International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228310

Evaluation of Machine Learning Approaches for Classification of Fake News

Reeya Baria[1], Sheshang Degadwala[2], Rocky Upadhyay[3], Dhairya Vyas[4]

[1]Research Student, Department of Computer Engineering, Sigma Institute of Engineering, Vadodara, Gujarat, India

[2]Associate Professor, Department of Computer Engineering, Sigma Institute of Engineering, Vadodara, Gujarat, India

[3]Assistant Professor, Department of Computer Engineering, Sigma Institute of Engineering, Vadodara, Gujarat, India

[4]Managing Director, Shree Drashti Infotech LLP, Vadodara, Gujarat, India

Article Info
Volume 8, Issue 3
Page Number 30-44

Publication Issue :
May-June-2022

Article History
Accepted: 01 May 2022
Published: 08 May 2022

ABSTRACT

Due to the obvious easy accessibility and exponential increase of information accessible on social media channels, differentiating between bogus and authentic information has become challenging. As a consequence, some scholars are concentrating on identifying bogus news. The bulk of saliency detection tools focus on the device's linguistic properties. However, they have problems recognizing particularly ambiguous false news, that could only be detected after establishing the content and most current linked information. To solve this problem, this research will provide a new Indian false news detection method based on a factual data base that's also generated and refreshed by human morality after accumulating evident facts. Our system takes a hypothesis and scans the Fact central database for conceptually similar stories in order to assess whether the given claim is false or not before contrasting it to the similar stories. To bypass these limits, the review will describe a unique matching strategy that takes use of all the article streamlining and entity discovery sets.

In this survey we learned different machine learning algorithms and its functionality with its benefits and downsides.

Keywords— Indian Language, social media, Sematic Feature, Entity Finding and Machine Learning

I. INTRODUCTION

In less than a generation, social media has evolved from a The transformation of direct electronic information interchange into a virtual meeting place, commerce platform, and critical 21st-century marketing instrument A "online community," according to Merriam-Webster, is defined as "a form of electronic communication (such as websites for social networking and micro blogging) through which

people can form online communities to share information and ideas, to send personal messages, and to post other types of content" (such as videos). [1,3-5] Throughout the remainder of this chapter, will examine the roots of social media, its relatively quick emergence as a social and economic force, and the changes it has brought about in the marketing industry. Following the introduction of blogging, social media has seen a meteoric rise in popularity. In the beginning, sites such as LinkedIn gained in popularity, while others allowed for the sharing of photos on the internet. With the advent of YouTube in 2005, a completely new method for people to interact and share content over great distances was unveiled. When Facebook and Twitter were launched in 2006, they were accessible to individuals all around the globe. These sites continue to be among the most popular social networking sites on the internet today. Among the earliest social networking sites to develop to fit specific social networking niches were Tumblr, Spotify, Foursquare, and Pinterest. [9] There are a variety of social networking sites accessible today, and many of them may be connected together to allow for cross-posting of information. This creates an environment in which users may interact with the biggest number of people possible while yet keeping the intimacy of one-on-one communication with those persons. As for the future of social networking, It can only speculate [12] about how it will develop. However, although it has many benefits, it also has certain disadvantages.

On social media, misinformation is not a new phenomenon to be found. Every day, will read a great deal of information on social media, some of which is legitimate, but the vast majority of which is not. As a consequence of this false or misleading information, fake news is created, which is made up of made-up articles that lack any confirmed facts, sources, or quotations to back them up. That information is manufactured with the intent of inducing or fooling the reader [15]. For the most part of the period Fake news is referred to by a variety of words, including misinformation, disinformation, and misinformation. In the context of misinformation, incorrect information is given by someone who thinks it to be true and distributes it to the public.

Fake news comes in a variety of forms [1-15]:

[1] “A satirical or parodic piece (No intention to cause harm but has potential to fool)”

[2] “When the headlines, images, or captions do not reflect the substance of the article, this is known as a false connection.”

[3] “Content that is deceptive (Misleading use of information to frame an issue or an individual)”

[4] “When actual material is presented with incorrect contextual information, this is referred to as false context.”

[5] “Content Created by Impersonators (When genuine sources are impersonated with false, made-up sources)”

[6] “Content that has been modified (When genuine information or imagery is manipulated to deceive, as with a doctored photo)”

[7] “Content that has been fabricated (New content is truly false, designed to deceive and do harm)”

II. Literature Study

Tao Jiang et.al [1] discussed how they obtained the text representation using tokenization approaches such as TF, TF-IDF, and embedding. Following that, they trained individual models on these text representation characteristics, including five machine learning models such as LR, DT, KNN, RF, and SVM, as well as three deep learning models such as LSTM, GRU, and CNN. They used a corrected version of McNemar's test to see whether the model with the greatest accuracy differed significantly from other models on both datasets in order to identify the best individual model. Finally, they introduced our stacking strategy of training another RF model based on the prediction outcomes of all individual models to increase each model performance.

Khubaib Ahmed Qureshi et.al [2]. claim that newly suggested source-based techniques depend on user information to detect fake news. With this concept,

many of the flaws of prior techniques may be easily remedied. As a consequence, they provide a source-based method that employs information about the source and the propagators of the information. By studying the context of the spreader's community on social media microblogs or social networks, the approach identifies fake news. The solution investigates such communities' connection patterns as well as the profile traits of their members. They look at two strategies that integrate propagation network characteristics with user profile-based information, one at the node level and the other at the community level (aggregated user and network features, aggregated network features). The trials' findings show that two methods for identifying bogus news are highly efficient.

HUNG-YU KAO et.al [3] introduced the MVAN which combines two attention processes, text semantic attention and propagation structure attention, to collect significant hidden hints and information in the source tweet text and propagation structure at the same time. MVAN has high performance and acceptable interpretation capabilities, according to the assessment findings utilising two public data sets. MVAN can also enable early identification of bogus news with good results. They think that MVAN can be utilised not only for detecting false news, but also for sentiment classification, subject classification, insult detection, and other text classification tasks on social media.

P. K. Verma et.al [4] Automatic false news identification is a thought-provoking subject in deception detection, with enormous political and societal implications in the real world. Deep Learning (GRU) was utilised in this study to analyse discourse segments and build a dependency tree that provided distinct properties for true and fraudulent news.

