Comparative Analysis of Fake News Detection Using Different Methodologies

Vandana .A[1], Dr. Clara Shanthi .D[2]

[1] Student, Department of Computer Science, Mount Carmel College Autonomous, Bengaluru, India

[2]Assistant Professor, Department of Computer Science, Mount Carmel College Autonomous Bengaluru, India

Abstract: ***In recent times, fake news and the influence it has become a growing cause of concern. Fake news is information that consists of news that is not well - researched. There is always a continuous need to check for news whether it's real or not that is received from various sources such ase - content, websites and blogs. Over the past few years, the news has been developed from written news to printed newspapers, magazines, tabloids and finally to digital news such as social media feeds,, blogs, online news platforms and other digital media. The great source of information in recent times is social media interaction especially the news that is spreading around the network. Fake news has become a major concern ever since the internet is boomed. This has also evolved into a network that allows us to stay informed about global events, as well as a breeding ground for malicious and fake news. It is critical to combat fake news since the world's view is based on this information. It is necessary to curb this because people not just make important decisions based on the information but also form their opinion. If the information taken is false, then it comes with a devasting consequence. Considering the widespread of social media platforms, people create and share more information than before few of them may be deceptive which will have no relevance to reality. The main question arises how do we authenticate the real news and the articles that is being circulated on various social media platforms like WhatsApp, Facebook, Instagram, Twitter and micro blogs and other social networking sites. This paper attempts identify the fake news by proposing a system that will be able to classify the fake news with help of few machine learning and deep learning algorithms.***

Keywords: Fake news detection, Social media, Machine Learning Algorithms, Internet, e - content, Websites, Blogs, Deep Learning Algorithms

1. Introduction

The first written news in the world came into existence in 8th BCE China, where officials gathered the report and eventually compiled as the Spring and Autumn Annals. The first newspaper emerged in Germany around 1600s. The world's first formalised paper was Relation aller Fürnemmen und gedenckwürdigen historian which means Account of all distinguished and commemorable stories, from 1605 [1]. Fake news originated around 1800s. The main reason for origin of fake news is sensationalism which always sold well. Around the year 1890s, William Hearst and Joseph Pulitzer, two rival newspaper publishers, battled for attention by sensationalizing stories and portraying rumors as truths. Their distrustful news played a major role in Spanish - American War of 1898. By19th century, modern newspapers came into existence, due to which circulation of fake news also increased. In the year 1835 the New York Sun's "Great Moon Hoax" purported there was an alien civilization on the moon, and that established "Sun" as a profitable newspaper [2]. The majority of adults rely on news that is primarily sourced from social media. The statistics shows that 59% are Twitter users, Facebook users are 66% and 70% users depend on the subsequent platforms for their news. Social media is the center of news for majority of people all over the world. The structure of the news from modern social media platform varies a lot from generic news platform such as newspapers. In recent times, with growth in the advancement of social media platforms, deceptive news for different commercial and political purposes have been showing up in very huge numbers and far - reaching in this online world. As we see how fake news has affected people around and the society from the time news has emerged/originated. It becomes important to curb fake news as it may lead to disasters like war, and it will also affect the decisions that has to be taken based of the news that is circulated.

2. Literature Survey

The problem to be handled is very relevant in this era of information, several works have been looked into from different perspectives, that focuses on various techniques, but the ultimate goal of all is to combat misinformation. Fake news detection examines the veracity of news received from various web sources such as e - content, blogs, and websites.

