Analyzing Social media's real data detection through Web content mining using Natural Language Processing and Machine learning techniques

C.Sangeetha[1], Dr S.P. Swornambiga[1]

[1]Ph.D Research Scholar CMS college of Science and Commerce. Coimbatore, Tamil Nadu.
[2]Associate Professor, Department of Computer Science CMS college of science and Commerce, Coimbatore, Tamil Nadu

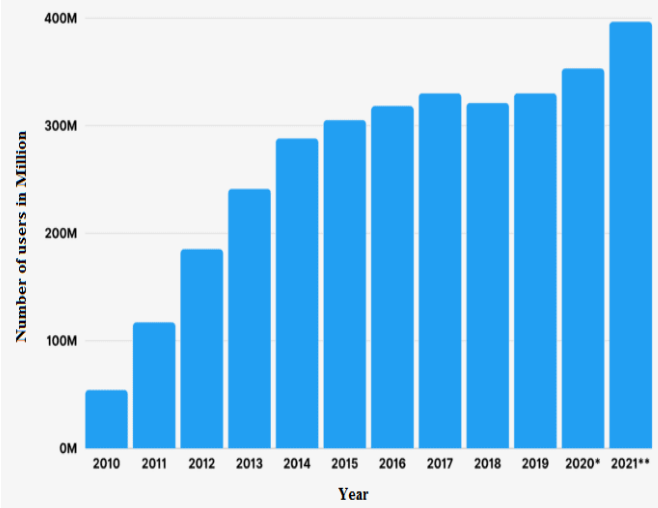
---***---

Abstract - *Online social media have developed into a significant informational resource for many individuals in recent years. The abuse of the media to distribute fake news multiplied along with the growth in social network usage. The accuracy of automated approaches is constrained by the semantic nature of the contents and frequently necessitates manual intervention. Online social networks generate enormous amounts of data, making it computationally difficult to perform the operation in real time. This study suggests a framework for spotting false content that draws on machine learning and natural language processing (NLP) ideas. The proposed method would make it possible to stop the dissemination of false material on the Twitter network. The Artificial Learning (ML)*

Key Words: **AdaBoost, Web content, Fake news, SVM, KNN**

1. INTRODUCTION

One of the more well-known microblogging platforms is Twitter. Real-time news dissemination is made possible through Twitter. Tweets, or brief communications with a character limit of 140, are the primary means of information dissemination on Twitter. By following someone, the system enables one to subscribe to their tweets. Retweeting the tweets one has received, it enables speedy information distribution. As a result of the ability to submit tweets from mobile devices like smartphones, tablets, and even SMS, Twitter has become the go-to informational platform for many consumers. Twitter is a medium for easily disseminating false information because of these features [1].



The accompanying figure [2] shows how there were more media consumers between 2010 and 2021. One of the most popular social networking services is Twitter. A significant portion of users now view Twitter as a global platform for news dissemination, opinion sharing, and interpersonal interaction. As a result, there is the potential for using such a large volume, high speed spike of Twitter data generated every second for important analytical and interpretative applications.

It has become extremely popular to spread incorrect information using Internet communication methods. With the advent of social networks, every user is now a self-publisher, without editing, fact-checking, or any sort of responsibility. For millions of users, seeing the facts on their screen serves as proof that the material is accurate even when it is delivered without authority. The book's discussion of the use of technology by individuals to encourage lies, deception, fraud, spin control, and propaganda has come to pass as a result of the misuse of online social networks like Twitter. It can be difficult and fraught with traps for both novice and seasoned Internet users to validate the info on the web.

This has to be adequately discovered. Only a small number of organizations verify the accuracy of the news before publishing it. Many data sets are being manually detected as fraudulent, yet they cannot be manually accepted or removed because of the sheer volume. Therefore, the news content should be qualified using an automatic approach. The suggested method could create a way to determine whether a certain tweet is phony or original based on the words, phrases, relevant sources, and title. The following contributions are made to this body of research:

- An automated system for determining whether popular Tweets are real or fraudulent,

- This paper evaluates the suggested model's predictive performance, demonstrating the model's efficacy.