A. Uppal et.al [5] is an example. Although accuracy is not usually the most important criteria for assessing a model, in this situation, high accuracy indicates that the model worked well since the datasets were balanced. TFIDF performed best on the Kaggle dataset for CNN. With the TFIDF, all three models performed well on the Kaggle dataset. When the dataset contains long paragraphs for each news item, it is possible to conclude that TFIDF is a good method.

P. Vlqjk et al [6] discuss example which attempts to solve problems with early detection of bogus news. While user comments might be beneficial in analysing news documents, they focused on the fact that there are few comments in the early phases of news distribution. As a consequence, they developed a Grover-based neural network model that provides comments to assist categorization. To assess the efficacy of our proposed approach in early detection, they experimented with making comments.

Y. Yanagi et.al [7] provide a Korean false news detection system based on a fact database generated and updated by human judgement in this study. Our algorithm takes a proposition as input and searches for comparable articles to determine whether the article discovered by the given proposition matches the proposition semantically. To do so, they employed the BiMPM model, which is a deep learning model for sentence matching.

K. H. Kim et.al [8], Despite the fact that several machine learning algorithms have shown to be effective in recognising fake news and communications. However, owing to the ever-changing characteristics and features of fake news on social media networks, classification of false news is getting increasingly complex. On the other hand, deep learning is well-known for its capacity to calculate hierarchical features. Many research works will implement deep learning methods, such as convolutional Neural Networks, deep Boltzmann machine, Deep neural network, and Deep autoencoder model, in a variety of applications, such as audio and speech processing, natural language processing and modelling, information retrieval, objective recognition, and computer vision, as a result of the recent implementation of deep learning research and applications.

S. I. Manzoor et.al [9] Because it uses a statistical approach to allow a machine to learn from data, Machine Learning is commonly used in the identification of false news. The techniques for collecting parameters and categorising different forms of news are also presented. According to the results, the dataset is first pre-processed using methods including stop word removal, tokenization, and stemming. The TF-IDF and probabilistic context-free grammar techniques for extracting features are also discovered. According to the literature study, the accuracy for predicting false news in social media is substantially greater than any other online news media, thus they've concentrated on online news media fake news detection and website verification.

S. Gaonkar et.al [10], for example, the novel and challenging problem of identifying false news in COVID-19 outbreaks on social media microblogs is the focus of this study. Despite the fact that there are numerous research on how to recognise fake news in other industries, there are few publications on how to detect false news in politics and entertainment. COVID-19 outbreaks. There are four primary approaches to dealing with this issue. Content-based approaches focus on the structure of the data. The Propagation/Network Ways are used to abuse the propagation network. The components of both systems are combined in hybrid techniques. Finally, Newly suggested source-based techniques depend on user information to detect fake news. This strategy is easy to put into practise. Many of the flaws of prior methods are eliminated. As a consequence, they provide a source-based method that takes use of information about the source as well as the propagators of the information. By studying the context of the spreader's community on social media microblogs or social networks, the approach identifies fake news.

Ankit Kesarwani et.al [11] offer a simple method for spotting bogus news on social media using the K-Nearest Neighbor classifier. This model has a classification accuracy of about 79 percent when evaluated against the Facebook news postings dataset.

A. Kesarwani et.al [12] to address this situation, HAGER SALEH and colleagues present fresh methodologies based on Machine Learning (ML) and Deep Learning (DL) for a false news identification system in [12]. The major goal of this research is to find the best model for achieving high accuracy. As a result, they offer a Convolutional Neural Network model that is optimised for detecting bogus news (OPCNN-FAKE). Using four fake news benchmark datasets, they compare the performance of the OPCNN-FAKE with the Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and the six regular ML techniques: Decision Tree (DT), logistic Regression (LR), K Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB). The parameters of ML and DL were optimised using grid search and hyperopt optimization approaches, respectively. In addition, standard ML models employed N-gram and Term Frequency Inverse Document Frequency (TF-IDF) to extract features from the benchmark datasets, whereas DL models used Glove word embedding to encode features as a feature matrix.

Nicholas Snell et.al [13] Detecting and alerting human readers to purposely false and manipulative news articles is critical to limiting the harm they may do. The dataset described in this study contains news items that have been manually recognised and categorised and may be used to train and evaluate classification algorithms that can distinguish between real and fraudulent and manipulated news stories.

Yuta Yanagi et.al [14] present a false news detector that can produce fake social situations, with the goal of detecting fake news early in its spread when few social contexts are accessible. A fake news generator model is used to create the bogus context. Using a dataset of news items and their social settings, this model was trained to produce comments. They also trained a classification model. This was done with the use of news items, real-time comments, and created

remarks. To assess the efficiency of our detector, they compared the performance of produced comments for articles with genuine and generated comments created by the classifying model. As a consequence, they believe that examining a manufactured comment aids in the detection of bogus news more effectively than evaluating just actual comments. It implies that their suggested detector would be successful in detecting bogus news on social media.

Terry Traylor, et.al [15] The findings of a fake news detection investigation that records the performance of a fake news classifier are reported in this article. The Text blob, Natural Language, and SciPy Toolkits were used to create a unique fake news detector that use cited attribution as a crucial feature in a Bayesian machine learning system to predict the chance of a news storey being fraudulent. The method precision as a consequence is 63.333 percent successful in determining whether or not an article with quotations is phoney. This procedure is known as influence mining, and it is touted as a way for detecting bogus news and even propaganda. The study procedure, technical analysis, technical linguistics work, classifier performance, and findings are all discussed in this publication. The study continues with a consideration of how the present system will develop into a system that mines influence.

Dwivedi, Sanjeev M et.al [17] concentrate on five important areas in particular. They begin by reporting on and debating the many definitions of false news and rumours that have been discussed in the literature. Second, they show how difficult it is to obtain useful data for detecting false news and rumours, as well as the numerous ways that have been used to collect this data, as well as publicly accessible datasets. Third, they go through the qualities that have been taken into account in false news and rumour detection methods. Fourth, they present a thorough examination of the many strategies used to identify rumours and false news. Finally, they discuss and determine future directions.