The study conducted in the year 2019 emphasizes on application of ML and NLP techniques for detection of fake news. The ML algorithms used are kNN, Naïve Bayes, SVM, Random Forest and XGB. The classification was performed with F1 score and ROC curve. The RF and XGB classifiers gave better results and statistically tied with 0.85 (±0.007), 0.86 (±0.006), 0.81 (±0.008) and 0.81 (±0.011) for AUC and F1 respectively [3]. In another study conducted it emphasizes on detecting fake news using the existing algorithms in Machine Learning. The classifiers used are Multinomial Naïve Bayes (MNB), RF, Gradient boosting and DT. It was shown that Gradient Boosting gave best results in classifying the multi - class dataset for fake news classification and has accuracy of 86% [4]. In one of the study conducted it emphasizes on classifying the fake news based on machine learning algorithms. The algorithms used are Naïve Bayes and SVM. It was shown that SVM gave the best results in classifying fake news using semantic analysis with accuracy of 75% [5]. In another study it emphasizes on predicting the fake news that is multi - dimensional in nature. The algorithms used are the mixed strategy of Naïve Bayes classifier, SVM and semantic investigation. The results of NLP and CNN was 76% and 86.65% respectively. The final results show that the algorithms used had an accuracy of 93.50% which showed best results compared to other models that were considered [6]. In the study conducted it emphasizes on detecting the fake news using different ML and DL Models. The ML & DL algorithms used are Bayesian Model, LR, SVM, RNN and LSTM. The accuracy is based on Count Vector and TF - IDF for the classifier models. The accuracy for the deep learning models is based on default and with kernel initialization for the classifier models. The experimental results showed that SVM gave best results compared to other ML and DL models with accuracy of 89.34 % and 89.34%. In one of the studies that was carried out it emphasizes on detecting fake news using Vectorizer's. The algorithms used for detecting fake news are Naïve Bayes, SVM, LR, DT and Neural network. The experimental results show the accuracy for the algorithms with respect to TF - IDF and Count Vectorizer. LR gave better results with respect to TF - IDF and Count Vectorizer with accuracy of 0.916 and 0.910 [7]. The other study emphasizes on detecting fake news using different ML and DL models. The datasets considered are Jru, Pontes, ClaimsKG, Kaggle, Liar, Newsfiles and Superset. The algorithms used are Naïve Bayes, Passive Aggressive and DNN. The experimental results show the accuracy for algorithms with respect to datasets as follows – for Jru dataset Naïve Bayes (NB) - 96%, Passive aggressive (PA) and DNN - 99%, for Pontes dataset NB - 96.6%, PA - 98.5% and DNN - 97%, for ClaimsKG dataset NB - 77.2%, PA - 73.5% and DNN - 77.9%, for Kaggle dataset Naïve Bayes - 85.1%, PA - 83.8% and DNN - 82.6%, for Liar dataset NB - 71.8%, PA - 64.7% and DNN - 63.9%, for News files dataset Naïve Bayes - 97.8%, PA - 99% and DNN - 98% and for Superset dataset NB - 84.6%, PA - 87.1% and DNN - 85%. The comparison between the algorithms show that DNN performed well with respect to all the datasets [8]. In one of the study it emphasizes on detecting semantic fake news using machine learning techniques. There are two different datasets considered Liar and Politifact respectively. The ML & DL algorithms used are Naïve Bayes, SGD Classifier, Logistic Regression, Random Forest, Decision tree, SVM, CNN, Basic LSTM, BiLSTM and GRU. For the experimental results the authors conclude that SVM gives better accuracy of 0.294 compared to other machine learning models, GRU gives better accuracy of 0.452 compared to other deep learning models [9]. In the final paper referred the study emphasizes on detecting fake news using different ML approaches. The dataset considered is taken from Kaggle and respective analysis is performed. The algorithms used are Random Forest, XGBoost, Naïve Bayes, kNN, Decision Tree and SVM. From the experimental results authors conclude that XGBoost and SVM performed well and showed better accuracy of 75 %compared to other models [10].

3. Methodology

3.1 Process Flow

The system architecture shown in Fig.1 depicts the flow of all work performed. Totally 15 research papers were reviewed before selecting appropriate models. Then the data cleansing and pre - processing was done for the same. Further, feature extraction and selection was done and the selected features were converted to numeric with suitable vectorizer function. The accuracy was checked and required adjustments were made and thenthe results were achieved for 4 classification models and compared which gave the best result out all of them.