- A three-dataset alignment that captures accuracy judgments for both true and false tweets.

2. LITERATURE REVIEW

Emma Cueva et al., proposed an effective model to identify the fake news spread on Twitter through the use of Artificial Intelligence (AI). To assess the effectiveness of different networks at identifying fake news, the analysis specifically looked at Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and Natural Language Processing (NLP) networks. Data was cleaned up and then utilized to develop AI systems. The NLP model was the only version that could recognise comedy as false news, despite the fact that all three models were successful in recognising fake news with high accuracy. This is why it was decided that the NLP model was the best option for spotting bogus news on Twitter [3].

SreeJagadeesh Malla et al., investigated COVID-19 fake data from various social media platforms such as Twitter, Facebook, and Instagram. The authors have tested various deep learning models on the COVID-19 fake dataset. Finally, deep learning models like BERT, BERTweet, AlBERT, and DistlBERT performed worse than the CT-BERT and RoBERTa models. Using the multiplicative fusion method, the proposed ensemble deep learning architecture beat CT-BERT and RoBERTa on the COVID-19 false news dataset. The multiplicative product of the final prediction values of CT-BERT and RoBERTa was used to assess the performance of the suggested model in this technique. This method eliminates the drawback of the inaccurate prediction nature of these CT-BERT and RoBERTa models. With 98.88% accuracy and a 98.93% F1-score, the suggested architecture beats both well-known ML and DL models [4].

Phayung Meesad et al., proposed a framework for robust Thai fake news detection. Information retrieval, natural language processing, and machine learning are the framework's three core parts. It has two phases, including the phase where data is collected and the phase where a machine-learning model is built. In order to extract useful features from web data, they used natural language processing techniques to examine data that they had collected from Thai online news websites utilizing web-crawler information retrieval. They chose a number of well-known machine learning models for classification for comparison, including Naive Bayesian, Logistic Regression, K-Nearest Neighbor, Multilayer Perceptron, Support Vector Machine, Decision Tree, Random Forest, Rule-Based Classifier, and Long Short-Term Memory. They implemented an automatic online web application for fake news detection after the comparative analysis of the test set revealed that Long Short-Term Memory was the best model [5].

Diaz-Garcia presented a solution based on Text Mining that tries to find which text patterns are related to tweets that refer to fake news and which patterns in the tweets are related to true news. To test and validate the results, the system faces a pre-labeled dataset of fake and real tweets during the U.S. presidential election in 2016. In terms of results, interesting patterns are obtained that relate the size and subtle changes of the real news to create fake news. Finally, different ways to visualize the results are provided [6].

3. System Overview

A user-friendly interface and a PYTHON software were employed to implement the system. Twitter datasets are employed in this system to get the experimental results.

To test the given dataset using ML classification methods, a Python script utilizing the Scikit-learn module was created. This script accurately identified the given data as belonging to the "real" and "fake" classes of the Twitter dataset. The Scikit-learn module was chosen because Python is one of the most widely used programming languages for scientific computing and contains a large selection of scientific libraries.

i. Data Collection

Having a thorough and trustworthy dataset is necessary before beginning the opinion mining process. Twitter media could be used to gather the required data. This study used a Twitter dataset about metformin and related branded medications to create a classification algorithm. Twitter provides customers with a streaming Application Programming Interface (API) that enables them to collect data in real-time access the Twitter Streaming API, you must first obtain your Twitter API keys (API key, API secret, Access token, and Access token secret). The "Tweepy" library files are used to access the Twitter Streaming API and retrieve Twitter data. This utility makes it simple to communicate between computer programmes and web services. R-Tool was used in this study to extract data from recently posted tweets by users.

Pre- Processing : To get decent results while analyzing tweets, it is crucial to use text pretreatment techniques [8]. The preprocessing stage involves a number of stages, including,

• URL Removal: Opinion analysis has nothing to do with URLs. Therefore, for successful analysis, URLs should be deleted from the tweets.