III. Proposed System

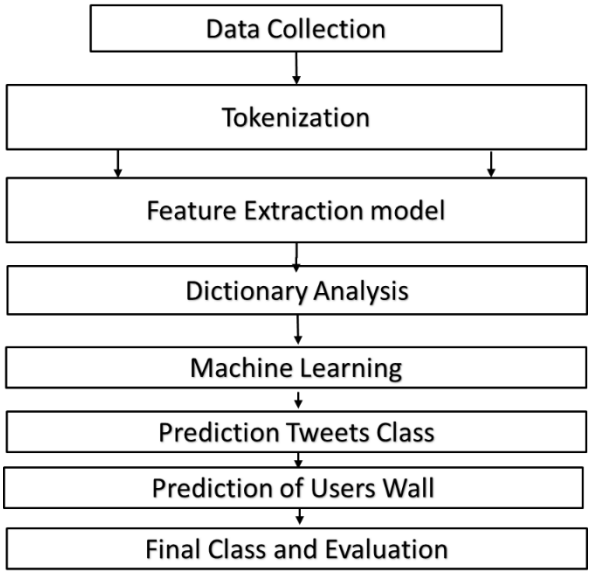

Fig.1 Proposed System

Step 1: Collect the data from the social media and offline sources according to given headline and location.

Step 2: Pre-process the collected data into meaningful and useful words and convert symbols(emoji) into text etc.

Step 3: Using various feature extraction model and dictionary, calculate score of each word or text.

Step 4: Using the word feature list, create feature vectors. This will be an input to our machine model.

Step 5: Creating a list of positive reviews and it further classified into very good and good reviews feature using the classification.

Step 6: Creating a list of negative reviews and it further classified into very bad and bad reviews feature using classification.

Datasets

Sentiment140 dataset contains 1,600,000 tweets extracted from Twitter by using the Twitter API.

The tweets have been categorized into three classes:

- 0:negative
- 2:neutral
- 4:positive

The information contained in the dataset:

- The polarity of the tweet
- id of the tweet
- date of the tweet
- query
- User that tweeted
- The content of the tweet.
- Dataset size: 305.13 MB

There are two types of geographical metadata when working with Tweet data: When a person discloses their location at the time of a Tweet, this information is available. Account Location - This is determined by the user's public profile's 'home' location. This is a free-form character field that may or may not contain geo referenced metadata. Users can select a location for individual tweets on Twitter. Through its many operators, Power Track provides multiple options to filter for Tweets based on Tweet-specific location data.

Tweets with a precise "Point" coordinate of latitude and longitude. Tweets with a Point coordinate are delivered from GPS-enabled devices and indicate the precise location of the Tweet in question. There is no contextual information about the GPS location being referenced in this sort of location.

Data pre-processing [1,11,14,15]

The information gleaned via Twitter is unorganised and erratic.

- Tweet Tokenization: The segmentation or tokenization of tweets is the first step in the data preparation process. In text analysis, tokenizing words is a key unit.
- Punctuation signs such as periods, semicolon, commas, ellipses, exclamation points and quotation marks are removed from the dataset in the second phase of twitter processing.
- There are no more "stop words." Stop words are the most frequently used words in the text. The significance of words that appear often in a piece of writing is negligible.
- Converting lowercase letters to uppercase: Upper and lower case characters are considered equal in text analysis. When they use both upper and lowercase letters in our training corpus, the number of feature words rises.
- Stemming: Stemming is a crucial part of data pre-processing. Words are reduced to their most basic form by the removal of affixes that save both time and space.

Feature Extraction [1-5,11-15]

There are a variety of methods for identifying and extracting characteristics.

TF-IDF (term frequency-inverse document function): TF-IDF (term frequency-inverse document function): Terms frequency and inverse document function are abbreviated as TF-IDF. When a phrase appears a lot in a document, its importance and weight are diminished.

According to the formula Wt,d, the weight of a word in the document d is determined as follows: To get the logarithm of (N/DFt): There are a total of TFt instances in Document D. The total number of documents in the corpus is N, and DFt is the number of documents that include the term t.

Words in a Pouch: The number of times a word appears in a document is represented in text by a "bag of words." Text characteristics may be easily and flexibly extracted using this approach. Without attention to grammar or word-order, it tracks word counts while ignoring them. They call this "bag of words" since there is no way to reconstruct the text's original order or organisation.

Emoticons Feature

People are increasingly using emoticons in their writing to communicate their feelings or recollect their remarks. Earlier machine learning algorithms focused primarily on the classification of text, emoticons, or graphics, with emoticons alongside text being overlooked, ignoring a wide range of emotions. This algorithm/method uses both text and emoticons to analyze sentiment. Machine learning techniques are used to find sentiments from twitter-based airline

data employing many features such as TF–IDF, N-gram, and emoticon lexicons in combination and separately. When emoticons are used with their most similar text, or their related sentiment dominates the sentiment indicated by textual data analysis, this feature will replace them.

Synonyms Feature

Synonyms are words that have the same or very similar meaning but differ in spelling and sound. "Aircraft" and "airplane," for example, are both synonyms for "plane." Synonyms are commonly used to adorn text in information retrieval to increase the likelihood that an acceptable query will match.

We can search the dictionary for comparable tweets using these synonyms and return the results using all of these synonyms.

N-Gram Model

An N-gram model predicts the most likely word to follow a sequence of N-1 words given a set of N-1 words. It's a probabilistic model that has been trained on a text corpus. Many NLP applications, such as speech recognition, machine translation, and predictive text input, benefit from such a model.

An N-gram model is created by counting the number of times word sequences appear in corpus text and then calculating the probability. Because a simple N-gram model has limitations, smoothing, interpolation, and backoff are frequently used to improve it.

One type of Language Model (LM) is an N-gram model, which is concerned with determining the probability distribution over word sequences.

Unigram, bigram, and trigram models are also available. Unigram is a model that depends solely on the frequency of a word without considering prior words. Bigram refers to a model that uses only the prior word to predict the following word. It's a trigram model if two previous words are taken into account.

Hybrid Dictionary

- VADER (Valence Aware Dictionary for Sentiment Reasoning) is a text sentiment analysis model that is sensitive to both emotion polarity (positive/negative) and intensity (strong). It's included in the package and may be used on unlabeled text data right away.

VADER sentimental analysis uses a dictionary to map lexical information to emotion intensities, which are referred to as sentiment scores. A text's sentiment score is calculated by adding the intensity of each word in the text.

