papers. The classifiers are Random Forest, CNN, SVM, KNN, Logistic Regression, Naive Bayes, Long Short Term memory and SGD. The accuracy obtained by using Random forest is 83 %, the accuracy obtained by using CNN is 97%, the accuracy obtained by using SVM is 94%, the accuracy obtained by using KNN is 79%, the accuracy obtained by using Logistic Regression is 97%, the accuracy obtained by using Naive Bayes is 90%, the accuracy obtained by using Long Short term memory is 97%, the accuracy obtaining by using the combination of SVM &NB is 78% and the accuracy obtained by using SGD is 77.2%. Compared to all CNN,LR and LSTM obtains high accuracy.

A. Random Forest:

It is a combination of decision trees. Here each tree will build a random subset of a training dataset. In each decision tree model, a random subset of variables is used to partition the data set at each node. Bhavika, Bhutani, Neha, Rastogi, Priyanshu, Sehgal, Archana and Purwar implemented a Sentiment Analysis technique for Fake news detection [1]. They used LIAR, George McIntire, Merged Datasets and they classified those datasets by using Random Forest, Naive Bayes classifiers. For detecting the fake news they proposed a new solution by taking Sentiment as an important feature to

Special Issue - 2021

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
ICACT – 2021 Conference Proceedings

improve the accuracy. In their approach the accuracy obtained by using TF-IDF with Cosine Similarity is better as compared to without using Cosine similarity. By using Random Forest they gain the maximum accuracy of 83%.

B. Convolutional neural networks

Rohit Kumar Kaliyar developed A Deep Neural Network techniques for fake news detection [2]. He used Fake or Real News Dataset for detecting the fake news by classifying with Convolutional Neural Networks, Long Short term memory, Naive Bayes, Decision Tree, Random Forest and K-Nearest Neighbour techniques. In this by increasing the depth of the network the accuracy is increased when using the CNN method. In this by using k-nearest neighbour algorithm the accuracy is decreased and also precision, recall, f1-score values are reduced. In this he gained maximum accuracy of 91.3% by using CNN algorithm. Belhakimi Mohamed Amine, Ahlem Drif, Silvia Giordano developed a Merging deep learning model for fake news detection [3]. They collected data from Kaggle:https://www.kaggle.com/c/fake-news for detecting the fake news by CNN with text only, CNN with text + title, CNN with text + author. They used layers present in convolution network for data preprocessing and feature extraction. The merged CNN reached accuracy of 96%. V. M. Krešňáková et al developed Deep learning methods for Fake News detection [4]. They used dataset available on https://www.kaggle.com/c/fake-news/data for detecting fake news by Feed forward neural network, CNN with one convolutional layer, CNN with more convolutional layer and LSTM. They used word embedding technique for data preprocessing. CNN with more convolutional layer achieves 97.15% accuracy.

C. Support Vector Machine

Tayyaba Rasool, Wasi Haider Butt, Arslan Shaukat and M. Usman Akram developed a Multi-layered Supervised Learning for Multi-Label Fake News Detection [5]. They used LIAR Dataset for detecting the fake news. They used Holdout testing using SVM, Testing on Test set using SVM, Cross Validation using SVM evaluation methods. By using multiple machine learning algorithms and hold-out, testing and cross validation for evaluation they gain an accuracy of 39.5%. Nicollas R. de Oliveira, Dianne S.V.Medeiros and Diogo M.F.Mattos developed a model for fake news detection [6]. They collected a fake news dataset from Boatos.org for detecting the fake news. In this they use Support Vector Machine for detecting the fake news. In this they use PCA and LSA methods for feature extraction. By using Support Vector Machine they obtained an accuracy of 86%. Karishnu Poddar, Geraldine Bessie Amali D, Umadevi K S developed a Comparison of Various Machine Learning Models for Accurate Detection of Fake News [7]. They used fake news dataset from kaggle.com for detection of fake news by Naive Bayes' Classifier, Logistic Regression, Decision Trees, Support Vector Machines and Artificial Neural Networks. They used Count Vectorizer and TF-IDF Vectorizer feature extraction methods. SVM shows better results with TF-IDF with 92.8% accuracy. Logistic regression performs equally for both count vectorizer and TF-IDF with accuracies of 91.6 and

91.0. Smitha, Bharat developed a Performance Comparison of Machine Learning Classifier for Fake News Detection [8]. They collected dataset from https://www.kaggle.com/mrisdal/fake-news. They use Word Embedding for preprocessing of data. They use Count vectorizer and TF-IDF for feature extraction. In their paper they used classifiers are Support Vector Machine, Logistic Regression, Decision Trees, Random Forest, XG-Boost, Gradient Boosting Neural Network for classifying the news as fake or real. By using classification algorithm the highest accuracy obtained is with SVM Linear classification algorithm with TF-IDF feature extraction with 94% accuracy.

