GMR Institute of Technology
An Autonomous Institute Affiliated to JNTUK, Kakinada

GMRIT
Training Tomorrow's Engineers Today

FAKE NEWS DETECTION USING MACHINE LEARNING ALGORITHMS

A Term paper report submitted in partial fulfillment of the requirement for the Award of degree

BACHELOR OF TECHNOLOGY IN
COMPUTER SCIENCE AND ENGINEERING

Submitted
By

P. SAI KRISHNA
19341A05C3

Under the esteemed guidance of

Dr. Balaji Maram
Associate Professor
Dept. of Computer Science & Engineering

Department of Computer Science and Engineering
GMR INSTITUTE OF TECHNOLOGY
(An Autonomous institute, affiliated to J.N.T.University kakinada)
NAAC "A" Graded, NBA Accredited, ISO 9001:2008 Certified Institution
G.M.R. Nagar, Rajam-532127, A.P

2021-22

INDEX

GMR Institute of Technology
An Autonomous Institute Affiliated to JNTUK, Kakinada

GMRIT
Training Tomorrow's Engineers Today

GMR INSTITUTE OF TECHNOLOGY

(An Autonomous institute, affiliated to J.N.T. University Kakinada)

NAAC "A" Graded, NBA Accredited, ISO 9001:2008 Certified Institution G.M.R. Nagar, Rajam-532127, A.P

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that term paper report titled **"FAKE NEWS DETECTION USING MACHINE LEARNING ALGORITHMS"** *submitted by* **P. Sai krishna** *bearing* **Reg. No:19341A05C3** *has been carried out in partial fulfillment for the award of* **B.Tech** *degree in the discipline of* ***Computer Science & Engineering*** *to* **JNTUK** *is a record of bona fide work carried out under our guidance and supervision.*

The report embodied in this paper has not submitted to any other university or institution for the award of any degree or diploma.

Signature of the Supervisor

Dr M. Balaji
Associate Professor
Department of CSE
GMRIT, Rajam.

Signature of the H.O.D

Dr. A V Ramana
Professor and Head
Department of CSE
GMRIT, Rajam.

ACKNOWLEDGEMENT

It gives me an immense pleasure to express deep sense of gratitude to my guide, **Mrs. N. Lakshmi Devi** **Assistant Professor**, Department of Computer Science and Engineering of whole hearted and invaluable guidance throughout the report. Without his/her sustained and sincere effort, this report would not have taken this shape. He/she encouraged and helped me to overcome various difficulties that I have faced at various stages of my report.

I would like to sincerely thank **Dr. A V Ramana,** Professor & HOD, Department of Computer Science and Engineering, for providing all the necessary facilities that led to the successful completion of my report.

I take privilege to thank our Principal **Dr. C.L.V.R.S.V.Prasad**, who has made the atmosphere so easy to work. I shall always be indebted to them.

I would like to thank all the faculty members of the Department of Computer Science and Engineering for their direct or indirect support and also all the lab technicians for their valuable suggestions and providing excellent opportunities in completion of this report.

P. SAI KRISHNA

19341A05C3

ABSTRACT

Automatic detection of fake news, which could negatively affect individuals and the society, is an emerging research area attracting global attention. The problem has been approached in this study from Natural Language Processing and Machine Learning perspectives. The evaluation is carried out for three standard datasets with a novel set of features extracted from the headlines and the contents. Social media interaction especially the news spreading around the network is a great source of information nowadays. From one's perspective, its negligible exertion, straightforward access, and quick dispersing of information that lead people to look out and eat up news from internet-based life. Twitter being a standout amongst the most well-known ongoing news sources additionally ends up a standout amongst the most dominant news radiating mediums. It is known to cause extensive harm by spreading bits of gossip previously. This study proposes a model for recognizing forged news messages from twitter posts, by figuring out how to anticipate precision appraisals, in view of computerizing forged news identification in Twitter datasets. Support Vector Machine, Naïve Bayes Method, Logistic Regression and Recurrent Neural Network models are the algorithms used in this study. The easy access and exponential growth of the information available on social media networks has made it intricate to distinguish between false and true information.

Keywords: Fake News, Twitter, Social Media, Data quality, Deep Learning, Machine Learning.