Learning Algorithm:

- SVM [1,3]: A help vector machine (SVM) is a regulated AI technique that is utilized for order. SVM develops a hyperplane or set of hyperplanes in a high dimensional space, which can be utilized for order or different assignments like recognition of exceptions from information. A decent arrangement is accomplished by the hyperplane that has the biggest separation to the closest preparing information purpose of any class.
- Decision Tree [9,10]: Decision trees orchestrate data in a tree-like structure, grouping the information into different branches. Each branch speaks to an elective choice. The tree-like model speaks to the choices and their potential results and utility. It very well may be additionally joined with different calculations.
- k-NN [1,10]: K-closest calculation (KNN) calculation is a regulated AI calculation. It tends to be utilized for both order and relapse prescient issues. An article is characterized dependent on its closest neighbor's democratic framework. In the k-NN, k is a client characterized steady. It is a non-parametric and apathetic learning calculation.
- RF [1-5]: Random timberland, similar to its name suggests, comprises of countless individual choice trees that work as an outfit. Every individual tree in the irregular woodland lets out a class expectation and the class with the most votes turns into our model's forecast.

- NB [7]: Naive Bayes is an AI model that is utilized for enormous volumes of information, regardless of whether you are working with information that has a large number of information records the suggested approach is Naive Bayes. It gives excellent outcomes with regards to NLP assignments, for example, wistful investigation. It is a quick and straightforward arrangement calculation.

IV. Results and Analysis

The data you use to train an algorithm or machine learning model to anticipate the outcome you want it to predict is known as training data. Test data is used to evaluate the algorithm you're using to train the machine's performance, such as accuracy or efficiency. Data is necessary for machine learning models to work. Even the most performant algorithms can be rendered useless without a foundation of high-quality training data. Indeed, when machine learning models are trained on insufficient, erroneous, or irrelevant data in the early stages, they might be handicapped. When it comes to training data for machine learning, the adage still holds garbage in, garbage out.

	target	comment_text
0	4	Reading my kindle2... Love it... Lee childs i...
1	4	Ok, first assesment of the #kindle2 ...it fuck...
2	4	@kenburbary You'll love your Kindle2. I've had...
3	4	@mikefish Fair enough. But i have the Kindle2...
4	4	@richardebaker no. it is too big. I'm quite ha...
...	...	...
492	2	Ask Programming: LaTeX or InDesign?: submitted...
493	0	On that note, I hate Word. I hate Pages. I hat...
494	4	Ahhh... back in a *real* text editing environm...
495	0	Trouble in Iran, I see. Hmm. Iran. Iran so far...
496	0	Reading the tweets coming out of Iran... The w...

497 rows × 2 columns

Fig.2 Reading Offline Tweet Dataset

	comment_text	Type
0	Reading my kindle2... Love it... Lee childs i...	Positive
1	Ok, first assesment of the #kindle2 ...it fuck...	Positive
2	@kenburbary You'll love your Kindle2. I've had...	Positive
3	@mikefish Fair enough. But i have the Kindle2...	Positive
4	@richardebaker no. it is too big. I'm quite ha...	Positive
...	...	...
484	Monday already. Iran may implode. Kitchen is a...	Negative
489	I just created my first LaTeX file from scratc...	Negative
493	On that note, I hate Word. I hate Pages. I hat...	Negative
495	Trouble in Iran, I see. Hmm. Iran. Iran so far...	Negative
496	Reading the tweets coming out of Iran... The w...	Negative

497 rows × 2 columns

Fig.3 Labeling Offline Tweet Dataset

	comment_text	Type	clean_tweet
0	Reading my kindle2... Love it... Lee childs i...	Positive	read kindl love child good read
1	Ok, first assesment of the #kindle2 ...it fuck...	Positive	first asses kindl fuck rock
2	@kenburbary You'll love your Kindle2. I've had...	Positive	kenburbari love kindl mine month never look ba...
3	@mikefish Fair enough. But i have the Kindle2...	Positive	mikefish fair enough kindl think perfect
4	@richardebaker no. it is too big. I'm quite ha...	Positive	richardebak quit happi kindl
...	...	...	...
484	Monday already. Iran may implode. Kitchen is a...	Negative	monday alreadi iran implod kitchen disast anna...
489	I just created my first LaTeX file from scratc...	Negative	creat first latex file scratch work well amand...
493	On that note, I hate Word. I hate Pages. I hat...	Negative	note hate word hate page hate latex said hate ...
495	Trouble in Iran, I see. Hmm. Iran. Iran so far...	Negative	troubl iran iran iran away flockofseagullswere...
496	Reading the tweets coming out of Iran... The w...	Negative	read tweet come iran whole thing terrifi incred

497 rows × 3 columns

Fig.4 Tokenization of Offline Dataset

K-Nearest Neighbors

	precision	recall	f1-score	support
Normal_User	0.82	0.66	0.73	150
Suspect_User	0.69	0.58	0.63	113
Criminal_User	0.65	0.85	0.74	159
accuracy			0.71	422
macro avg	0.72	0.69	0.70	422
weighted avg	0.72	0.71	0.71	422

```
[[ 99  16  35]
 [ 10  65  38]
 [ 11  13 135]]
```

Fig.5 K-Nearest Neighbors

Liner SVM

	precision	recall	f1-score	support
Normal_User	1.00	0.91	0.95	155
Suspect_User	1.00	0.89	0.94	114
Criminal_User	0.85	1.00	0.92	153
accuracy			0.94	422
macro avg	0.95	0.93	0.94	422
weighted avg	0.95	0.94	0.94	422

```
[[141   0  14]
 [  0 102  12]
 [  0   0 153]]
```

Fig. 6 Liner SVM

Decision Tree

	precision	recall	f1-score	support
Normal_User	1.00	0.89	0.94	157
Suspect_User	1.00	0.91	0.95	120
Criminal_User	0.84	1.00	0.91	145
accuracy			0.93	422
macro avg	0.95	0.93	0.94	422
weighted avg	0.94	0.93	0.93	422

```
[[140   0  17]
 [  0 109  11]
 [  0   0 145]]
```

Fig. 7 Decision Tree

Random Forest

	precision	recall	f1-score	support
Normal_User	1.00	0.90	0.95	154
Suspect_User	1.00	0.89	0.94	117
Criminal_User	0.84	1.00	0.92	151
accuracy			0.93	422
macro avg	0.95	0.93	0.94	422
weighted avg	0.94	0.93	0.93	422

```
[[139   0  15]
 [  0 104  13]
 [  0   0 151]]
```

