FAKE NEWS IDENTIFICATION TO CREATE TRUSTED MEDIA USING MACHINE LEARNING

S. THILAGAMANI and R. SHARUMATHI

[1]Professor and Head, [2]PG Scholar

Department of Computer Science and Engineering

M. Kumarasamy College of Engineering

Thalavapalyam, Karur, Tamil Nadu-639113

E-mail: sthilagamani11@gmail.com

  sharumathi0205@gmail.com

Abstract

Real records is very rich, and in state of-the-art world it's far very difficult to distinguish reality from falsehood in this age of technology improvement, spreading false records and portraying it as authentic become common. False facts is constantly dangerous and it is able to cause lot of problems in our society. In recent times most of the false statistics is unfold intentionally through social media it is used for one's selfishness and political benefit, which causes extremely good thread to our society. This creates a worry a number of the people and the use of this some humans meets their desires. It's miles tough for the character to discover the falsely spread messages which is created and unfold thru era. In this situation, it's miles necessary to observe the authenticity of the information to be had to us. Our goal is to analyse the authenticity of the news unfold on twitter account and to separate the fake message from the Twitter account and detecting the source of false records and preventing it from spreading similarly. It allows to create a accept as true with among net users to make sure the safety of social media.

1. Introduction

In the present contemporary nation, online networks and platforms are being very useful and with their beneficial nature helps the customers to examine and explore their opinion and share their views over certain inclusive of politics, schooling, and medical field. but, these systems are maybe used with a bad notations through positive models commonly for

economic advantage and in different instances for creating several two faced critiques, influencing the mindsets, and spreading preposterous and unreasonable news. Such phenomenon is generally referred as fake information [1].

Faux information makes the speciality of classifying the credibility of a tweet publish. It makes and gives a few rankings and their involvement and elucidated. On line information use their power to persuade tens of crores of human beings, however there aren't any ways to establish what it's far proper and what is false [2]. Research display that everyone has issues with identifying faux news, brushing off standards inclusive of age or schooling.

We have been facing a speedy boom in the unfold of fake information in the remaining past few years, these are largely prominent during the times of election. Such escalation of spreading articles on line which don't agree to information that has brought about several troubles not simply restrained to political affairs but overlaying numerous different domain names which includes sports, fitness, and additionally science [3]. One such vicinity suffering from fake information is the monetary environment, in which a lie could have caused devastated outcomes and can deliver the marketplace to a end. Our capacity to take a correct opinion depends more on the certain facts we devour; our view about world is shaped on the premise of statistics we take in consideration. It may be more proof that people have shown certain actions of preposteration to information which at the end seemed to be faux. One current scenario is the unfold of deadly covid, where faux information has reached several people through the net approximately the starting place and its severity. The situation changed as greater members read about the fake details spread in the social media. Figuring out such is a frightening venture.

Luckily, there are some processes that may be used to record the sure records as faux on the premise of their virtual recorded content. Maximum number of these strategies use truth websites which includes "PolitiFact" and "Snopes" [4]. There are some storage units that contains storage by way of experts that provides as with list of web sites which can be diagnosed as ambitious and fake. The problem about this kind of resources include person understanding which in turn is needed to maintain the articles as faux. greater important in this is, the truth validating the webs include summary

from precise node along with political issues and which are not concerned to discover fake news articles based on multiple domains together with enjoyment, games, and generation. The WWW contains some information or data about the huge set of codes with their needed documents or videos and sometimes audio. News that are posted internet in an undefined format (which includes information, movies, and videos) is highly hard and difficult to discover and categories them as these strictly requires customers understanding. But, computational methods which includes natural language processing (NLP) may be applied to stumble on anomalies that divides a textual content details this is defined in nature from the details that depends on records. The various techniques contain the analysis of propagation of faux information in contract with true information. More selectively, the method utilizes how a fake news details is propagated in a different way on a network relationship to a true article. The response that a piece of writing gets may be separated and detailed at a level of search and documentation to categorise the article as actual or fake. A more hybrid technique also can be used to investigate the social reaction of an editorial along with searching the textual functions to take a look at whether or not an editorial is devious in nature or no longer.

2. Machine Learning

Machine learning is a subset of artificial intelligence which uses software applications to often more appropriate outcome prediction without any programming needed to often them. In this kind of learning, we use past data as input which allows the machine to predict the output from these datas. [5] Artificial intelligence is a wider from of machine learning which in turn is a wider from deep learning. The most popular form of application of machine learning includes image recognition, speech recognition, fraud detection and threat detection and for some automation purposes.