D. K-Nearest Neighbour

Ankit Kesarwani, Sudakar Singh Chauhan and Anil Ramachandran Nair developed a K-Nearest Neighbour Classifier technique for Fake News Detection on Social Media [9]. In this they use Buzz Feed news. It contains the information about the Facebook news. In this the model has achieved maximum accuracy when the value of K taken between15 to 20. In this they gain the maximum accuracy of 79% tested against Facebook news dataset.

E. Logistic Regression

It is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes and success) or 0 (no and failure). Uma Sharma, Sidarth Saran, Shankar M. Patil developed a Fake News Detection using Machine Learning Algorithms [10]. They used liar dataset for detecting if fake news by Naive Bayes Classifier, Logistic Regression, Random Forest. They used Bag-Of-Words, N-Grams, TF-IDF. Logistic regression shows better results with accuracy of 65%. Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf and Muhammad Ovais Ahmad implemented a model for fake news detection on social media [11]. They worked on Logistic Regression, SVM and KNN models using social media and fake news datasets. LIWC method is used for feature extraction. On their experiment they found that Logistic Regression shows high accuracy 91% compared to SVM 67% and KNN-68%. Vanya Tiwari, Ruth G. Lennon and Thomas Dowling developed some Machine learning Algorithms for Fake news detection [12].

In this they use factcheck.csv, fake or real news.csv datasets for detecting the fake news by using Logistic Regression, K-Nearest Neighbour, Decision Tree and Random Forest classification techniques. In this they use Count Vectorizer, term frequency and inverse document frequency Vectorizer, Hashing Techniques. The logistic regression algorithm when implemented after extracting feature with term frequency and inverse document frequency gave the highest accuracy of 71% while testing the model. Sruthi. M. S,Rahul R,Rishikesh G implemented An Efficient Supervised Method for Fake News Detection using Machine and Deep Learning Classifiers [13]. They used news channel data from Kaggle.com. In this paper they used various methods like Naive Bayes, SVM and LSTM. In this paper they used stop-word removal, tokenization, sentence segmentation, a lower

Special Issue - 2021

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
ICACT – 2021 Conference Proceedings

casing and punctuation removal for Preprocessing of data and tendency technique is used for feature extraction. Among the all methods LSTM gives the more accuracy of 94.53%.

F. Naive Bayes

It uses probabilistic approaches and based on Bayes theorem. They deal with probability distribution of variables in the dataset and predicting the response variable of value. An advantage of naïve Bayes classifier is that only requires less bulk of training data to access the parameters necessary for classification. Mykhailo Granik and Volydimyr Mesyura developed a Naive Bayes classifier technique for fake news detection [14]. In this they use Buzz feed news which contains the information of Facebook content. In this the classification accuracy for true is 75.59% and for false is 71.73% and accuracy for total is 75.40%.Rahul M, Monica R, Mamathan N, Krishana R developed a machine learning model for fake news detection by using FND-jru, Pontes Rout, News Files datasets [15]. After experimenting on different dataset they found that each model shows variant accuracies on different datasets among them Naïve Bayes, Passive Aggressive and DNN gave better accuracies of 90%, 83% and 80% .They used TF-IDF, Bag-of-Words, Count Vectorizer for feature extraction

G. Long Short Term Memory

Long short-term memory (LSTM) units are the building blocks for the layers of a recurrent neural network (RNN). An LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell is responsible for "remembering" values over a vast time interval so that the relation of the word in the starting of the text can influence the output of the word later in the sentence. Deepak S, Bhadrachalam Chitturia proposed a model for fake news detection [16] using FNN and LSTM. In this they used George McIntire dataset and Bag-of-Words, Words2Vec, Glove are used for feature extraction. On their experiment they found that LSTM shows better performance of 91% compared to FNN which is 84% accurate on classification. Nerissa Pereria, Simran Dabreo implemented Comparative Analysis of Fake News Detection using Machine Learning and Deep Learning Techniques [17]. They collected dataset from kaggle.com. In this paper they used various methods like Naive Bayes, SVM, LSTM and Neural Networks. In this paper they used Stemming, Tokenization and POS Tagging for Preprocessing of data and also used Count Vector, TF-IDF for feature extraction. Among all the methods LSTM gives more accuracy of 93%. Arush Agarwal, Akhil Dixit developed a Fake News Detection: An Ensemble Learning Approach [18]. They used two datasets are Liar dataset and fake news dataset for more experiment data. They proposed an ensemble based learning Classifier based on multiple base Classifier include SVM, convoluted neural network (CNN), LSTM, KNN, and NB. They used bag-of-words, N-Grams, TF-IDF, Word2Vec and parts of speech tagging for feature extraction. They were observed that LSTM showed better accuracy with 97%.