Fig. 8 Random Forest

ExtraTreesClassifier

	precision	recall	f1-score	support
Normal_User	1.00	0.89	0.94	155
Suspect_User	1.00	0.93	0.96	116
Criminal_User	0.86	1.00	0.92	151
accuracy			0.94	422
macro avg	0.95	0.94	0.94	422
weighted avg	0.95	0.94	0.94	422

```
[[138   0  17]
 [  0 108   8]
 [  0   0 151]]
```

Fig. 9 Extra Trees

	user	date	text	user_loc
0	gulf_news	Tue May 10 03:40:00 +0000 2022	#Pulitzer Prize: Slain photographer #DanishSid...	United Arab Emirates
1	AhsanAbbasShah	Tue May 10 03:39:57 +0000 2022	Govt to build varsities in Diplo, Sakrand: @AA...	Lahore, Pakistan
2	DailyhuntApp	Tue May 10 03:39:54 +0000 2022	'Apna chashma badal': Suniel Shetty to Twitter...	India
3	patro_anup	Tue May 10 03:39:53 +0000 2022	@rdmodisha @OdishaVigilance @MoSarkar5T @dmkor...	Jeypore Koraput Odisha
4	DrPrashantkorat	Tue May 10 03:39:49 +0000 2022	Noted Kannada writer Bhairappa praises PM Modi...	Gandhinagar, Gujarat
...	...	...	...	...
95	otvnews	Tue May 10 03:33:04 +0000 2022	#CycloneAsani: ECoR takes precautionary measur...	Bhubaneswar, India
96	techfoogle	Tue May 10 03:33:03 +0000 2022	7 Latest 3D Printer Inventions 2022 YOU WISH Y...	India
97	MarketsCafe	Tue May 10 03:33:03 +0000 2022	Rainbow Children's Medicare listing expected t...	India
98	telugustop	Tue May 10 03:33:03 +0000 2022	Global Covid caseload tops 517.8 mn https://t....	India
99	HealthSite4U	Tue May 10 03:33:02 +0000 2022	Delhi Reports Big Drop in Covid Cases On Monda...	Mumbai

100 rows × 4 columns

Fig. 10 Live Data Tweets

	user	date	text	user_loc	clean_tweet
0	gulf_news	Tue May 10 03:40:00 +0000 2022	#Pulitzer Prize: Slain photographer #DanishSid...	United Arab Emirates	pulitz prize slain photograph danishsiddiqui a...
1	AhsanAbbasShah	Tue May 10 03:39:57 +0000 2022	Govt to build varsities in Diplo, Sakrand: @AA...	Lahore, Pakistan	govt build varsiti diplo sakrand
2	DailyhuntApp	Tue May 10 03:39:54 +0000 2022	'Apna chashma badal': Suniel Shetty to Twitter...	India	apna chashma badal suniel shetti twitter user ...
3	patro_anup	Tue May 10 03:39:53 +0000 2022	@rdmodisha @OdishaVigilance @MoSarkar5T @dmkor...	Jeypore Koraput Odisha	daylight loot govt revenu highlight
4	DrPrashantkorat	Tue May 10 03:39:49 +0000 2022	Noted Kannada writer Bhairappa praises PM Modi...	Gandhinagar, Gujarat	note kannada writer bhairappa prais modi say o...
...	...	...	...	...	...
95	otvnews	Tue May 10 03:33:04 +0000 2022	#CycloneAsani: ECoR takes precautionary measur...	Bhubaneswar, India	cycloneasani ecor take precautionari measur fo...
96	techfoogle	Tue May 10 03:33:03 +0000 2022	7 Latest 3D Printer Inventions 2022 YOU WISH Y...	India	latest printer invent you wish you had right n...
97	MarketsCafe	Tue May 10 03:33:03 +0000 2022	Rainbow Children's Medicare listing expected t...	India	rainbow children medicar list expect mute rain...
98	telugustop	Tue May 10 03:33:03 +0000 2022	Global Covid caseload tops 517.8 mn https://t....	India	global covid caseload top coronavirus covid ge...
99	HealthSite4U	Tue May 10 03:33:02 +0000 2022	Delhi Reports Big Drop in Covid Cases On Monda...	Mumbai	delhi report big drop covid case monday covidc...

100 rows × 5 columns

Fig. 11 Clean Live Data

Fig. 12 Emoticons Feature

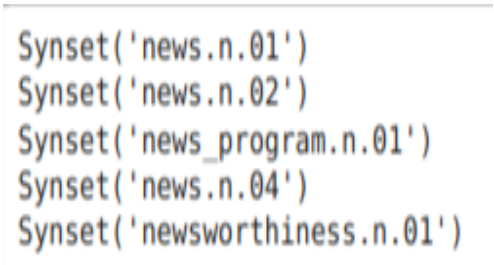

Fig. 13 Synonyms Feature (synonyms: ‘News’)

