DETECTION OF FAKE NEWS USING CLASSIFICATION ALGORITHMS OF MACHINE LEARNING

Pawan Kumar, Anu Chaudhary, Shashank Sahu, Vivek Agarwal

Department of Computer Science & Engineering , Ajay Kumar Garg Engineering College Ghaziabad
drpawancse@gmail.com, hodcse@gmail.com, Sahushashank75@gmail.com, vivelagarwal292@gmail.com

Abstract: Spreading of fake news by various social media or online digital media platforms is increasingly rapidly and sometime it causes the lot of social and national damages. Now, handling of such fake news is becoming very serious matter of concern for the society due to its destructive impacts. Many researchers are already suggested the various fake news detecting approaches. Our research work used various supervised classification algorithms for classifying the fake news and these detection classifiers are classifying fake news as true or false. The proposed research work not only help in classify fake news but also helps in analysing or exploring to select best machine learning classifier for detection of fake news. The research work used feature extraction technique for text data extraction. The accuracy of logistic regression and random forest model is scored minimum and maximum with scored value of 87.35% and 98.46% respectively.

Keywords: Fake news, Online deception, Disinformation, Misinformation and social networks.

I. INTRODUCTION

Fake News contains some misleading or useless information that are needed to be check. Due to this misleading information, sometimes countries have to face many problems. It has been noticed that mostly countries in Arabic spring are always in unrest due to spreading of fake news. The House of Commons organization try to handle the fake news but due to its limited storage capacity and it is not possible to manage the huge increasing digital network data. As we know, billions of articles are published or removed every minute on social network and it is very difficult to manage manually. A smart solution is needed to manage the spreading of such misleading information [1][2].

Sometime, fake news is generally created and circulated for getting political gains. Often government's agencies and own party may be involved. It has been seen in many countries such as China and Russia where manipulation is circulated through internet to public and control is completely in the hand of government. Main aim of proposed research work to implement or design a methodology by using classification algorithms of supervised machine learning algorithms that help in creating a ML classifier that detect a fake article. The authenticity of articles is identified by using phrases, words, titles and sources by applying various supervised classification machine learning algorithms on given available dataset The dataset is collected from online repository. The Pearson Correlation technique is used for selection of right features from available dataset and that for select best fit features that are help in providing highest accuracy and precision result [3][4]. The final accuracy of proposed model is tested on unseen data and its final outcome result is plotted accordingly. The proposed model is used for detection and classification of fake news articles and that can be applied and integrated with any system and can be applied in future use.

II. BACKGROUND STUDY

A. Social Media and Fake News

Social Media Platforms contains information in form of various social media tools such as social websites, social bookmarking, wikis, micro blogging etc. Now, social media tools have become very powerful tool for spreading fake news dissemination. Sometimes, spreading of fake due to social media causes the accidental issues. As we noticed, during COVID -19, the spurt of fake and real news included lots of misinformation. Such misinformation's spreads at very fast speed throughout globe and prove a deadly when it is reinforcing misplaced or personal bias against the all trustworthy evidences Accordingly, the joint statement of the National Academies that was posted on July, 2021. The social media, has to improved ease with broadcasting of real-time information [5][6].

B. Natural Language Processing (NLP)

As we discussed above that fake news are the information that are presented on the web in different forms such as in form long stories, images, videos etc. NLP is very powerful tool for managing or analysing text and combat this issue. It provides us with tools and techniques to analyse text and raise flags for any kind of fake news. NLP plays very important role for detecting the fake news by using various different techniques that are used to analyse text data such

as classification of texts, sentiment and semantics analysis, fact checking, network analysis etc. NLP models are well trained for classifying news articles or digital media posts and identified the real or fake news based on patterns such as structure of word and sentences or other features. NLP model also analyses sentiments that are expressed in news articles and try to detect any misleading information or content by examine emotional language and tone used in given text.NLP models also help in understanding the text semantic meaning by using relationship of words and phrases. It collects real time information from data in real time and flags of contextual clues those are indicates misinformation [7][8].

C. Data Mining
Data mining techniques are used in machine learning for extracting useful or meaningful information from available dataset. It can be applied for both supervised and unsupervised classification problems and it is very good technique that help in identifying, recognizing or discovering hidden patterns from given dataset and play a very crucial role in model designing 9][10].