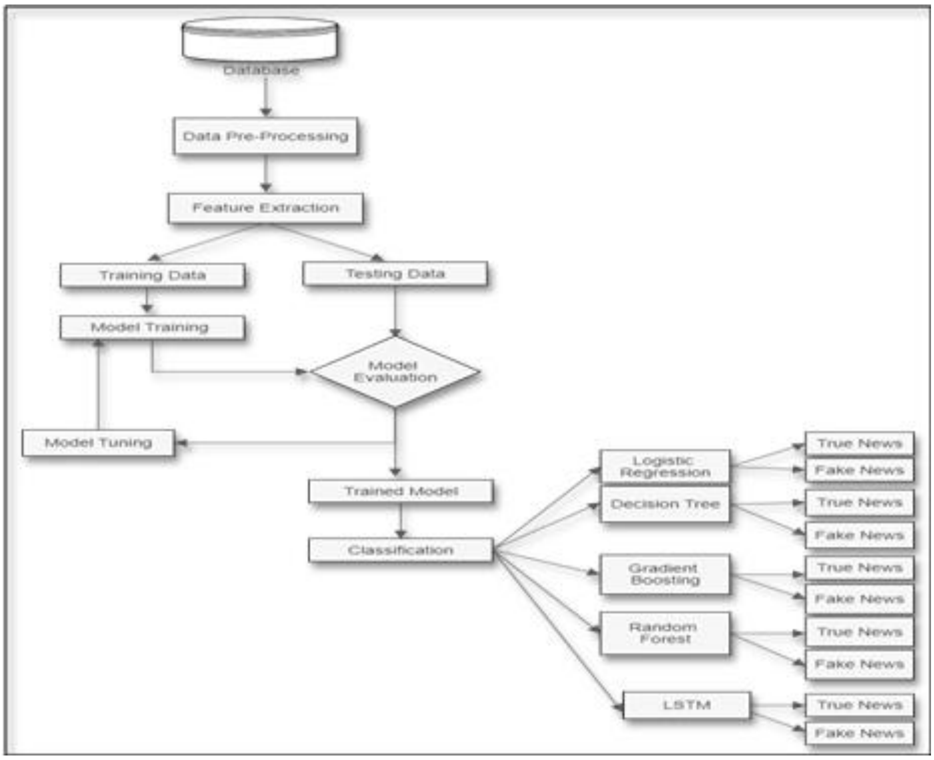
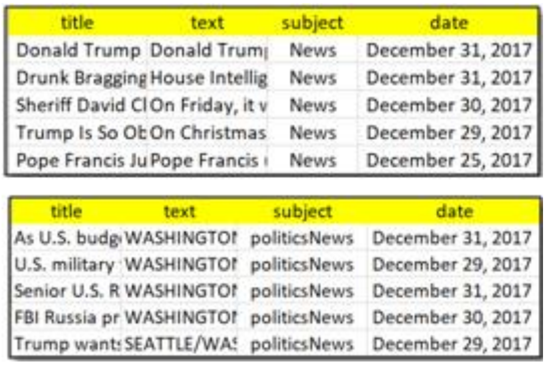


Figure 1: System Architecture

3.2 Pre – Processing of Data and Data Collection

Raw data was obtained by the website Kaggle. com, which had two different csv files namely Fake and True. The dataset contains different news articles like general news, political news for the years 2015 to 2017. Since the raw data was received, there was a lot of cleaning process that had to be done to make the data usable. Since the live data couldn't be extracted from the online websites, the same recorded data was downloaded Kaggle. com. Extra target variable class was added to the table while pre - processing. To the data that was obtained, we had to add the target / class variable to the dataset which helped to classify the fake news. There are 2 excel files that are extracted from the dataset – the first excel sheet consists of news articles that are fake and is named Fake. csv and the second excel sheet consists of news articles that are true and is named True. csv. Both the datasets consists of 4 columns namely title, text, subject and date



title	text	subject	date
Donald Trump	Donald Trum	News	December 31, 2017
Drunk Bragging	House Intellig	News	December 31, 2017
Sheriff David Cl	On Friday, it v	News	December 30, 2017
Trump Is So Ob	On Christmas	News	December 29, 2017
Pope Francis Ju	Pope Francis	News	December 25, 2017

title	text	subject	date
As U.S. budg	WASHINGTO	politicsNews	December 31, 2017
U.S. military	WASHINGTO	politicsNews	December 29, 2017
Senior U.S. R	WASHINGTO	politicsNews	December 31, 2017
FBI Russia pr	WASHINGTO	politicsNews	December 30, 2017
Trump wants	SEATTLE/WAS	politicsNews	December 29, 2017