Recently, the machine learning plays an important role in the many trends which offers many chances for customers behaviour and provides business requirements as well as supports the obtaining new products. Many of the present days big companies, such as Facebook, Google and Uber, uses this kind of learning in their products creation [6] [7]. This create huge competition between the companies and huge opportunity for customers.

Normally, machine learning signifies categorization of algorithm and allows it learn and provide more accuracy and making it more effective. The four types of machine learning includes, supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The data scientists select the methods based the type of data needed for them and their prediction level.

Supervised Machine Learning: This type of machine learning, data scientists uses already existing labeled data and trains these data with the necessary variables and also provides the output. This method helps to predict the output based on the history. This supervised machine learning includes regression and classification.

Unsupervised Machine Learning: This type of machine learning provides the data scientist with the set of unlabelled datasets or inputs. This method helps to find the hidden connection between input and output rather than the predicting the output. This includes clustering and association.

Semi Supervised Machine Learning: The blend of both supervised and unsupervised machine learning. In this type, data scientists use labelled data as input but doesn't depend on the output of past data which allows the machine to explore new outputs.

Reinforcement Machine Learning: In this type of machine learning, data scientist makes use of this method in a long process of input output prediction which includes multi step with pre defined rules and they accomplish this task by programming the algorithm considers whether it is positive or negative.

3. Related Work

Researchers of the topic an enhanced graph based semi supervised learning algorithm in their work tries to predict the fake news and its users from the past twitter datas. To carry out this, Researchers used Scrapy for obtaining the tweets and extracting the important features which includes, retweets, their average length, number of URLs and time gap between the tweets. They also mentioned that fake news creates more miscommunication and in convenience for the commoners to find the truth [8].

The work of researchers like Deepak and Bhadrachalam, they worked to overcome the fake news and identifying them using data mining. The methods they used in their work include FNN and LSTM. They mixed these algorithms with certain word vectors. The data mining part combines these and uses them to find the primary information. They provide us with prediction based on the calculation of Recall, precision, accuracy [9].

The authors of another work provide us with a two step method for fake news detection. The first step includes combining pre-processing methods with data sets and helps in converting it to unstructured format based on TF and Document term matrix.

The second step includes several supervised machine learning for the datas from the first step. The authors calculated prediction based on four methods like accuracy, precision, recall, and F1-measure. The e algorithm gives the best accuracy with 65.5% for the BuzzFeed Political News Data set [10].

4. Proposed System

In this proposed system, we are using some models of supervised machine learning. We used the subsequent studying algorithms along side this method to calculate and predict the performance of the algorithm in fake news detection.

Logistic Regression: Linear Regression comes under the supervised machine learning which helps to provide the mapping of input to the output data or information. As we are categorizing the textual content at the idea of a border characteristic set, with a output which can be binary (authentic/false or actual article/faux article), a logistic regression (LR) method is applied, because it offers the intuitive equation to separate the troubles into two or more than one instructions. We programmed hyperparameters stimulated to get the excellent result for all single datasets, whereas many parameters are examined before obtaining the most precision from this regression.

LR can be calculated like,

$$LR(X) = \frac{1}{(1 + e - (a0 + a1Y))} \tag{1}$$

The Equation 1 shows that logistic regression allows to access a sigmoid feature to convert the output to a proportional price; the aim is to limit the characteristic cost to gain a most excellent proportionality.

Support Vector Machine: Support vector machine(SVM) is one of supervised model which is used for binary class problem and is to be had in diverse set of capabilities. The goal of this model of version is to calculate a hyperplane (or choice limits) on the idea of feature group which helps to divide the record points. The measurement of hyperplane differs in keeping with a wide variety of features. As this would provide us with several chances for the hyperplane to develop into a N dimensional sector, this model helps to perceive a aircraft that divides the set of records factors of instructions with large margin. In this model, we used sigmoid SVM, polynomial SVM, Gaussian SVM, and basic linear SVM models. Mathematically, SVM is given as,

$$Svm(\theta) = \frac{1}{2} \sum_{i=1}^{j} \theta^2 \qquad (2)$$

In Equation 2, we have used linear kernels, in order to make it inseparable.

K Nearest Neighbours: KNN is an another model in unsupervised machine learning model where an established variable is not required to expect the result on a certain information. We offer sufficient education records to this system and allows them to check which specific neihbourhood a facts that belongs to a certain point. KNN version calculates the distance of a each records factor to its closest buddies, and the cost of k calculates most of the people of its associates 'votes; if the cost of k is 1, then the brand recent facts point is alloted to a category which has the closest distance.