III. MACHINE LEARNING CLASSIFICATIONS MODEL

Machine Learning (ML) is subpart of artificial intelligence and contains several numbers of supervised and unsupervised algorithms that help in achieving more accurate and correct results. ML models are applied in various areas and used for their problem solving such education, health sectors, banking and insurance, social media related problems etc. Data scientist's analysis the specific data and designed or developed a specific accurate predicted model for respective problems. The predicted model generated the target output either in continuous or discrete value depend upon the specific problems. The research proposal used various classifications algorithms for fake new detections. The description of algorithms is discussed below. There are four algorithms are used for in proposed work to classify fake news[11][12].

A. Decision Tree (DT)
Decision tree is one of most popular algorithm of solving classification problem that was based on flow chart structure. DT has different number of choices that help it to take right easy decisions. Every internal node of DT checks condition on every individual attribute and its branches are performing on basis of test conditions and final result is evaluated. Finally, leaf node of tree specifies label of classes that evaluated after computing values of all attributes. The distance from top root to leaf shows classification rule. DT model are performing good result for identifying most valuable variables. They are generally used for Creating a new features and variables that are used for data exploration and it also predicts output (target) variables that are worked quite efficiently. DT are supervised learning algorithm that widely used for prediction modelling and perform high accurate result. DT algorithm perform well for both linear and nonlinear dataset and are used for solving both classification or regression problems [13][14].

B. Random Forest (RF)
RF is collection of various DT in which all DT are work together and basis of their voting result to make final prediction. RF algorithm is very powerful tree algorithm that is mostly applied to large data. The accuracy of this RF model is very high as compared to other classifications algorithms. RF is used for solving both classification and regression problems. RF is a classifier that take decision on output of many DT and make final decision [15].

In RF, the available is divided into randomly into several decision tree and model is trained separately by number of DT tree and their final combine average outcomes results help in making accurate or right decision. This combined approach helps in improving accurate predictions of model. RF algorithms can be a good choice for detecting fake news for large dataset.

C. Gradient Boosting Classifier
Gradient Boosting is also a very powerful boosting technique of classification algorithm that was worked on combining by several weak learners into a very strong learner, in which each new model is prepared or trained for minimizing loss function and its accuracy is measured by confusion matrix, precision call and other accuracy score matrix. Gradient Boosting model has three main components that are used for classification are as:

Loss Function: This function is mostly used for to estimate how our model is good for making predictions on given available data. The higher value of loss function indicated that model is performing good and accuracy of model is high. Less value of loss function shows that there is large difference between predicted value and targeted value.

Weak Learner: It classifies given data poorly and it has also high rate of error. These are mostly used in decision trees. This is also known as decision stumps and are less complicated than DT.

Additive Model: This model is used for adding number of trees sequentially and performed step by step. After every and each number of iteration, we try to reach nearly to final destination. During each or every iteration value of loss function will be reduced.

D. Logistic Regression (LR)
Logistic regression can be applied for both classification as well as regression problems. LR model used sigmoid function

for binary classification and soft-max function for multi-classification. The value of targeted output can be one 0r zero in case of binary classification. LR model are good model for classification of linearly separable and small data set [16] [17] [18].

IV. PROPOSED METHODOLOGY

In proposed work designed a methodology for classification of fake news by using various supervised classification algorithms of machine learning. In proposed work, a software tool is designed for detecting the fake articles. The initially step was taken the collection of data and then data is classified by using phase, text etc step by step by using various phases of ML included pre-processing, features selection, training and testing of available dataset and final result in form of targeted output. Figure [1] explain the working flow of proposed worked. The designed methodology used random forest, decision tree, logistic regression and ensemble techniques for classification of fake news. The accuracy of each model is calculated by using precision, recall and F1 score matrixes. The combined result of each decision tree help in detecting accurate fake news.

The main objective of research work is applied several classification algorithms for obtaining a good classification model that are used to finding fake news data by python application. The, appropriate Python code produce an optimized result code.
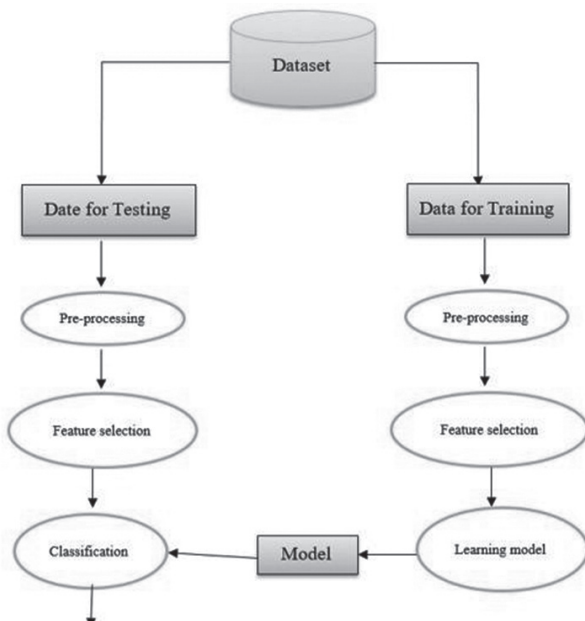
models, we can select the best accurate model for fake news classification and reach on final output decision. The model creation process of proposed work starting from collection of news dataset, to perform pre-processing (in this process removed all rough noise by using natural Language Toolkit), features selections, splitting of data to creation of models. Testing is model is checked by test dataset, and final results is verified, the final step is to calculate or monitor precision output of model. The model is also tested on unseen data that was selected independently by user. The model was created on dataset that was included 50% data was fake and 50% was real data. For model creation, 90% of available data is used for model training and remaining 10% data is used for testing.