	user	date	text	user_loc	clean_tweet	unigram	bigrams	trigram	Score
0	gulf_news	Tue May 10 03:40:00 +0000 2022	#Pulitzer Prize: Slain photographer #DanishSid...	United Arab Emirates	pulitz prize slain photograph danishsiddiqui a...	[pulitz, prize, slain, photograph, danishsiddi...	[pulitz prize, prize slain, slain photograph, ...	[pulitz prize slain, prize slain photograph, s...	{'neg': 0.0, 'neu': 0.413, 'pos': 0.587, 'comp...
1	AhsanAbbasShah	Tue May 10 03:39:57 +0000 2022	Govt to build varsities in Diplo, Sakrand: @AA...	Lahore, Pakistan	govt build varsiti diplo sakrand	[govt, build, varsiti, diplo, sakrand]	[govt build, build varsiti, varsiti diplo, dip...	[govt build varsiti, build varsiti diplo, vars...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
2	DailyhuntApp	Tue May 10 03:39:54 +0000 2022	'Apna chashma badal': Suniel Shetty to Twitter...	India	apna chashma badal suniel shetti twitter user ...	[apna, chashma, badal, suniel, shetti, twitter...	[apna chashma, chashma badal, badal suniel, su...	[apna chashma badal, chashma badal suniel, bad...	{'neg': 0.181, 'neu': 0.819, 'pos': 0.0, 'comp...
3	patro_anup	Tue May 10 03:39:53 +0000 2022	@rdmodisha @OdishaVigilance @MoSarkar5T @dmkor...	Jeypore Koraput Odisha	daylight loot govt revenu highlight	[daylight, loot, govt, revenu, highlight]	[daylight loot, loot govt, govt revenu, revenu...	[daylight loot govt, loot govt revenu, govt re...	{'neg': 0.0, 'neu': 0.625, 'pos': 0.375, 'comp...
4	DrPrashantkorat	Tue May 10 03:39:49 +0000 2022	Noted Kannada writer Bhairappa praises PM Modi...	Gandhinagar, Gujarat	note kannada writer bhairappa prais modi say o...	[note, kannada, writer, bhairappa, prais, modi...	[note kannada, kannada writer, writer bhairapp...	[note kannada writer, kannada writer bhairappa...	{'neg': 0.294, 'neu': 0.706, 'pos': 0.0, 'comp...
...	...	...	...	...	...	...	...	...	...
95	otvnews	Tue May 10 03:33:04 +0000 2022	#CycloneAsani: ECoR takes precautionary measur...	Bhubaneswar, India	cycloneasani ecor take precautionari measur fo...	[cycloneasani, ecor, take, precautionari, meas...	[cycloneasani ecor, ecor take, take precaution...	[cycloneasani ecor take, ecor take precautiona...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
96	techfoogle	Tue May 10 03:33:03 +0000 2022	7 Latest 3D Printer Inventions 2022 YOU WISH Y...	India	latest printer invent you wish you had right n...	[latest, printer, invent, you, wish, you, had,...	[latest printer, printer invent, invent you, y...	[latest printer invent, printer invent you, in...	{'neg': 0.0, 'neu': 0.787, 'pos': 0.213, 'comp...
97	MarketsCafe	Tue May 10 03:33:03 +0000 2022	Rainbow Children's Medicare listing expected t...	India	rainbow children medicar list expect mute rain...	[rainbow, children, medicar, list, expect, mut...	[rainbow children, children medicar, medicar l...	[rainbow children medicar, children medicar li...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
98	telugustop	Tue May 10 03:33:03 +0000 2022	Global Covid caseload tops 517.8 mn https://t....	India	global covid caseload top coronavirus covid ge...	[global, covid, caseload, top, coronavirus, co...	[global covid, covid caseload, caseload top, t...	[global covid caseload, covid caseload top, ca...	{'neg': 0.0, 'neu': 0.66, 'pos': 0.34, 'compou...
99	HealthSite4U	Tue May 10 03:33:02 +0000 2022	Delhi Reports Big Drop in Covid Cases On Monda...	Mumbai	delhi report big drop covid case monday covidc...	[delhi, report, big, drop, covid, case, monday...	[delhi report, report big, big drop, drop covi...	[delhi report big, report big drop, big drop c...	{'neg': 0.174, 'neu': 0.826, 'pos': 0.0, 'comp...

100 rows × 9 columns

Fig. 14 Hybrid Dictionary

	user	date	text	user_loc	clean_tweet	unigram	bigrams	trigram
0	gulf_news	Tue May 10 03:40:00 +0000 2022	#Pulitzer Prize: Slain photographer #DanishSid...	United Arab Emirates	pulitz prize slain photograph danishsiddiqui a...	[pulitz, prize, slain, photograph, danishsiddi...	[pulitz prize, prize slain, slain photograph, ...	[pulitz prize slain, prize slain photograph, s...
1	AhsanAbbasShah	Tue May 10 03:39:57 +0000 2022	Govt to build varsities in Diplo, Sakrand: @AA...	Lahore, Pakistan	govt build varsiti diplo sakrand	[govt, build, varsiti, diplo, sakrand]	[govt build, build varsiti, varsiti diplo, dip...	[govt build varsiti, build varsiti diplo, vars...
2	DailyhuntApp	Tue May 10 03:39:54 +0000 2022	'Apna chashma badal': Suniel Shetty to Twitter...	India	apna chashma badal suniel shetti twitter user ...	[apna, chashma, badal, suniel, shetti, twitter...	[apna chashma, chashma badal, badal suniel, su...	[apna chashma badal, chashma badal suniel, bad...
3	patro_anup	Tue May 10 03:39:53 +0000 2022	@rdmodisha @OdishaVigilance @MoSarkar5T @dmkor...	Jeypore Koraput Odisha	daylight loot govt revenu highlight	[daylight, loot, govt, revenu, highlight]	[daylight loot, loot govt, govt revenu, revenu...	[daylight loot govt, loot govt revenu, govt re...
4	DrPrashantkorat	Tue May 10 03:39:49 +0000 2022	Noted Kannada writer Bhairappa praises PM Modi...	Gandhinagar, Gujarat	note kannada writer bhairappa prais modi say o...	[note, kannada, writer, bhairappa, prais, modi...	[note kannada, kannada writer, writer bhairapp...	[note kannada writer, kannada writer bhairappa...
...	...	...	...	...	...	...	...	...
95	otvnews	Tue May 10 03:33:04 +0000 2022	#CycloneAsani: ECoR takes precautionary measur...	Bhubaneswar, India	cycloneasani ecor take precautionari measur fo...	[cycloneasani, ecor, take, precautionari, meas...	[cycloneasani ecor, ecor take, take precaution...	[cycloneasani ecor take, ecor take precautiona...
96	techfoogle	Tue May 10 03:33:03 +0000 2022	7 Latest 3D Printer Inventions 2022 YOU WISH Y...	India	latest printer invent you wish you had right n...	[latest, printer, invent, you, wish, you, had,...	[latest printer, printer invent, invent you, y...	[latest printer invent, printer invent you, in...
97	MarketsCafe	Tue May 10 03:33:03 +0000 2022	Rainbow Children's Medicare listing expected t...	India	rainbow children medicar list expect mute rain...	[rainbow, children, medicar, list, expect, mut...	[rainbow children, children medicar, medicar l...	[rainbow children medicar, children medicar li...
98	telugustop	Tue May 10 03:33:03 +0000 2022	Global Covid caseload tops 517.8 mn https://t....	India	global covid caseload top coronavirus covid ge...	[global, covid, caseload, top, coronavirus, co...	[global covid, covid caseload, caseload top, t...	[global covid caseload, covid caseload top, ca...
99	HealthSite4U	Tue May 10 03:33:02 +0000 2022	Delhi Reports Big Drop in Covid Cases On Monda...	Mumbai	delhi report big drop covid case monday covidc...	[delhi, report, big, drop, covid, case, monday...	[delhi report, report big, big drop, drop covi...	[delhi report big, report big drop, big drop c...