H. Hybrid models

Ms.Smita vinit implemented a hybrid model for fake news detection. It is a combination of SVM and Naïve Bayes techniques [19]. They used manual dataset for training the model and Word Count, Glout techniques are used for feature extraction. The hybrid model gave a good accuracy of 78%.
Shlok Gilda implemented a Evaluating Machine Learning Algorithms for Fake News Detection [20]. They collected dataset from signal media. In this model they used Bounded Decision Tree, Gradient Boosting, Random Forest, Stochastic Gradient Descent, and SVM. In this paper they used TF-IDF, PCFG for feature extraction. Among all the methods the best performing models are Stochastic Gradient Descent models are trained on TF-IDF feature set only and gives accuracy of 77.2%.
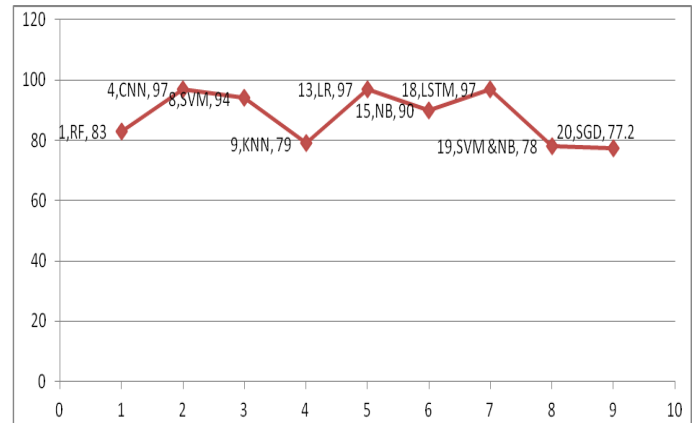

Fig2: Comparative study of existing Techniques

III. TYPICAL FRAMEWORK FOR FAKENEWS DETECTION

In our work, We want to implement the some machine learning methods for detecting the news is fake or real as mentioned in below table from google.com. For that first we collect the data and after that raw texts of news required some pre-processing .so that we want to pre-process that data. The performance of machine learning models depends on a great deal on feature design. So we want to extract a wide range of features and then by using the above methods we want to train the data for classifying the data. After that we want to classify the data which is fake or real.
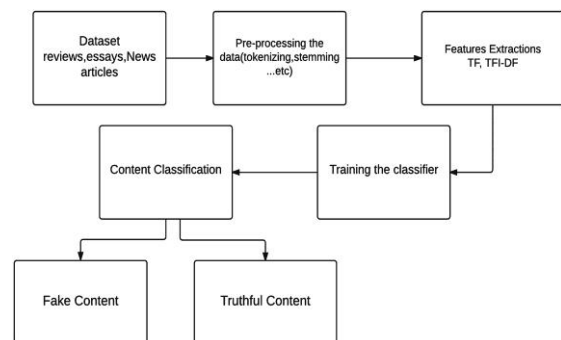

Fig 3: Typical Framework for fake news detection using machine leaning techniques

In our work, we currently use Naive Bayes, Random Forest, Decision Tree, Logistic Regression and Support Vector Machine on Liar Dataset.

Naive Bayes method is a set of supervised learning algorithms based on applying Bayes theorem. Bayes theorem is used to find the probability of an event occurring given the probability of another event that has already occurred. The most important assumption that Naive Bayes makes is that all the features are independent of each other.

Random forest or random decision forest is a supervised machine learning algorithm based on ensemble learning for classification and regression that operate by constructing a multitude of decision trees at training time and obtaining mode of classes as outputs. If there are higher number of trees in the forest and prevents the model from over fitting and gives best results. Those results are merged together in order to get more accurate prediction. Each decision tree is trained separately based on data .If a new data point is introduced in the dataset it doesn't affect the overall algorithm.

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes and success) or 0 (no and failure).