V. RESULT ANALYSIS

The proposed work used political data of news of fake news detection and that is very famous dataset and known as New Benchmark Dataset and it categorised into two labels fake news or trust news. The final result analysis on given datasets was calculated by using four classifiers

- Logistic Regression
- Decision Tree Classifier
- Gradient Boosting Classifier
- Random forest

The confusion matrix of logistic regression is given in figure 2 and figure 3 respectively.
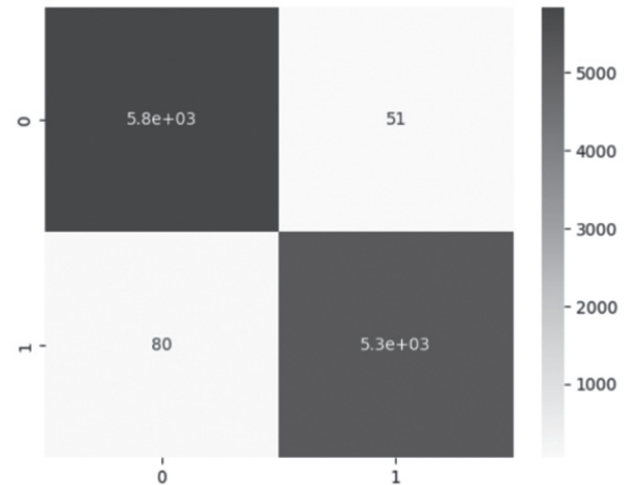

Figure 1. Working Flow Diagram of Proposed Model

Figure 2. Logistic Regression

All designed classifiers are performed well for classification of fake news and after comparative analysis of accuracy of each

The accuracy of Random forest is maximum among all classifiers with score value of 97.37% and LR model was achieved minimum accuracy scored with value 87.54%. The Accuracy table of each model is given in Table 1.

Table 1. Accuracy Comparison

S.No	Model	Accuracy Score
01	LR	87.54%
02	DT	92.13%
03	RF	97.37%
04	Gradient	95.45%

VI. CONCLUSION

Research work focuses on calculating or detecting fake news by using various classification classifiers. The news are fake or real are analysis on the basis of speech characteristics and designed model is good model for predicting target output. Gradient Boosting Classifier achieved good accuracy scored with value of 95.45% for detecting real or fake new or messages. LR model achieved minimum accuracy score with minimum value of 87.54% and random forests scored highest accuracy scored value 97.37%.

In future ,we can also improve the accuracy of model by applying some deep learning algorithms. As we know , accuracy of model is highly depended on size of data and deep learning algorithms are performed well for large dataset as compared to machine learning algorithms.
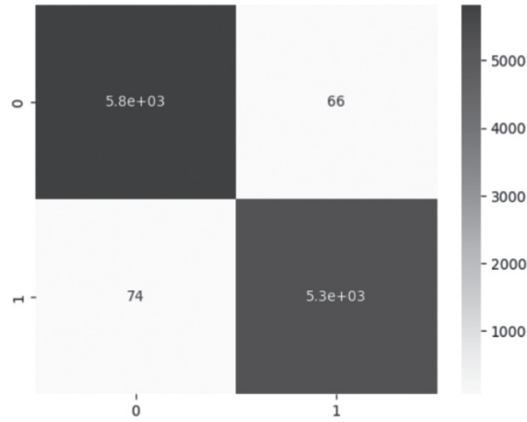

Figure 3. Random Forest

REFERENCES

[1]. Radianti J. et al ,” An Overview of Public Concerns Duringthe Recovery Period after a Major Earthquake Nepal Twitter Analysis”, Hawaii International Conference on System Sciences (HICSS) (Washington, DC, USA: IEEE) 136-145,2016.

[2]. Alkhodair S A, Ding S H.H,” Detecting breaking news rumors of emerging topics in social media “,Inf. Process. Manag Fung B C M and Liu J, 57 -64, 2020.