• Lowercase Text Conversion: Tweets will contain text that combines upper- and lower-case letters [6]. As a result, Twitter data is changed to lowercase making it simpler to analyze.

• User name Removal: Nearly every sentence in a Twitter text has a user name. There is no opinion in their presence. Therefore, removing it during the pre-processing stage is a crucial step.

- ***Removing the Punctuations (#, @, etc,):*** Punctuations do not share any contribution toward analyzing the opinion of a person. Hence, they should be removed to make the analysis process easy.
- ***Remove Blank spaces:*** This step is used to remove the unwanted blank space which helps for the tokenization of the tweets.
- ***Tokenization:*** Tokenization means breaking the sentence into words.
- ***Removing stop words:*** The proposed method loads a complete list of English stop words that is gathered in Python programming language and removes it to reduce the computational load.
- ***Lemmatization:*** This step aims to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.
- ***Stemming:*** It refers to a basic experimental process that chops off the ends of words.

ii. Feature extraction Algorithms

A feature extraction method is used to decrease unintended fluctuations in the twitter data and prevent computationally expensive operations in order to make efficient use of the learning system. Enhancing the robustness of a feature typically requires less work from the classifier to improve the effectiveness of the classification system. The process of feature extraction begins with a base set of measured (preprocessed) data and creates derived values that are meant to be useful. This technique speeds up the next learning phases and, in some situations, improves human interpretations [9]

a. Term Frequency- Inverse Document Frequency (TF-IDF)

It is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction [10]. TF-IDF is a weight metric which determines the importance of word for that document.

Term Frequency:

Term Frequency measures number of times a particular term t occurred in a document d. Frequency increases when the term has occurred multiple times.

Inverse Document Frequency:

TF processes simply the frequency of a word/term 't ' present in dataset. Some terms like stop words occur multiple times but may not be useful. Hence Inverse Document Frequency (IDF) is used to measure term's importance

The above equations are tried to explain mathematical concept behind the all process like tf-idf calculation by using Term Frequency and Inverse Document Frequency. Usually TfidfVectorizer consider overall document weightage of a word. It aids us in allocating with maximum occurrence words. TfidfVectorizer weights the word counts by a measure of how often they appear in the documents. It is text vectorization based on the Bag of words (BoW) model. This method does well than the BoW model as it considers the significance of the word in a dataset (text data).

iii Machine Learning (ML) approach:

Machine learning studies the ways in which computers may learn from data. Computer programmes that automatically recognize complicated patterns and provide informed conclusions from data are a major field of research.

Opinion mining is solved by ML algorithms as a typical text classification problem using scientific features. We have a collection of training records D=X1, X2,..., Xn, each of which is labeled with the classes "A" and "B." The features in the testing record are connected to one of the class labels by the classification model. Next, a class label is predicted for a particular instance (features) of an unknown class using the model. The existence of labeled training documents is a prerequisite for supervised learning methodologies [11].

Data labeling:

In this study system, the classification procedure is carried out using a supervised machine-learning model. The learning algorithm examines the practise data and generates an inferred function that may be applied to the mapping of unknowable data (testing data). As a result, in the best case scenario, the algorithm will be able to accurately identify the class labels for instances that are not visible, such as legitimate tweets, which are labelled as 0, and fraudulent tweets, which are labelled as 1..

a. K-Nearest Neighbor (KNN)

• The KNN algorithm is crucial to machine learning systems. It falls under the category of supervised learning and has a wide range of uses in pattern recognition, intrusion detection, and other fields. These KNNs are used in situations with realistic outcomes when non-parametric techniques are required [12].

These methods don't make any assumptions about how data is distributed. The KNN approach divides the correlatives in the given dataset into clusters that can be identified by a certain trait. The main benefit of this strategy is that it produces comparable results for comparable training data. For the input population, the closest value that can classify all or some of the samples is found.

b. Support Vector Machines (SVM)

SVM is a supervised learning model that is generated for binary classification in both linear and nonlinear forms. Usually, datasets are nonlinearly inseparable, thus the main goal of the SVM method is to catch the finest available surface to make a separation between positive and negative training feature samples depending on the experimental threat (training set also test set error) reduction principal. This method can try to describe a decision boundary with the hyper-planes in a high-dimensional feature space. This hyperplane distinguishes the vectorized data into 2 classes also finds an outcome to take a decision depending on this support vector [13].