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{k} (x_i, y_j)2} \qquad (2)$$

The Equation 3 shows that distance can be measured using Euclidean distance formula.

Decision Tree: Gaining knowledge of by means of decision tree develops a techniques based on the significance of a decision tree as it is a estimative version. Decision tree method has an application in mining especially

statistics and machine learning. In this decision tree structures, the leaves shows the goal variable values and the branches lists the values of combinations of certain enter variables. In mastering and information mining, a selection tree shows the records however no longer the choices of them, the tree could be used as a starting point for the decision tree technique.

Random Forest: The Another method for detection fake news in machine learning is Random Forest algorithm which is occasionally considered and converted as a forest of decision trees is a method in which set of condition are there that reduces the difference in various forecasts that are formed in the decision tree. This helps thereby to improve their performance. More decision trees are combined in form of bagging in order to achieve higher performance. In its maximum traditional method, it plays sequential learning on a couple of choice tree randomly built and educated on one of a kind of facts. The correct ser of number of trees are necessary as they act as step to move on into few hundred or thousand, which plays an important role.

Naive Bayes: Bayes theorem contains many category one of which is easy probabilistic method and it is based on the idea of independent rules and conditions. The bayes theorem are executed using a classifier which is mainly a linear classifier. Naive Bayes Classifier is a famous supervised machine learning method. Its miles a Supervised getting to know set of rules used for class. It's far specially beneficial for textual content categorization issues.

Gradient Boosting: Gradient boosting is a one of supervised learning for regression and class issues. It helps to create a model of prediction within the shape of a set of weak prediction models, normally selection of tree. Here, the steps are carried out in levels for alternative bagging methods. This method specifies these by allowing more accurate option of an true difference loss in function.

Multilayer Perceptron: The multilayer MLP perceptron comes under artificial neural network is divided into a few basic categories input layer in output field only; therefore which is a direct broadcast net (supplier). Every single level of this is made of neurons called as, endothelial neurons (called outgoing) in being global system results. It is a type of supervised learning algorithm studying job by training in databases. Used for Tabular Data sets

(containing data in column format as table website) and Planning, Depression, Prediction Problems.

As said in beginning, for this research there is various machine algorithms that are Logistic Regression: LR, Decision Tree: DT, Random Forest: RF, Support Vector Machines: SVM and the neural network Multi Linear Perceptron: MLP, Naive Bayes: NB, Gradient Boosted Tree: GBT.

Our only goal is selecting the model that best suits (giving the good prediction results) will be used to detect faux tweets in real time where, - TP: True Positive: Accurately predicted tweets that are real - FP: Inaccurately predicted tweets that are real - FN: False Negative: Inaccurately predicted tweets that are fake - TN: Incorrect Fact: Accurately predicted tweets that are Fake.

Table 1. Combination Matrix.

Classification into true or false	True after Prediction	False after Prediction
Original true	True Positive(T+)	False Negative(F-)
Original false	False Positive(F+)	True Negative(T-)

Table 1 represents the combination matrix includes True positive, True negative, False positive, False negative. Based on the combination matrix, we can calculate,

Accuracy: Accuracy means sum of all part to the sum of all tweets.

$$\text{Accuracy} = \frac{(T+) + (T-)}{(T+) + (T-) + (F+) + (F-)} \tag{4}$$

The Equation 4 represents the calculation of Accuracy.

Recall: also called sensitivity or a good level of truth. Recall is exact true and positive to all forms of true. This method shows how true a model is.

The Equation 5 represents the calculation of Recall.

$$\text{Recall} = \frac{(T+)}{(T+) + (T-)} \tag{5}$$

Precision: Accurate metrics indicate the precision of the REAL class.

Precision is the value of only true positive to all positive.

$$\text{Precision} = \frac{(T+)}{(T+) + (F+)} \tag{6}$$

The Equation 6 represents the precision calculation.

F1 Score: It shows the relationship between the precision and recall and provides the accuracy.

$$F1\ Score = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

The Equation 7 represents the calculation of F1 score using precision and recall.

5. Research Design

Research design includes elaboration of datasets, pre processing of data, feature extraction, training the classifier and tweets classification. The datasets module includes gathering of data and data selection. After the data gathering, raw data are converted into useful data by the pre processing method. Then, the classifier is trained. Finally, qualitative analysis for this model is evaluated. The below figure 1 shows the architectural design:
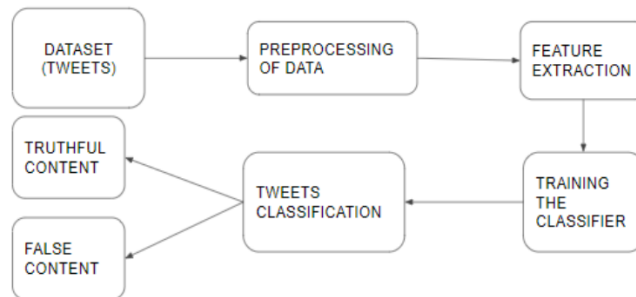

Figure 1. Architectural Design.