1. INTRODUCTION

1.1 Fake news detection using Machine Learning

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such **as** image recognition**,** speech recognition**,** email filtering**,** Facebook auto-tagging**,** recommender system**,** and many more. Machine learning is an application of AI which provides the ability to system to learn things without being explicitly programmed. Machine learning works on data and it will learn through some data. 'Supervised learning' is one of the type of machine learning. Supervised learning means we trained our model with labeled examples so the machine first learns from those examples and then performs the task on unseen data. This machine learning tutorial gives you an introduction to machine learning along with the wide range of machine learning techniques such as Supervised**,** Unsupervised, and Reinforcement learning.

Fake news is false or misleading information presented as news It often has the aim of damaging the reputation of a person or entity, or making money through advertising revenue. However, the term does not have a fixed definition, and has been applied more broadly to include any type of false information, including unintentional and unconscious mechanisms, and also by high-profile individuals to apply to any news unfavourable to their personal perspectives.

Fake news's simple meaning is to incorporate information that leads people to the wrong path. Nowadays fake news spreading like water and people share this information without verifying it. This is often done to further or impose certain ideas and is often achieved with political agendas. For media outlets, the ability to attract viewers to their websites is necessary to generate online advertising revenue. So it is necessary to detect fake news. This fake news can be detected by using machine learning algorithms.

2. LITERATURE SURVEY

[1] Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita, "Fake News Detection Using Machine Learning approaches: A systematic Review", IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8 ©2019 IEEE.

The easy access and exponential growth of the information available on social media networks has made it intricate to distinguish between false and true information. The easy dissemination of information by way of sharing has added to exponential growth of its falsification. Machine learning has played a vital role in classification of the information although with some limitations. This paper reviews various Machine learning approaches in detection of fake and fabricated news. In this paper author discussed various fake news detection methods like Linguistics basis Deception modelling, Clustering, Predictive modelling, Content cue-based methods, non-text cue-based methods. The author used the following algorithms on data sets such as Naïve Bayes, Decision trees, SVM, Neural Networks, Random Forest, XG Boost.

The various types of fake news by Authors of paper, in their recent paper is summarize below.

1. Visual-based: These fake news posts use graphics a lot more in as content, which may include morphed images, doctored video, or combination of both.

2. User-based: This type of fabricated news is generated by fake accounts and is targeted to specific audience which may represent certain age groups, gender, culture, political affiliations.

3. Knowledge-based: These types posts give scientific (so called) explanation to the some unresolved issues and make users to believe it is authentic.

4. Style-based: posts are written by pseudo journalists who pretend and copy style of some accredited journalists

5. Stance-based: It actually is representation of truthful statements in such a way which changes its meaning and purpose.

Although there is evident success in detection of fake news and posts using various Machine learning approaches. However everchanging characteristics and features of fake news in social media networks is posing a challenge in categorization of fake news. However the main characteristic feature of deep learning is to compute hierarchical features.

[2] Adrian M. P. Bra¸soveanu, Razvan Andonie, Semantic Fake News Detection: A Machine Learning Perspective", c Springer Nature Switzerland AG 2019 I. Rojas et al. (Eds.): IWANN 2019, LNCS 11506, pp. 656–667, 2019.

Teaching an automated system to recognize fake news is a challenging task, especially due to its interdisciplinary nature. In this paper author approach for semantic information extraction and then describe how we use the extracted information to classify fake news. We present techniques related to: metadata collection, extraction of relations, and inclusion of embeddings to neural classifiers.

In this paper author used a procedure to detect fake news which includes the following steps:

- **Metadata collection.** The first step is to simply collect the sentiment, entities and additional metadata available from third party tools.
- **Relation Extraction**. A second pass will collect both (i) the general relations found in a KG, and (ii) those computed from the current texts.
- **Embeddings.** Last step refers to the adaptation of various neural models (e.g., by adding a layer of embeddings) for improving fake news detection.

In order to perform semantic fake news detection, some additional statements like the past truth history of a speaker or the relations between speakers and publishers should be considered if possible. The idea of using past inaccuracies for each speaker was introduced with the Liar data set and named credit history, but it is rarely used in practice. In this paper author use the Liar data set for the experiments. It contains politics-related articles classified based on the degree of truth, while also offering credit histories that tracks the accuracy of the speaker statements. The data set is split into three partitions (train, test and validation) and includes six classes that need to be predicted: False, Barely-true, Half-true, Mostly-true, True, Pants on fire. The initial paper about the Liar data set identified SVMs as best classical models and CNNs as the best Deep Learning classifiers. A follow-up paper indicates that LSTMs would be even better. Since our focus is not on credit history (five counts for all the classes that are not True including the score for the current statement) but on the impact of the relational features, we do not reproduce those results and do not compare with them. Multinomial Naive Bayes, SGD Classifier, Logistic regression (One Vs Rest), Random forest, Decision trees are the algorithms used by the author for the fake news detection.