100 rows × 8 columns

Fig. 15 N-gram

	user	date	text	user_loc	clean_tweet	unigram	bigrams	trigram	Score	Class K-Nearest Neighbors	Class Liner SVM	Class Decision Tree	Class Random Forest	Class ExtraTreesClassifier
0	gulf_news	Tue May 10 03:40:00 +0000 2022	#Pulitzer Prize: Slain photographer #DanishSid...	United Arab Emirates	pulitz prize slain photograph danishsiddiqui a...	[pulitz, prize, slain, photograph, danishsiddi...	[pulitz prize, prize slain, slain photograph, ...	[pulitz prize slain, prize slain photograph, s...	{'neg': 0.0, 'neu': 0.413, 'pos': 0.587, 'comp...	Negative	Negative	Neutral	Negative	Negative
1	AhsanAbbasShah	Tue May 10 03:39:57 +0000 2022	Govt to build varsities in Diplo, Sakrand: @AA...	Lahore, Pakistan	govt build varsiti diplo sakrand	[govt, build, varsiti, diplo, sakrand]	[govt build, build varsiti, varsiti diplo, dip...	[govt build varsiti, build varsiti diplo, vars...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	Positive	Negative	Neutral	Negative	Neutral
2	DailyhuntApp	Tue May 10 03:39:54 +0000 2022	'Apna chashma badal': Suniel Shetty to Twitter...	India	apna chashma badal suniel shetti twitter user ...	[apna, chashma, badal, suniel, shetti, twitter...	[apna chashma, chashma badal, badal suniel, su...	[apna chashma badal, chashma badal suniel, bad...	{'neg': 0.181, 'neu': 0.819, 'pos': 0.0, 'comp...	Neutral	Negative	Negative	Negative	Neutral
3	patro_anup	Tue May 10 03:39:53 +0000 2022	@rdmodisha @OdishaVigilance @MoSarkar5T @dmkor...	Jeypore Koraput Odisha	daylight loot govt revenu highlight	[daylight, loot, govt, revenu, highlight]	[daylight loot, loot govt, govt revenu, revenu...	[daylight loot govt, loot govt revenu, govt re...	{'neg': 0.0, 'neu': 0.625, 'pos': 0.375, 'comp...	Negative	Negative	Neutral	Negative	Neutral
4	DrPrashantkorat	Tue May 10 03:39:49 +0000 2022	Noted Kannada writer Bhairappa praises PM Modi...	Gandhinagar, Gujarat	note kannada writer bhairappa prais modi say o...	[note, kannada, writer, bhairappa, prais, modi...	[note kannada, kannada writer, writer bhairapp...	[note kannada writer, kannada writer bhairappa...	{'neg': 0.294, 'neu': 0.706, 'pos': 0.0, 'comp...	Negative	Positive	Positive	Positive	Positive
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	otvnews	Tue May 10 03:33:04 +0000 2022	#CycloneAsani: ECoR takes precautionary measur...	Bhubaneswar, India	cycloneasani ecor take precautionari measur fo...	[cycloneasani, ecor, take, precautionari, meas...	[cycloneasani ecor, ecor take, take precaution...	[cycloneasani ecor take, ecor take precautiona...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	Negative	Negative	Neutral	Negative	Neutral
96	techfoogle	Tue May 10 03:33:03 +0000 2022	7 Latest 3D Printer Inventions 2022 YOU WISH Y...	India	latest printer invent you wish you had right n...	[latest, printer, invent, you, wish, you, had,...	[latest printer, printer invent, invent you, y...	[latest printer invent, printer invent you, in...	{'neg': 0.0, 'neu': 0.787, 'pos': 0.213, 'comp...	Negative	Negative	Negative	Negative	Negative
97	MarketsCafe	Tue May 10 03:33:03 +0000 2022	Rainbow Children's Medicare listing expected t...	India	rainbow children medicar list expect mute rain...	[rainbow, children, medicar, list, expect, mut...	[rainbow children, children medicar, medicar l...	[rainbow children medicar, children medicar li...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	Neutral	Positive	Neutral	Neutral	Neutral
98	telugustop	Tue May 10 03:33:03 +0000 2022	Global Covid caseload tops 517.8 mn https://t....	India	global covid caseload top coronavirus covid ge...	[global, covid, caseload, top, coronavirus, co...	[global covid, covid caseload, caseload top, t...	[global covid caseload, covid caseload top, ca...	{'neg': 0.0, 'neu': 0.66, 'pos': 0.34, 'compou...	Positive	Negative	Neutral	Negative	Neutral
99	HealthSite4U	Tue May 10 03:33:02 +0000 2022	Delhi Reports Big Drop in Covid Cases On Monda...	Mumbai	delhi report big drop covid case monday covidc...	[delhi, report, big, drop, covid, case, monday...	[delhi report, report big, big drop, drop covi...	[delhi report big, report big drop, big drop c...	{'neg': 0.174, 'neu': 0.826, 'pos': 0.0, 'comp...	Negative	Neutral	Neutral	Negative	Neutral

100 rows × 14 columns

Fig. 16 Final Result

Table I: Comparative Analysis

Model	Precision	Recall	F1-Score	Accuracy
KNN	72	69	70	71
SVM	95	93	94	94
Decision Tree	95	93	94	93
Random Forest	95	93	94	93
Extra Tree	95	94	94	94