Data Selection: In this study, we are using the old Twitter database. The website contains tweets history. The tweets are collected by researchers. Researchers have a sample of online election datasets based on how voters have created to the parties and election of that time and used a sensible snowball sample method to collect relevant hashtags. The hashtag tagged # is used to target specific topic available on Twitter. It allows people to easily

follow topics according to their interests. This type of sampling method is carried on so as to add the additional relevant data to the original datasets present. Generally, already available original datasets helps to access or find the similar datasets needed for further study. The datasets are collected on specific conditions which includes hashtag terms, users comment, real time information for that hashtags and medias influence. The text removed some existing tags from the embedded tweets, which were assigned to compliance. When a marker exceeds a specific limit, it is added to the tag list and which in turn will help to add the new tag list to them.

Datasets: Initial datasets which we used here are available in online and that are free of charge. The datasets includes the set of true and false data which gathered from different domains. The true set of data in the datasets include mostly the real world events and the false data from datasets include the unsubstantiated lie. The compatibility of faux news from the political base of more of these topics can be checked by humans by referring them to various other websites and by comparing them. We have collected and used different datasets for this research to give a detailed representation.

The first data set is referred as the "ISOT Fake News Dataset" (which can be called as dataset D1) containing both true and false articles published on the WWW. The factual articles published on the popular reuters.com news website, while false articles are published from many sources, especially websites that have been flagged from different websites [14]. The database are formed of a total of 45,000 articles, 21,000 are true texts and 24,000 are false topics. The content of the companies contains various articles on the domains, but with a strong focus on political issues.

The second set of data is available at Kaggle (which are called as D2). It contains of a total of 21,000 articles that are trained and only 5,000 documents is selected for testing. Database is made up of many online resources [11]. Normally, political topic can't be restrained to single websites as their opinion may differ from website to website which have both true and false data. The topics are not limited to one domain as political as include both false and true articles from various sources domains.

The next set of data is taken from Kaggle (hereafter is called as D3); which includes total number of 3400 articles which can true and false. The authentic information are taken from several journals and publications like

New York times [12], while the headlines are false published on unreliable news websites. The most of its domains which includes games, health, and political affairs.

Combination of both data set is a group of data from various data sets this mixed dataset is referred as D4 [13]. This dataset is different from other dataset as it is not original and derived from several sets to cover multiple domains in one database to improve the accuracy and efficiency [15].

Training the Classifier: It is necessary to carry certain performance in order to find the quality content analysis on specific datasets. To do these tasks, many ML techniques are done in this phase. ML is one of the AI model where the person find the input and its output without any assistance with use of past data itself. These methods help the computer to study the output without the human intervention.

Data sampling: The segment is designed to identify potentially fraudulent messages on Twitter. The separator is built using machine-readable format. The role of Supervised machine learning method is to provide the input and train them based on the already provided past output. The dataset is divided into two training set and test sets. The training set train the model and the test set check, whether the data is learned by the machine. In this set of data few tweets are taken into consideration and then they are selected as training data. Two different classes were used to record data. The beginning stage is taken as true tweets. These tweets which are good messages sent to people. These are marked '1'. Another category is false tweets. These tweets are sent by bots or are incorrect and creates misinformation. These tweets are marked '0'.

Sample tweets were chosen by researchers in unorderly manner for the purpose of training and they are labelled. First, few true tweets were collected. This database made in the study to recruit other voters on Twitter. Database was refined for their study The 300 collected and labeled for messaging used as a training- as well as a test set of supervised machine learning algorithms.

Preliminary processing of data: The data collected from the past datasets are always raw data which needs to be cleaned and extracted before it is used for the any further process of machine learning methods. These raw

data can't be used straightly into the machine. Many technique require numbers such as input. Therefore, many techniques require vectors of the details of numbers that do not change to green text with variable lengths, such as tweets.

Before any further step, data has to be cleaned and extracted properly to improve the efficiency of the any chosen machine learning methods. In every comment, before next step, punctuation mark has to be cleared to decrease the size of the comment. Next special characters are also cleared for being more efficient. Then all the letters converted into lowercase.