[3] Marina Danchovsky Ibrishimova and Kin Fun Li,'' A Machine Learning Approach to Fake News Detection Using Knowledge Verification and Natural Language Processing", © Springer Nature Switzerland AG 2020 L. Barolli et al. (Eds.): INCoS 2019, AISC 1035, pp. 223–234, 2020.

In this paper author applied hybrid framework for fake news detection, which repurposes the machine learning model for incident classification. Author used the logistic regression model on incident classification as our machine learning model for fake news detection and author trained it and validated it on Kaggle Fake News Dataset. In this paper author find that the most promising framework for fake news detection uses a combination of source and fact verification and NLP analysis, and author propose a hybrid framework based on automating incident classification. The automated knowledge verification proposed in this paper relies heavily on the notion of similarity. In particular, it relies on establishing semantic similarity. Several state-of the-art algorithms designed for this purpose have been introduced over the last 20 years, namely latent semantic analysis, latent relational analysis, explicit semantic analysis, temporal semantic analysis, distributed semantic analysis. Future work in automating fake news detection using our proposed framework would involve evaluating these algorithms and deciding on one that best fits authors purpose. Establishing the factual accuracy of a claim is crucial in determining whether it is "fake news" by most definitions of "fake news". Several manually generated tools for identifying the factual accuracy of a given claim exist. However, such approaches rely on humans who may or may not be objective. One method uses Recognizing Textual Entailment (RTE) where "RTE-based models assume that the textual evidence to fact check a claim is given" as part of the claim. Another method relies on checking a claim against a knowledge database of proven facts. Yet another method attempts to verify claims by profiling their source and implementing "credit history" of individual sources. In this paper we define fake news in the context of information warfare. Authors briefly study the socio-political implications of fake news and we investigate previous efforts in automating fake news detection. We find that the most promising framework for fake news detection uses a combination of source and fact verification and NLP analysis, and authors propose a hybrid framework based on our previous work in automating incident classification.

[4] Arvinder Pal Singh Bali(&) , Mexson Fernandes, Sourabh Choubey, and Mahima Goel," Comparative Performance of Machine Learning Algorithms for Fake News Detection", © Springer Nature Singapore Pte Ltd. 2019 M. Singh et al. (Eds.): ICACDS 2019, CCIS 1046, pp. 420–430, 2019.

The proliferation of internet and social websites have led to the exponential growth in opinion spams and fake news, in recent times. The Fake news which diffuse faster than the real news on social websites like Google Plus, Facebook, Twitter etc. Authors have been defined as 'fabricated information' which resemble news media content 'in form but not in organizational process or intent'. Further subdivided fake news into three distinct categories viz. false news, fake satire news and poorly written news articles. The objective of the paper is to compare performances of seven machine learning (ML) algorithms on three standard datasets using a novel set of features and statistically validate the results using accuracies and F1 scores. In this paper author introduced a new n-gram model to detect automatically fake contents, particularly focusing reviews and news. Results of two different feature extraction techniques viz. tf, tf-idf and six machine learning classification techniques were reported by the authors. Linear classifiers viz. linear Support Vector Machine (SVM), Stochastic Gradient Descent (SDG) and Logistic Regression (LR) achieved better results than nonlinear ones for both fake reviews and news. Author extensively reviewed the detection of fake news on social media, from a data mining perspective, evaluation metrics and representative datasets. In this paper the model was tested on three different datasets: (i) Open Sources dataset having 9,408,908 articles out of which 11,161 articles from categories fake and reliable were selected (ii) Kaggle dataset on fake news consisting 20,800 articles and (iii) GitHub repository for fake or real news dataset by George McIntire, having two sections, headlines and text of the news. The author selected the following ML algorithms like Random Forest (RF), Support Vector Classifier (SVC), Gaussian Naïve Bayes (GNB), AdaBoost (AB), K-Nearest Neighbour (KNN), Multi-Layer Perceptron (MLP) and Gradient Boosting for model evaluation. For building a Machine Learning model, feature selection is of utmost importance for optimum performance of the system the author used the features (i) n grams count feature and (ii) tf-idf. In this paper author described XGB (gradient boosting) classifier can efficiently detect fake news with 88% mean accuracy and 0.91 F1 score, outperforming other ML classifiers.