Table 1: Sample raw data from Fake. csv and True. csv

3.3 Machine and Deep Learning Models

Machine Learning is subset of AI that will help the model to improve and learn by using the previous data that is available. The major goal of machine learning is for system to learn by itself automatically through algorithms with human intervention. The ML algorithms can be classified into 3 kinds –Unsupervised learning Supervised learning and Reinforcement learning. Supervised learning is a algorithm that will apply their learnings from the previous data to predict the events that will occur in the future. We use unsupervised learning algorithms when data is not labelled. It doesn't give us the correct output or predict, instead the data is explored and will draw analysis and the inference will be used to explain the hidden structure from unlabelled data. Reinforcement learning algorithms learns on its own. DL is a subset of AI and ML which tries to mimic the human brain to gain certain types of knowledge. In the proposed work, we have used supervised learning algorithms such as LR, DT, RF, unsupervised learning algorithm gradient descent and deep learning algorithm LSTM.

Logistic regression is easy to interpret, implement and efficient for training. LR is mainly used to predict categorical variables that are dependant with help of set of independent variables. Since the prediction that is made is categorical the outcome will be either 0 or 1. Decision tree can be used to solve problems related to both classification and regression, but classification is more preferred. It is split in the form of a tree to arrive at the classification. On the other hand, the RF is a combination of many decision trees. It will create multiple DTs and considers the majority poll of predictions among all the DTs and will predict the result. The greater the number of trees means better accuracy and it will also prevent overfitting. Gradient descent is a ML algorithm that helps us to solve optimization problems. It multiplies the gradient by a number that will help to determine the next point. LSTM is deep learning algorithm and is a type of RNN that is capable of learning order dependence in sequence problem prediction.

4. Experimental Results

The experimental results obtained after model building and analysis are given below –

Algorithm	Precision	Recall	f1 score	Time Complexity	Space Complexity	Accuracy
Logistic Regression	0.99	0.99	0.99	1 sec	40287973	0.9838
Decision Tree	0.99	1.00	0.99	26 sec	129721270	0.9946
Gradient Boosting	1.00	1.00	1.00	224 sec	182681301	0.9947
Random Forest	0.99	0.99	0.99	62 sec	129567253	0.9870

Algorithm	Train Accuracy	Test Accuracy	Time Complexity	Space Complexity
LSTM	98.68	98.66	15 sec	134436574

The ML algorithms mainly work based on TF - IDF vectorizer. When we feed a lot of data to the machine it learns from the pattern and gives the prediction. But when it comes to textual data it gets difficult for the machine to understand the textual data so here is when feature extraction comes into picture. We convert the text into vectors it is easier for the machine to understand and interpret. TF - IDF helps us in mapping from textual to numerical data it becomes easier for the machine to interpret. After using algorithms such as LR, DT, GBC, RFC and LSTM, we can conclude that GBC is more accurate and reliable than other algorithms. Coming to the dataset, the dataset used had fake and true comma delimited files (. csv). Prior to the training, an extra step was added which removed all punctation, stop words which increases the system's accuracy. Precision tells us about the positive prediction made by the model. Recall tells us about how successful was the information retrieved from precision. f1 score is just the mean of precision and recall.