The Scikit-lern package allows with predata processing usage. We have used Name Bag Representation Model. Provides tools for token entry, calculation, generalization and data exposure. The model discards all order details in messages, but checks the appearance of characters in the text. Every different name is allotted to a special number, a symbol. It is not necessary to remove the shortcuts. This form of word may appear more times which means nothing to the coded vectors. However, as this approach remarks the emergence of characters in the text, which has been considered.

We may list character frequencies, another way for the use pf a standard word bag. The TfidfVectorizer function is used. The work is based on „frequent off-text texts, abbreviated TFIDF. It is widely used in character analysis with a greater character chorus. The value of number of floating points has to be measured for the every characters present in the corpus.

The calculation for vectorization is says:

$$tf - idf(t, d) = tf(t, d) \times idf(t) \tag{8}$$

Based on equation 8, the word frequency means the sum of the total occurrence of a character of the particular database, is repeated in this section. The total number of messages is considered as Nd. The sum of the messages from particular name is considered as $df(d, t)$.

Machine Training the Classifier: In the study, various different machine learning techniques were analysed. Based on these techniques datasets are tested and trained using available datasets, their 80:20 scale used in response training. to check. 10k certification was made to verify the validation of the results in the test. The separator is configured using the

Python. We are using The Python Scikit-learn package. It has some practical as well as tool for various machine learning problems. It is used for data mining. This package is selected for various reasons because it is easy for naïve bayes problem, for managing the datasets and using them for more machine learning problems and it is also used to differentiate the tests. Its purpose in this research is more and they are mostly open source, the package can be re selected and re used for as many as times it is needed.

After these processes, data can be categorized into either it is true or false. True means information is true one and false refers the false dataset. The true data are mostly sent by humans and they may be indicated as 1, whereas the false seems to be bot loaded which is wrong and they are indicated as 0.

Every technique is tested and analysed based on a 10-step verification model. The most efficient technique may be accessed to divide the data. The functionality of every ML techniques are discovered.

6. Result and Discussion

The algorithm detected potentially false message. Deliberately posted content or messages which created by users who are bots. Tweets after find out by whom it was created, the quality content analysis carried out to find the connection between them. After this phase, we have to separate the false tweets and group them into different category.

Results of Classification: Performance measures were used for separate testing algorithms. In order to select which technique is suitable for this kind of model below shows the probability of different ML techniques.

Table 2. determine false and true percentage in precision(P), recall(R), F1 score(F).

	False(0)			True(1)		
	P	R	F	P	R	F
LSVM	0.95	0.75	0.84	0.82	0.97	0.89
B-NB	0.66	0.89	0.76	0.86	0.59	0.7
M-NB	0.79	0.82	0.81	0.84	0.81	0.83

G-NB	0.88	0.79	0.83	0.83	0.91	0.87
DT	0.86	0.89	0.88	0.9	0.88	0.89
RF	0.69	0.86	0.76	0.84	0.66	0.74
ET	0.77	0.82	0.79	0.83	0.78	0.81
SGD	0.63	0.86	0.73	0.82	0.56	0.67

The Table 2 explains about the determination of false and true percentages based on F-Score which is harmonic accuracy measurement and memory level. Accuracy is a measure of true points in all points, while memory is a measure of true points in all well-organized messages. Three algorithms get maximum F1- Score. Based on the various findings, it is clear that decision tree provides us with maximum accuracy of F1 score that is 87%. And the second best model is considered as Bayes model which provides us with 84%. And Random forest shows very poor accuracy and precision which was really unexpected. The algorithm received a rating with a fact of F1 Score of 74% and was deleted by the bayes theorem. In random tree model each node is built with sample. During parting a tree in development, partition what will be selected remains no longer the better division in every aspects. As a result of this inevitability, forest bias tends to rise slightly, which is often the case compensation due to scale. In our database, this has been the case may not adequately overcome.

The Combination shows both the tree model for detecting the fake information. They have a small faux positive. But the naïve bayes seems to more false fake ratings. 45% of actual tweets are considered as faux method. As a result, Bernoulli has a very low price with an F-Score weight of 73%. In finding the real messages, SVM has more positive result with less false information. But gaussian and naïve bayes and decision tree show next better performance and efficiency which represents little false number. Linear SVM had 3% lie level, while G-NB has a 9% rating and Decision Tree has 12.5%. During you compare these past algorithms, it can be seem. During this, we can find that decision tree is best in both the models. It has two highs F-Score, by comparison, seems to be the worst in the decision any false messages. The DT algorithm is much better distinguishing both true and false messages, which is why FScore is so high.