[5] Adrian M. P. Bra¸soveanu, R ˘azvan Andonie," Integrating Machine Learning Techniques in Semantic Fake News Detection", Accepted: 3 October 2020 © Springer Science Business Media, LLC, part of Springer Nature 2020.

Detecting fake news is an interdisciplinary problem, as it requires us to examine which methods where used to disseminate the news (e.g., social networks), the links between the various actors involved (e.g., by using the information available in public Knowledge Graphs like Wikipedia), the propaganda tools (e.g., language can often be examined through the lens of semantics) or even the geopolitics (e.g., as proven by the Cambridge Analytica scandal, some news might be targeted to some specific groups who might be more likely to respond to it). In this paper the author discusses a semantic fake news detection method built around relational features like sentiment, entities or facts extracted directly from text. In this paper the author used the Liar data sets contains politics-related short texts extracted from the Politifact API and classified based on the degree of truth, while also offering credit histories that track the accuracy of the speaker statements. The data set is split into three partitions (train, test and validation) and includes six classes that need to be predicted: True, Mostly True, Half True, Barely True, False and Pants-on-fire. The initial paper about the Liar data set identified SVMs as best classical models and CNNs as the best Deep Learning classifiers. In this paper author used other dataset named Politifact data set also contains short texts extracted from the Politifact API, but, as opposed to the Liar data set, does not provide any additional features except for the identity of the speaker. The data set is split into two partitions (train and test) and includes the following classes that need to be predicted: True, Mostly True, Half True, Mostly False, False and Pants-on-fire. The paper that introduced the data set shows that LSTMs with text and no additional features perform best in both 2-classes (when the data is split only in True and False) and 6-classes (when keeping the original annotations) scenarios. In this paper Multinomial NaiveBayes, SGD Classifier, Logistic regression (One Vs Rest), Random forest Decision trees , SVM are the algorithms used by the author on datasets. Author described SVM is best classic ML classifier among Logistic regression and decision trees and Random forest to detect the fake news.

[6] Abdullah-All-Tanvir, Ehesas Mia Mahir, Ehesas Mia Mahir, Mohammad Rezwanul Huq," Detecting Fake News using Machine Learning and Deep Learning Algorithms",, 2019 7th International Conference on Smart Computing & Communications (ICSCC), 978-1-7281-1557-3/19/$31.00 ©2019 IEEE 2019 .

In this paper author discussed to use machine learning algorithms because this model has 2 classes in news, whether it's real or fake one. To classify a news, we need to understand the problem definition first, then we go for our model and evaluate the result. Machine Learning is replete with its algorithms but some of them are really good for "Fact or Fiction" detection and some are on an average scale. In this paper author used Chile earthquake 2010 Datasets to perform on the model. There are 20,360 data in the dataset. This dataset includes the Label class as H and N where 'H' stands for harassment indicating Fake and 'N' stands for non-harassment indicating Real or Fact. Author used the following different ideas to process the dataset. These are as follows: Count vectors, TF-IDF, word embedding. In this paper for this model author used 5 different types of machine learning algorithms and for the implementation work, we used Python 3.6.5 as our programmable language. The classification models that we implemented using the above- mentioned dataset are Bayesian Model, Logistic Regression & also Support Vector Machine - two most famous deep learning methods RNN Recurrent Neural Network and Long Short-Term Memory were also implemented to see how well our data fit into the model. These algorithms are good for different classifications and they got their own properties and performance based on different datasets. As Naïve Bayes, Logistic Regression and SVM are more commonly used algorithms for classification problems. In this paper, authors analyzed a computerized model for checking the verification of news extracted from Twitter which gives general answers for information accumulation and expository demonstration towards fake news recognition. After having an idea from the supervised models, a deep learning-based model is proposed to identify fake news. In this paper this work exhibits a programmed model for identifying fake news in well-known Twitter strings. Such a model could be important to a huge number of social media users by expanding their own credibility decisions. In this paper the author proposed that SVM performs best for characterization technique among Naive Bayes (NB), Recurrent Neural Network (RNN), Logistic Regression (LR), Long Short Term Memory (LSTM).

3. METHODOLOGY/RECENT TECHNOLOGY

3.1 Proposed method:

The proposed Fake news detection and identification of fake news using classification based feature selection is outlined in the below flow chart.