C. Decision Tree (DT)

Quinlan's DT classifier is one of the most well-known ML techniques. A "DT" is created with leaf nodes and "decision nodes." Each decision node has a number of branches, each of which contains the results of the test "X," and is related to a test "X" over a single element of the input data. Each leaf node denotes a group that is the result of a case's decision [14].

d. Proposed methodology: Ensemble Techniques

The fundamental tenet of ensemble research is that, in many cases, a combination of classifiers will result in more accurate and stable predictions than a single classifier. This is especially true when using neural networks, where the component classifiers are prone to instability. Even though the use of ensembles in ML research is still relatively new, it is well established that ensembling will enhance the performance of unstable learners. When various individual learners are combined to form only one learner then that particular type of learning is called ensemble learning [15].

Unstable learners are learners where small changes in the training data can produce quite different models and thus different predictions. Thus, a ready source of diversity **AdaBoost with Ensemble**

SVM-KNN algorithm

In classification algorithms, each one has its own advantage and disadvantage. So, AdaBoost with Ensemble SVM-KNN algorithm is compared with above-mentioned algorithms to achieve the highest accuracy than others [16]. The working mechanism of the proposed algorithm is explained in the below section. In proposed technique, K nearest neighbor technique finds the distance between test sample and training sample. An important task of KNN is to find out the neighbors first, and then it will classify the query sample on the majority class of its nearest neighbors. The proposed KNN-SVM ensemble classification approach can be used effectively for pancreatic cancer detection with low computational complexity in the training and detection stage. The lower computational complexity property is gained from KNN classification approach that does not require the construction of feature space. KNN algorithm has been used in the proposed hybrid approach KNN-SVM as the first step in the pancreatic tumor detection, and then the SVM method is employed in the second stage as a classification engine of this hybrid model.

Adaboost is an iterative boosting approach to improve the classification of weak classifiers. At the initial stage, the Adaboost algorithm will allocate variant weights to each observations. After a few iteration, the weight imposed on the misclassified observations will increase, and vice versa, the correctly classified will have lesser weights. The weights on the observations are the indicators as to which class the observation belongs to, thus lower the misclassification of the observations while extremely improve the performance of the classifiers at the same time. That mainly aims at reducing variance, boosting is a technique that consists in fitting sequentially multiple weak learners in a very adaptive way: each model in the sequence is fitted giving more importance to observations in the dataset that were badly handled by the previous models in the sequence.

4. Result and discussion

This part provides a detailed explanation of the experiment's methodology and reviews the outcomes of the algorithms used in this study to determine the most effective method for detecting fake news on Twitter in terms of feature extraction. Twitter real-time datasets were used to conduct the tests.

Donald Trump-related Twitter datasets are downloaded from Twitter and used in various tests to determine if user tweets are phony or real. This section goes into further detail about how the data is gathered, how the features are selected, and how they are divided for training and testing when using a machine learning classifier to predict fake news.

Experimental Setup

The experimental environment is a 64-bit macOS system, Intel 2.6GHz 8-core i7 CPU, 16GB 2400MHz DDR4 memory, Radeon Pro 560X 4GB GPU. The compilation environment is Python 3, and NLP processing system is selected as the machine learning framework. Various steps involved in this system are as explained as follows.

Step 1: Create twitter account

Step 2: Now we have to authenticate with twitter by using consumerKey, consumerSecret, accessToken, accesSecret.

Step 3: Fetching tweets from twitter and saving tweets into .csv files

Step 4: There is lot of noise in the collected data like "@" "/" ":" "##", which has to be removed from the data, therefore data cleaning/pre-processing is required.

Step 5: Surplus data is present in the preprocessed data, but using this given data as it is, may not produce desired output as data contains irrelevant items which makes it tough to handle. So it is always suggested to feature extraction process which can used for classification process.