Results for quality content analysis: Mathematical Features Since there is no metadata available in the database, we have tried to do so remove

those facts from text model by measuring the following:

- Part of tweets that are resent (these tweets begins with "RT")

- The average value of each tweets

- Average number of URLs mentioned in every tweets.

- Average number of hashtags mentioned per tweet
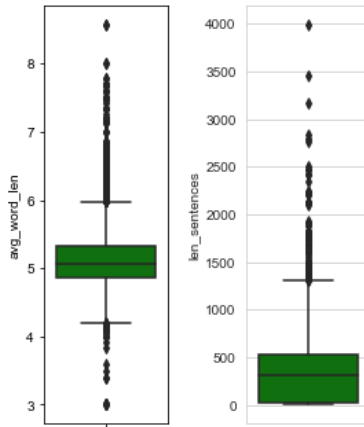
- Medium length of every tweets based on Figure 2

Figure 2. word length of tweets.

The Decision Tree algorithm is designed to detect lies messages. With false quality content analysis tweets posted in the past, different characters and elements of faux model were selected and divided into six categories. The effects of the division may be analyzed in here. These are shown in figure 3 and 4.

Total of about, 600,000 tweets were categorized by ML technique. Among these 325,000 considered as true and other 225,000 was considered as false. Quality content analysis was carried on the false tweets received. They are scanned for different features and features. The purpose for which the messages were sent was investigated. Different symbols were obtained and assembled to develop into six sections.

The six categories are:

· Statistical messages

- Marketing messages

- Linking messages

- Hashtag messages

- Resend

- Negative messages

In the following, every categories was evaluated.

Statistical messages: Sarcastic messages comes under this category of messages. Messages under these are considered as false but they are developed by peron who sent them. These kind of messages doesn't bring more or direct influence on the internet. These tweets are posted in net to please the recipient. Satirical Twitter accounts available on the site posing as celebrities or acting as a news channel.

Marketing messages: The next category of messages are marketing messages. This kind of tweets are used by certain people to show their model to others through internet and influence the people by using the model they developed. It can be carried out by seeing the hashtags mentioned and slogans used in the messages which refers to the product sales, people have been encouraged to view corporate platforms where goods or their services can be used to purchase. But these messages doesn't have any connection with the products selling and their companies. The hashtags of parties election articles are carried out for the spreading of ad, not for sharing stories, facts or opinions about political affairs.

	label	statement	subjects	speaker	speakersJobTitle	state_info	party_afflin	bt_c	f_c	ht_c	mt_c	pof_c	context
0	false	wisconsin is on pace to double the number of l...	jobs	katrina-shankland	state representative	wisconsin	democrat	2	1	0	0	0	a news conference
1	false	says john mccain has done nothing to help the ...	military,veterans,voting-record	donald-trump	president-elect	new york	republican	63	114	51	37	61	comments on ABC's This Week.
2	half-true	suzanne bonamici supports a plan that will cut...	medicare,message-machine-2012,campaign-adverti...	rob-cornilles	consultant	oregon	republican	1	1	3	1	1	a radio show
3	pants-fire	when asked by a reporter whether hes at the ce...	campaign-finance,legal-issues,campaign-adverti...	state-democratic-party-wisconsin	NaN	wisconsin	democrat	5	7	2	2	7	a web video
4	true	over the past five years the federal governmen...	federal-budget,pensions,retirement	brendan-doherty	NaN	rhode island	republican	1	2	1	1	0	a campaign website

	title	text	date	country	label	URL	avg_word_len	len_sentences	source
0	A video shows a fortune teller predicting the coronavirus pandemic in December on Spanish TV.	Circulating on social networks a video that shows an excerpt from a Spanish television show, supposedly issued December 24, 2019, in which it appears a woman (who claims to be psychic) ??to make "predictions". In this video, the woman describes a set of events that have been interpreted as a detailed forecast of Covid-19 pandemic that has hit the world. It is, however, a fake video, at least as regards the date of issue. The video has been being disseminated on the Internet with a date and not tampered with the real.	08-04-2020	Portugal	False	https://observador.pt/factchecks/fact-check-uma-vidente-previu-a-pandemia-da-covid-19-na-televisao-espanhola-em-dezembro/	4.795181	83	https://observador.pt

Figure 3. Classification of tweets and feature extraction.

Linking messages: The data set contained links to many interesting tweets. Tweets contain a # with a URL. Usually, these are different messages are delivered by different persons. Influencing account usually sends same messages several times to influence them in very little time. Person that posted all Twitter messages on the website are made available in internet for at least 3 days before they were removed. Based on the Twitter guidelines, sending URL in a without explanation is considered illegal. Many URLs to the messages have been removed. Links to these many tweets available on our website refer to online platforms that developed on every domain. These nodes may be used because these nodes are therefore no payment information are required. Therefore, they are difficult to track the first recipient of these messages.