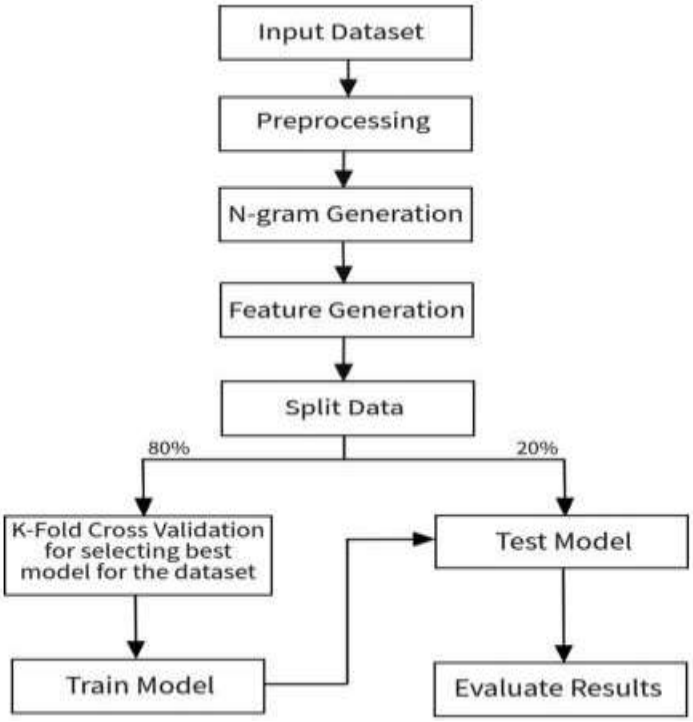

Fig. 1. Flowchart of the classification process

3.2 Pre-processing:

The removal of extraneous information is crucial. Articles that are used, contained links, numbers, and other symbolic contents that are not required for feature analysis. Source URLs if any are removed from the article content. The text of headline and body are then tokenized and stemmed. Finally unigrams, bigrams and trigrams are created out of the list of tokens. These grams and the original text are used by the different feature extractor modules. Lowercase Text, URL Removal, Contraction Splitting are preprocessing techniques.

3.3 Feature Selection:

For building a Machine Learning model, feature selection is of utmost importance for optimum performance of the system.

Features used in the proposed model are as follows:

i) n-grams Count Feature

- These features are used for counting occurrences of n-grams in the title and body of the news, and various ratios of the unique n-gram and total word count given by Eq
- ratio of unique n-gram = total unique n-gram /total n-gram
- Thereafter it uses specific binary refuting words, like ‘fake’, ‘fraud’, ‘hoax’, ‘false’, in the headline.

ii) tf-idf: Term Frequency- Inverse Document Frequency:

- It consists of two terms tf and idf. Term Frequency is how many times a word occurs in a given document given by Eq.

$$tf(t,d) = 0.5 + 0.5.\frac{f_{t,d}}{\max\{f_{t',d}:t' \in d)\}}$$

- Inverse Document Frequency (idf) is the number of times a word occurs in corpus of documents. This facilitates to understand which words are important.
- Finally, Cosine similarity of these normalized tf-idf vectors are calculated for headlines and the contents. This gives the measure of how correlated the headlines and their corresponding article contents are. Since cosine similarity considers only those vectors which have non-zero dimension, its calculation is quite fast.

$$Similarity = Cos\,\theta = \frac{A.B}{\|A\|\|B\|}$$

SUPPORT VECTOR MACHINE:

Support Vector Machine is one of the most popular Supervised Learning algorithms , which is used for classification as well as Regression problems. However, primarily it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine, there are two different categories that are classified using a decision boundary or hyperplane. The basic principle of the algorithm is to use samples to form a lattice in the high-level feature space, select sample points as the support vector near the boundary between the two types of sample points and use the support vector to make the decision. And finally achieve the purpose of classification and identification.

A support vector machine (SVM), which can be used interchangeably with a support vector network (SVN), is also considered to be a supervised learning algorithm. SVMs work by being trained with specific data already organized into two different categories. Hence, the model is constructed after it has already been trained. Furthermore, the goal of the SVM method is to distinguish which category any new data falls under, in addition, it must also maximize the margin between the two classes. The optimal goal is that the SVM will find a hyperplane that divides the dataset into two groups. The kernel used in this application is RBF as it is best suited for large applications like a corpus of news articles. The Radial Basis function on two samples and it is given by:

$$K(x,x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$$

- Where ||x-x'||^2 represents the squared Euclidean distance and is a free parameter.