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

IV. RESULT ANALYSIS

We have used Liar dataset [22] for experimenting the machine learning models to find the fake news. In this experiment we verified the accuracies of Logistic Regression, Support Vector Machine, Naïve Bayes model, Random Forest and Decision tree models. Experiments result in Random Forest better performed compared to remaining models.
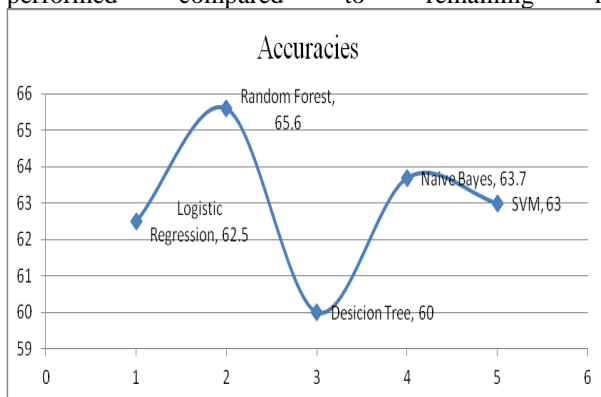

Fig 4: Typical Framework for fake news detection using machine leaning techniques

Figure 4 on the y axis marks accuracies and x axis various machine learning models experimented.

V. CONCLUSION

In these days more people are continuously consuming news from social media rather than the traditional media. This fake news develops a strong negative impact on individual users and the society. Therefore for detecting the fake news we analyse different research papers and identifies Word Embedding, Tokenization and Parts of speech tagging are best for Pre Processing of data and also identifies TF-IDF and Count Vectorizer are best for feature extraction. So, for better approach Further we want to use those methods for PreProcessing, feature extraction and also we want to implement the Random Forest classifier , Convolutional Neural Networks , Long Short Term Memory for high accuracy and an Ensemble Learning Approach for high accuracy

REFERENCES

[1] Bhavika, Bhutani, Neha, Rastogi, Priyanshu, Sehgal, Archana and Purwar implemented a Sentiment Analysis technique for Fake news detection.

[2] Rohit Kumar Kaliyar developed A Deep Neural Network techniques for fake news detection.

[3] Belhakimi Mohamed Amine, Ahlem Drif, Silvia Giordano developed a Merging deep learning model for fake news detection.

[4] The merged CNN reached accuracy of 96%. V. M. Krešňáková et al developed Deep learning methods for Fake News detection.

[5] Tayyaba Rasool, Wasi Haider Butt, Arslan Shaukat and M. Usman Akram developed a Multi-layered Supervised Learning for Multi-Label Fake News Detection.

[6] Nicollas R. de Oliveira, Dianne S.V.Medeiros and Diogo M.F.Mattos developed a model for fake news detection.

[7] Karishnu Poddar, Geraldine Bessie Amali D, Umadevi K S developed a Comparison of Various Machine Learning Models for Accurate Detection of Fake News.

[8] Smitha, Bharat developed a Performance Comparison of Machine Learning Classifier for Fake News Detection.

[9] Ankit Kesarwani, Sudakar Singh Chauhan and Anil Ramachandran Nair developed a K-Nearest Neighbour Classifier technique for Fake News Detection on Social Media.

[10] Uma Sharma, Sidarth Saran, Shankar M. Patil developed a Fake News Detection using Machine Learning Algorithms.

[11] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf and Muhammad Ovais Ahmad implemented a model for fake news detection on social media.

[12] Vanya Tiwari, Ruth G. Lennon and Thomas Dowling developed some Machine learning Algorithms for Fake news detection.

[13] Sruthi. M. S,Rahul R,Rishikesh G implemented An Efficient Supervised Method for Fake News Detection using Machine and Deep Learning Classifiers.

[14] Mykhailo Granik and Volydimyr Mesyura developed a Naive Bayes classifier technique for fake news detection.

[15] Rahul M, Monica R, Mamathan N, Krishana R developed a machine learning model for fake news detection by using FND-jru, Pontes Rout, News Files datasets.

[16] Deepak S, Bhadrachalam Chitturia proposed a model for fake news detection.

[17] Nerissa Pereria, Simran Dabreo implemented Comparative Analysis of Fake News Detection using Machine Learning and Deep Learning Techniques.

[18] Arush Agarwal, Akhil Dixit developed a Fake News Detection: An Ensemble Learning Approach.

[19] Ms.Smita vinit implemented a hybrid model for fake news detection. It is a combination of SVM and Naïve Bayes techniques.

[20] Shlok Gilda implemented a Evaluating Machine Learning Algorithms for Fake News Detection.

Special Issue - 2021

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
ICACT – 2021 Conference Proceedings

[21] Website: https://libguides.pace.edu/fakenews Author: Sarah Cohn, Instructional Services Librarian Last Access on: 16th May 2021.

[22] Datasets: Liar dataet source: https://huggingface.co/datasets/liar acessed on:16th May 2021 uploaded by: Hugo Abonizio