Hashtag messages: Many messages that are spread has hashtags# to show the part of tweets and show its main theme. That are to send the types of suspended messages. The messages are used for increase brand awareness and political party awareness.

Re-tweaking: Bots are used for reposting posts for ordinary person. The messages bots are used for many purposes. The maximum number of times that messages spread shows us that whether the information or messages true. If a tweets has a huge number of messages that are resent, it seems that most people agree with the tweet message. It will provide the sender by showing that must look knowledgeable about the title they speak. Unluckily, resend messages are difficult to identify with a ML technique, as well as they have no connection with the original message. Using only the details of the resend messages, we can't decide whether they are human or bot. Additional details about the account reposted message.

```
Epoch 2/10
1139/1139 [==============================] - 19s 17ms/sample - loss: 0.4552 - accuracy: 0.8323 - val_loss: 0.3775 - val_accurac
y: 0.8740
Epoch 3/10
1139/1139 [==============================] - 24s 21ms/sample - loss: 0.4516 - accuracy: 0.8323 - val_loss: 0.3745 - val_accurac
y: 0.8740
Epoch 4/10
1139/1139 [==============================] - 20s 17ms/sample - loss: 0.4459 - accuracy: 0.8323 - val_loss: 0.3784 - val_accurac
y: 0.8740
Epoch 5/10
1139/1139 [==============================] - 19s 17ms/sample - loss: 0.4439 - accuracy: 0.8323 - val_loss: 0.3674 - val_accurac
y: 0.8740
Epoch 6/10
1139/1139 [==============================] - 19s 16ms/sample - loss: 0.4389 - accuracy: 0.8323 - val_loss: 0.3940 - val_accurac
y: 0.8740
Epoch 7/10
1139/1139 [==============================] - 19s 16ms/sample - loss: 0.4351 - accuracy: 0.8323 - val_loss: 0.3764 - val_accurac
y: 0.8740
Epoch 8/10
1139/1139 [==============================] - 19s 16ms/sample - loss: 0.4295 - accuracy: 0.8323 - val_loss: 0.3697 - val_accurac
y: 0.8740
Epoch 9/10
1139/1139 [==============================] - 19s 16ms/sample - loss: 0.4317 - accuracy: 0.8323 - val_loss: 0.3713 - val_accurac
y: 0.8740
```

Figure 4. Accuracy percentage.

Negative messages: Negative messages are developed in order to influence the people not to believe any particular unit. This is used when any party needs to accuse any other party and create bad opinion on them. The recipient will try to influence the normal people which does not allow them to cast vote for someone or a party during the election, or to hurt the person. These are often use touching, unreliable, or specially crafted articles for enhance reading, internet sharing, and online financial needs.

7. Conclusion

In this study, we learned how false messages were used on Twitter during the 2012. The filter was upgraded to find the active fake articles. The Twitter messages of 2012 has been listed and are split based on our demand. These split tweets was done based on quality content analysis. The conclusion part includes, the results for the work which are discussed, as the research has its own shortcomings and other potential activities.

The eight different machine learning algorithms are analyzed and compare and contrast for these are done. These supervised learning methods are trained and tested for a sample of tweets data of the 2012 elections. The sample data contains a same number of 'true' and 'false' tweets. With these samples most of the machine learning methods are trained and tested. F1 score is used to calculate the precision and performance of the methods.

Decision tree algorithm is considered as the best out of all the methods of machine learning. Tweets which are referred as false are given more importance and analysed carefully. These false tweets can be classified and

categorized into six different tweets. The results concludes that decision tree remains as the best method of prediction in this work, which can easily categorize the difference as well as true and false tweets. The maximum percentage of precision or accuracy created using F score seems to 89%. The algorithm worked very well with the F-Score weighing 88%. The second best seems to be the support vector machine which provides with the maximum precision of about 85%. In F score percentage. From this, it assures the SVM remains best in the case which provides with more computer aided input as well as depends on the length of the input provided. So, this work leans more towards the decision tree as provides access to the large datasets and little independent inputs.

Based on the above statements, the decision tree seems to separate the tweets and analyze them on certain criteria and divide them into either true or false. We took into consideration of about 70,000 tweets in total. They was separated into two, i.e., true or false. From these, 400,000 was divided into true and 300,000 are grouped into false. After grouping, the messages needs to be analyzed. The messages cne of six types which are statistical, connect, poor polling, hashtag and marketing messages. Based on the research, false are more which seems to affect the quality factor every tweets and mainly false information are to be labelled separately.