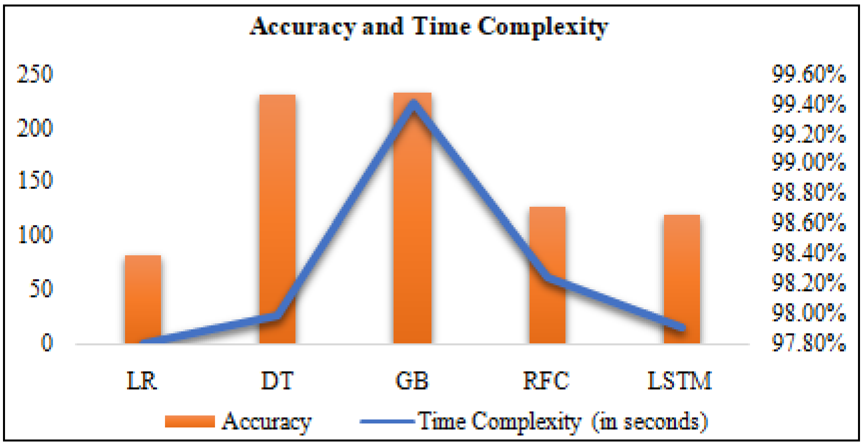


Figure 2: Accuracy graph of ML and DL algorithms

5. Conclusion and Future Enhancement

This project consists of different classification algorithms. The work carried outgave us better results by providing us with accurate predictions for both real and fake news, and hence we can say it can be a feasible approach. We have taken data from Kaggle. com. In this paper, Gradient Boosting Classifier model gave most accurate predictions for the present dataset. We have used 4 ML and 1 DL algorithm to predict fake news. Despite the short study period, we were able to test 5 models, yielding good results. The work completed can be expanded upon by extracting live data from different social media websites, e - magazines and other online platforms with help of more deep learning models which will help in detecting the fake news more accurately and more advanced visualization techniques can also be added. In this document we have used few ML and DL algorithms, other methods can be used to achieve better results.

References

[1] https: //www.bbc. co. uk/bitesize/articles/zwcgn9q

[2] https: //www.cits. ucsb. edu/fake - news/brief - history

[3] Julio C. S. Reis, Andre Corriea, Fabricio Murai, Adriano Veloso, Fabricio Benevenuto, “Supervised Learning for Fake News Detection”, *International Conference on Data Science, E - learning and Information Systems*, 2021, pp.185 - 192. https: //doi. org/10.1145/3460620.3460753

[4] Rohit Kumar Kaliyar, Anurag Goswami and Pratik Narang, “Multiclass Fake News Detection using Ensemble Machine Learning”, *IEEE*, 2019, pp.103 - 107, doi: 978 - 1 - 7281 - 4392 – 7

[5] Arun Nagaraja, Soumya K. N, Prajwal Naik, Anubhav Sinha, Jain Vinay Rajendrakumar, “Fake News Detection Using Machine Learning Methods”, *International Conference on Data Science, E - learning and Information Systems*, 2021, pp.185 - 192. https: //doi. org/10.1145/3460620.3460753

[6] Anjali Jain, Avinash Shakya, Harsh Khatter and Amit Kumar Gupta, “A smart system for fake news detection using machine learning”, IEEE, 2019, doi: 978 - 1 - 7281 - 1772 – 0

[7] Abdullah - All - tanvir, Ehesas Mia Mahir, Saima Akhter and Mohammad Rezwanul Huq. “Detecting Fake News using Machine Learning and DeepLearning Algorithms*”*, *ICSCC*, 2019, doi: 978 - 1 - 7281 - 1557 – 3

[8] Rahul R Mandical, Mamatha N, Shivakumar N, Monica R and Krishna A N, “Identification of Fake News Using Machine Learning”, *IEEE*, 2020, doi: 978 - 1 - 7281 - 6828 – 9

[9] Adrain M. P. Brasoveanu and Razvan Andonie. “Integrating Machine Learning Techniques in Semantic Fake News Detection”, *Springer*, 2020.

[10] Z Khanam, B N Alwasel, H Sirafi and M Rashid, “Fake News Detection Using Machine Learning Approaches”, *IOP Conf. Series: Materials Science and Engineering*, 2020, doi: 10.1088/1757 - 899X/1099/1/012040