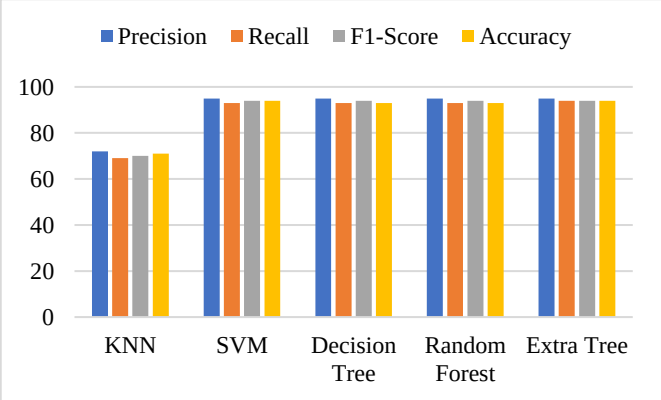

Fig. 17 Analysis Graph

V. Conclusion

Many machine learning algorithms, however, have proven to be effective in detecting fake news and messaging. Due to the ever-changing characteristics and aspects of fake news on social media networks, classifying false news is difficult. On the other hand, deep learning is well-known for its capacity to

calculate hierarchical features. Many research projects will use pattern recognition, text classification and model construction, knowledge discovery, observable recognition, and computer vision, as well as multidimensional and multi-task learning in the categorization of news posts, thanks to the recent deployment of computational intelligence research and applications. The proposed method includes a Dictionary-based feature as well as an additional tree classifier. To get the most out of tweeting news. The technique can be used for online media news, bogus news, and multi-classification in the future.

VI. References

1) T. Jiang, J. P. Li, A. U. Haq, A. Saboor, and A. Ali, "A Novel Stacking Approach for Accurate Detection of Fake News," IEEE Access, vol. 9, pp. 22626–22639, 2021, doi: 10.1109/ACCESS.2021.3056079.
2) K. A. Qureshi, R. A. S. Malick, M. Sabih, and H. Cherifi, "Complex Network and Source Inspired COVID-19 Fake News Classification on Twitter," IEEE Access, vol. 9, pp. 139636–139656, 2021, doi: 10.1109/ACCESS.2021.3119404.
3) S. Ni, J. Li, and H. Y. Kao, "MVAN: Multi-View Attention Networks for Fake News Detection on Social Media," IEEE Access, vol. 9, pp. 106907–106917, 2021, doi: 10.1109/ACCESS.2021.3100245.
4) P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding over Linguistic Features for Fake News Detection," IEEE Trans. Comput. Soc. Syst., vol. 8, no. 4, pp. 881–893, 2021, doi: 10.1109/TCSS.2021.3068519.
5) A. Uppal, V. Sachdeva, and S. Sharma, "Fake news detection using discourse segment structure analysis," Proc. Conflu. 2020 - 10th Int. Conf. Cloud Comput. Data Sci. Eng., pp. 751–756, 2020, doi: 10.1109/Confluence47617.2020.9058106.
6) P. Vlqjk et al., ") Dnh 1Hzv ' Hwhfwlrq D Frpsdulvrq Ehwzhhq Dydlodeoh ' Hhs / Hduqlqj Whfkqltxhv Lq Yhfwru Vsdfh," pp. 5–8.
7) Y. Yanagi, R. Orihara, Y. Sei, Y. Tahara, and A. Ohsuga, "Fake News Detection with Generated Comments for News Articles," INES 2020 - IEEE 24th Int. Conf. Intell. Eng. Syst. Proc., pp. 85–89, 2020, doi: 10.1109/INES49302.2020.9147195.
8) K. H. Kim and C. S. Jeong, "Fake News Detection System using Article Abstraction," JCSSE 2019 - 16th Int. Jt. Conf. Comput. Sci. Softw. Eng. Knowl. Evol. Towar. Singul. Man-Machine Intell., pp. 209–212, 2019, doi: 10.1109/JCSSE.2019.8864154.
9) S. I. Manzoor, J. Singla, and Nikita, "Fake news detection using machine learning approaches: A systematic review," Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019, no. Icoei, pp. 230–234, 2019, doi: 10.1109/ICOEI.2019.8862770.
10) S. Gaonkar, S. Itagi, R. Chalippatt, A. Gaonkar, S. Aswale, and P. Shetgaonkar, "Detection of Online Fake News : A Survey," Proc. - Int. Conf. Vis. Towar. Emerg. Trends Commun. Networking, ViTECoN 2019, pp. 1–6, 2019, doi: 10.1109/ViTECoN.2019.8899556.
11) Y. Yanagi, R. Orihara, Y. Sei, Y. Tahara, and A. Ohsuga, "Fake News Detection with Generated Comments for News Articles," INES 2020 - IEEE 24th Int. Conf. Intell. Eng. Syst. Proc., pp. 85–89, 2020, doi: 10.1109/INES49302.2020.9147195.
12) A. Kesarwani, S. S. Chauhan, and A. R. Nair, "Fake News Detection on Social Media using K-Nearest Neighbor Classifier," Proc. 2020 Int. Conf. Adv. Comput. Commun. Eng. ICACCE 2020, pp. 0–3, 2020, doi: 10.1109/ICACCE49060.2020.9154997.

13) H. Saleh, A. Alharbi, and S. H. Alsamhi, "OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection," IEEE Access, vol. 9, pp. 129471–129489, 2021, doi: 10.1109/ACCESS.2021.3112806.
14) N. Snell, W. Fleck, T. Traylor, and J. Straub, "Manually classified real and fake news articles," Proc. - 6th Annu. Conf. Comput. Sci. Comput. Intell. CSCI 2019, pp. 1405–1407, 2019, doi: 10.1109/CSCI49370.2019.00262.
15) T. Traylor, J. Straub, Gurmeet, and N. Snell, "Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator," Proc. - 13th IEEE Int. Conf. Semant. Comput. ICSC 2019, pp. 445–449, 2019, doi: 10.1109/ICOSC.2019.8665593.
16) Singh, Vernika, Raju Shanmugam, and Saatvik Awasthi. "Preventing Fake Accounts on Social Media Using Face Recognition Based on Convolutional Neural Network." Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020 57 (2021): 41.
17) Dwivedi, Sanjeev M., and Sunil B. Wankhade. "Survey on fake news detection techniques." In International Conference on Image Processing and Capsule Networks, pp. 342-348. Springer, Cham, 2020.

Cite this article as :

Reeya Baria, Sheshang Degadwala, Rocky Upadhyay, Dhairya Vyas, "Evaluation of Machine Learning Approaches for Classification of Fake News", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 3, pp. 30-44, May-June 2022. Available at doi : https://doi.org/10.32628/CSEIT228310
Journal URL : https://ijsrcseit.com/CSEIT228310