Step 6: Supervised machine learning techniques used to take labeled data, which already have categorized in to the available classes. That's why we need to assign label such as "0" or "1" to the reviews according to their characteristics.

Step 7: Finally, classification algorithms are implemented to predict the fake or real tweets about Donald trump.

Applying text preprocessing steps before analyzing the tweets is very important for achieving good results. The purpose of the preprocessing is to illuminate the input data for further analysis. The following table illustrated the output of pre-processing,

i. Evaluation of accuracy

• In the proposed system, the prediction accuracy has been assessed based on the various training messages that are then optimized by the glow worm swarm optimization algorithm. The accuracy attained using various feature selection techniques for the SVM algorithm is displayed in the following table.

Performance Evaluation

The proposed technique is evaluated with other techniques by measuring the following evaluation parameters [15],

- Accuracy: It is the percentage of test set tuples that are correctly classified by the classifier.
- Precision: It is the ratio of predicted positive examples which really are positive
- Recall or sensitivity: it measures how much a classifier can recognize positive examples
- F measure: It is to combine precision and recall into a single measure.

From the below table and figures, it is observed that proposed feature extraction technique have highest values when compared with TF-IDF and N-Grams algorithms for the SVM classification. The precision, recall and F1-Score obtained by different techniques are shown in below figures respectively.

Table1. Comparison of Proposed Model with SVM Algorithms using Accuracy value

S.No	Algorithms	Accuracy
1	DT	84
2	KNN	87
3	SVM	89
4	Proposed	90

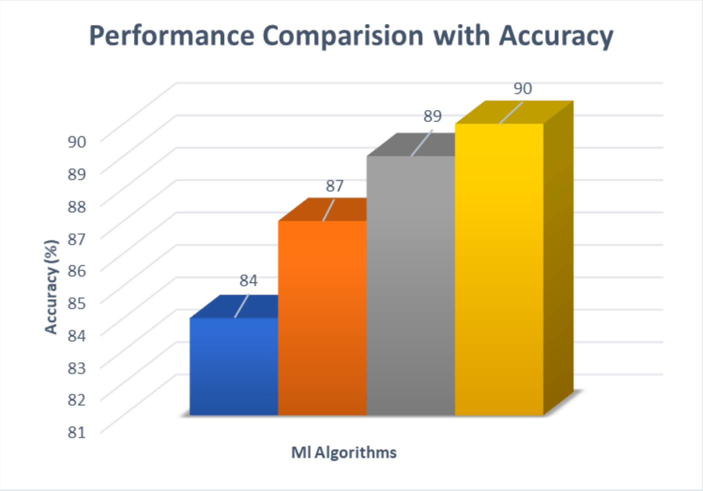
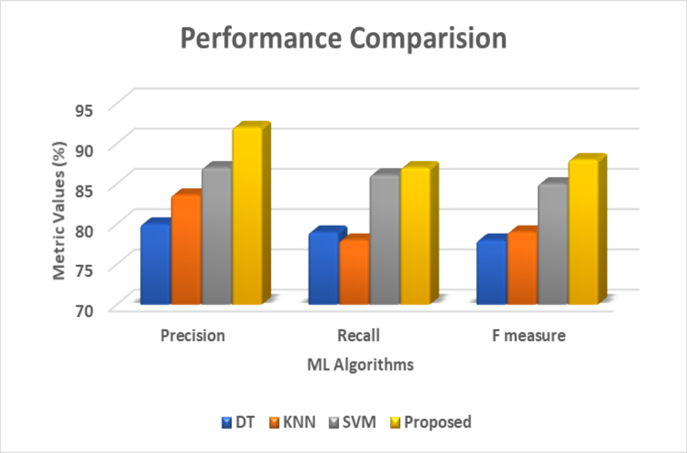


Table: Performance comparison of various machine learning with various evolution parameters

Algorithm	Precision	Recall	F measure
DT	80	79	78
KNN	83.6	78	79
SVM	87	86.1	85
Proposed	92	87	88



The above figure clearly shows accuracy of the proposed algorithm has a maximum accuracy when compared with the other ML techniques. Classification performance of the proposed classifier is better than the other classifiers and it is illustrated from the results in above tables. The proposed classifier yields better classification accuracy, because it has a regularization parameter, which avoids over-fitting. Above Table exhibits the performance metrics comparison of the various classifiers, for fake / real case respectively.