[3]. Jeonghee Yi et al. “Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques.” In Data Mining, ICDM 2003. Third IEEE International Conference 427-434,2003.

[4]. Tapaswi et al ,” Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit m sentences “, Software Engineering (CONSEG), on Software Engineering (CONSEG) (IEEE),2012.

[5]. Ranjan et al ,” Part of speech tagging and local word grouping techniques for natural language parsing in Hindi.”, Proceedings of the 1st International Conference on Natural Language Processing (ICON 2003).

[6]. Diab Mona et al.,” Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks “,Proceedings of HLT-NAACL 2004: Short Papers (Boston, Massachusetts, USA: Association for Computational Linguistics) 149-152,2004.

[7]. Rampersad G and Althiyabi T ,” Fake news: Acceptance by demographics and culture on social media”, J. Inf. Technol. Politics 17 1-11 2020.

[8]. Sirikulviriya Naphaporn and Sinthupinyo Sukree,” Integration of Rules from a Random Forest “, International Conference on Information and Electronics Engineering (Singapore) 194-198 ,2011.

[9]. Kevric Jasmin et al.,” An effective combining classifier approach using tree algorithms for network intrusion detection Neural Computing and Applications 1051-1058. ,2017.

[10]. Granik Mykhailo and Mesyura Volodymyr,” Fake news detection using naive Bayes classifier”, First Ukraine Conference on Electrical and Computer Engineering (UKRCON) (Ukraine: IEEE),2017.

[11]. Gilda S.,” Evaluating machine learning algorithms for fake news detection “, 15th Student Conference on Research and Development (SCOReD) (IEEE) 110-115,2017.

[12]. Jain Akshay , Kasbe Amey,” Fake News Detection”, IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS) (Bhopal, India: IEEE) 2018.

[13]. Gupta Arushi , Kaushal Rishabh,” Improving spam detection in Online Social Networks “,International Conference on Cognitive Computing and Information Processing (CCIP) ,2017.

[14]. Khanam Z. ,” Analyzing refactoring trends and practices in the software industry”, Int. J. Adv. Res. Comput. Sci. 10 0976-5697,2018.

[15]. Aphiwongsophon,Supanya et..al.,” Detecting Fake News with Machine Learning Method”, 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) (Chiang Rai, Thailand, Thailand: IEEE),2018.

[16]. Kaur Prabhjot et al .,” Hybrid Text Classification Method for Fake News Detection”, International Journal of Engineering and Advanced Technology (IJEAT) 2388-2392,2018.

[17]. Looijenga M. S. ,” The Detection of Fake Messages using Machine Learning”, Twenty Student Conference on IT, Jun. 6th, Enschede The Netherlands (Netherlands),2018

[18] Neeharika Tripathi, Amrita Tiwari, Enhancing Policing Efficiency: A Machine Learning-Based Crime Management System, *Glimpse Journal of Computer Science, 3(2), pp. 29 33, 2024*

ABOUT THE AUTHORS

Dr. Pawan is Associate Professor in the department of CSE at AKGEC, Ghaziabad, UP, India since February 2024. He has received his Master and Doctorate degree in field of Computer Science and Engineering from MDU, Rohtak in 2008 & 2017 respectively. The major areas of his interest and research are Cloud Computing, Machine Learning and & Deep Learning etc. He has published several research papers in various conferences & journals.

Dr. Anu Chaudhary is working as Head and Professor at Ajay Kumar Garg Engineering College, Ghaziabad in the Department of Computer Science and Engineering. He has vast experience of more than 25 years of academic and administration. He has published 4 patents and more than 30 papers in various international and national journals/conferences. He has guided 3 Ph.D students and M.Tech students. His area of specialization is high speed data network, Database Management and Software Engineering.

Dr Shashank Sahu is working as Professor in Department of Computer Science and Engineering at Ajay Kumar Garg Engineering College Ghaziabad. He is B.Tech., M.Tech. and Ph.D. in Computer Science and Engineering discipline. He has published more than 55+ research papers in National/International journals and conferences. He has also published patents and book chapters. He is professional member of the IEEE. He has guided more than 12 M.Tech. dissertations and 1 Ph.D. student.

Mr. Vivek Agarwal is Assistant Professor in the department of CSE at AKGEC, Ghaziabad, UP, India since August 2024. He received his Bachelor's degree in Computer Science Engineering from UPTU in 2009 and Master degree in Computer Science and Engineering in 2014 from UPTU. The major fields of his interest and research are IoT, Cloud Computing, Machine Learning and & Soft Computing etc.