Certain limitations are calculated in this research, which can be eventually addressed in the near future. The dataset are made access from the several datasets available for the research. These datasets are mostly based from the past tweeter account. Certain tweets which are considered as false are old tweets so, these tweets are not available now. Hence they are considered as false. Most accounts are been extinct this shows that information about the tweets or account, such as Twitter followers and the number of messages sent, could not be retrieved. This could be good to find or do research about the fake news or message detection. As studied by, potentially fraudulent messages can be detected using a large amount of account data. An algorithm trained with a combination of tweet content and account data can have high potential.

Due to limited user account information, it has been difficult to do a lot of training machine learning machine training. A comprehensive data set can analyse the probability of the Machine Learning algorithm and thus reduce the number of categories that are not true.

References

[1] Wikipedia, https://en.wikipedia.org/wiki/COV ID-19_ p{and}emic_in_Morocco.

[2] A. Onan, Sarcasm identification on Twitter: A machine learning approach, In: R. Silhavy, R. Senkerik, Z. Kominkova Oplatkova, Z. Prokopova, P. Silhavy, editors, Artificial intelligence trends in intelligent systems, CSOC 2017, Advances in intelligent systems and computing, Cham: Springer; Vol. 573 2017, http://dx. doi.org/10.1007/978-3-319-57261-1_37.

[3] A. Onan, Topic-enriched word embeddings for sarcasm identification, In: Silhavy R, editor, Software engineering methods in intelligent algorithms, CSOC 2019, Advances in intelligent systems and computing, Vol. 984, Cham: Springer; 2019, http://dx.doi.org/10.1007/ 978-3- 030-19807-7_29.

[4] A. Onan, Sentiment analysis on Twitter based on ensemble of psychological and linguistic feature sets, Balkan J. Electr. Comput. Eng. 6(2) (2018), 69-77.

[5] A. Onan, Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks, Concurr Comput Pract Exp 2020;e5909. http://dx.doi.org/10.1002/cpe.5909.

[6] A. Onan, Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach, Comput. Appl. Eng. Educ 2020 1-18. http://dx.doi.org/10.1002/cae.22253.

[7] Aldwairi Monther and Alwahedi Ali, Detecting fake news in social media networks, In The 9th international conference on emerging ubiquitous systems and pervasive networks (EUSPN 2018).

[8] D. Zhuk, A. Tretiakov, A. Gordeichuk and A. Puchkovskaia, Methods to identify fake news in social media using artificial intelligence technologies, In: D. Alexandrov, A. Bukhanovsky, A. Chugunov, Y. Kabanov, O. Koltsova, editors, Digital transformation and global society, DTGS 2018, Communications in computer and information science, Cham: Springer; (858) 2018.

[9] Atodiresei Costel-Sergiu, Tănăselea Alexandru, Iftene Adrian, Identifying fake news and fake users on twitter, In International conference on knowledge-based and intelligent information and engineering systems, KES2018, 3-5 2018, Belgrade, Serbia.

[10] S. Deepak and Chitturi Bhadrachalam, Deep neural approach to fake-news identification, In: International conference on computational intelligence and data science (ICCIDS 2019), Procedia computer science 167 (2020), 2236-2243.

[11] Nyow Ning Xin and Chua Hui Na, Detecting fake news with tweets properties, In IEEE Conference on the Application, Information and Network Security (AINS), (2019).

[12] Shu Kai, Mahudeswaran Deepak, Suhang Wang, Lee Dongwon and Liu Huan, Fakenewsnet: A data repository with news content, social context and spatiotemporal information for studying fake news on social media, 2019, arXiv:1809.01286v3 [cs.SI].

[13] M. BalaAnand, N. Karthikeyan, S. Karthik, et al. An enhanced graph-based semisupervised learning algorithm to detect fake users on Twitter, J. Supercomput 2019 75 6085-6105. http://dx.doi.org/10.1007/s11227-019-02948-w.

[14] Zhang Chaowei, Gupta Ashish, Kauten Christian, V. Deokar Amit and Qin Xiao, Detecting fake news for reducing misinformation risks using analytics approaches, European J. Oper. Res. 279(3) (2019), 1036-52. http://dx.doi.org/10.1016/j.ejor. 2019.06.022.

[15] Stahl Kelly, Fake news detection in social media, B. S. Candidate, Department of Mathematics and Department of Computer Sciences, California State University Stanislaus, University Circle, Turlock, CA 95382.