5. Conclusion

In recent years, due to the booming development of online social networks, fake news for various commercial and political purposes has been appearing in large numbers and is widespread in the online world. An important goal in improving the trustworthiness of information in online web content is to identify fake news timely. The proposed system was developed to classify the real or fake data in twitter dataset. In this work various steps are involved for this process such as data collection, preprocessing, feature extraction and classification. The Precision, Recall and F1 score metrics are applied to find the best ML approach for classification. In a more extensive, this research proposes a novel ensemble machine learning algorithm for fake news detection in web content to increase the accuracy and reduce misclassification rate.

6. Reference

1. Kumar, K.P.K., Geethakumari, G. Detecting misinformation in online social networks using cognitive psychology. Springer, Hum. Cent. Comput. Inf. Sci. 4, 14 (2014).

2. https://www.statista.com/topics/1164/social-networks/#topicHeader__wrapper

3. E. Cueva, G. Ee, A. Iyer, A. Pereira, A. Roseman and D. Martinez, "Detecting Fake News on Twitter Using Machine Learning Models," *2020 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 2020, pp. 1-5, doi: 10.1109/URTC51696.2020.9668872.

4. Malla, S., Alphonse, P.J.A. Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news. Springer, Eur. Phys. J. Spec. Top. (2022).

5. Meesad, P. Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning. Springer, SN COMPUT. SCI. 2, 425 (2021).

6. Diaz-Garcia J.A., Fernandez-Basso C., Ruiz M.D., Martin-Bautista M.J. (2020) Mining Text Patterns over Fake and Real Tweets. In: Lesot MJ. et al. (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2020. Communications in Computer and Information Science, vol 1238. Springer, Cham.

7. M. Čišija, E. Žunić and D. Đonko, "Collection and Sentiment Analysis of Twitter Data on the Political Atmosphere," 2018 14th Symposium on Neural Networks and Applications (NEUREL), 2018, pp. 1-5.

8. E. Nugraheni, "Indonesian Twitter Data Pre-processing for the Emotion Recognition," 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2019, pp. 58-63.

9. R. S. Karan, K. K. Shirsat, P. L. Kasar and R. Chaudhary, "Sentiment Analysis on Twitter Data: A New Aproach," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1-4.

10. K. U. Manjari, S. Rousha, D. Sumanth and J. Sirisha Devi, "Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm," 2020 4th International Conference on Trends in

Electronics and Informatics (ICOEI)(48184), 2020, pp. 648-652.

11. Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. Springer, Electron Markets 31, 685–695 (2021).

12. Alotaibi, A.S. Hybrid Model Based on ReliefF Algorithm and K-Nearest Neighbor for Erythemato-Squamous Diseases Forecasting. Arab J Sci Eng 47, 1299–1307 (2022).

13. Goswami M., Sebastian N.J. (2022) Performance Analysis of Logistic Regression, KNN, SVM, Naïve Bayes Classifier for Healthcare Application During COVID-19. In: Raj J.S., Kamel K., Lafata P. (eds) Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 96. Springer, Singapore.

14. Ligthart, A., Catal, C. & Tekinerdogan, B. Systematic reviews in sentiment analysis: a tertiary study. Artif Intell Rev 54, Springer, 4997–5053 (2021).

15. Priya Iyer, K.B., Kumaresh, S.: Twitter sentiment analysis on coronavirus outbreak using machine learning algorithms. Eur. J. Molecul. Springer, Clin. Med. 07(03), 2663 (2020).

16. Luo, S., Dai, Z., Chen, T. et al. A weighted SVM ensemble predictor based on AdaBoost for blast furnace Ironmaking process. Appl Intell 50, Springer, 1997–2008 (2020).