Flowchart:

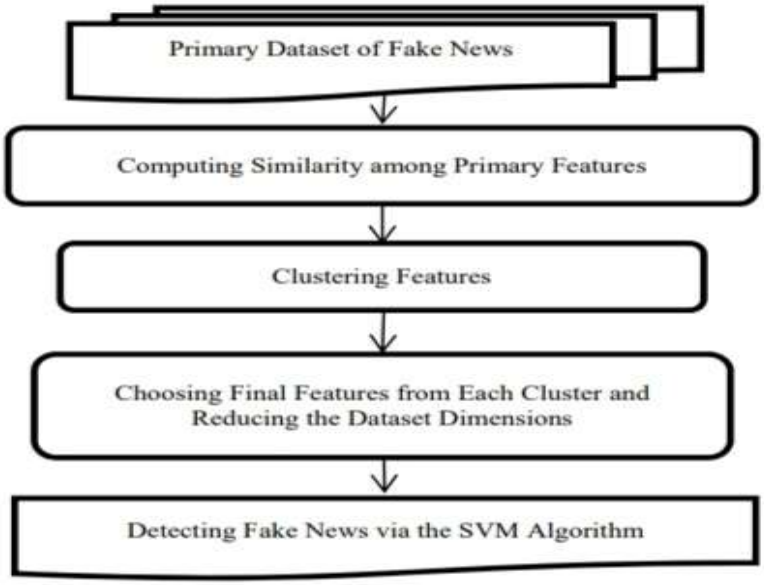

Fig. 2: Flowchart for detecting fake news using SVM

The flowchart states that detecting of the fake news detection using svm algorithm.

Graph Analysis for SVM:

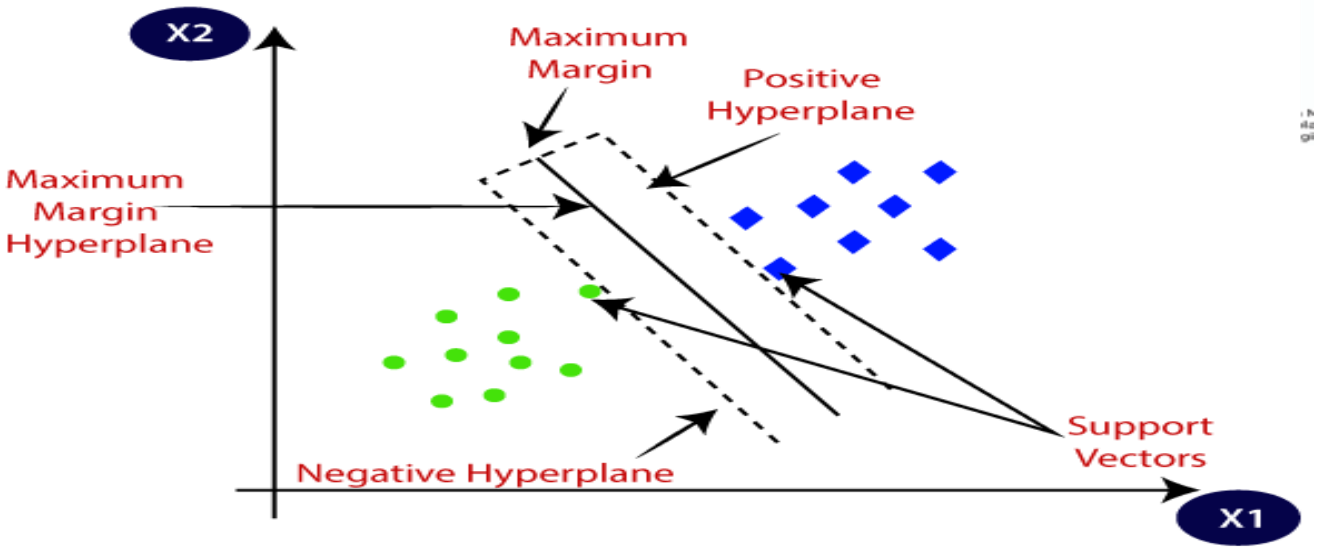

Fig. 3: SVM Graph

NAIVE BAYES

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in *text classification* that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability. The formula for Bayes' theorem is given as:

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

There are three types of Naive Bayes Model, which are given below:

Gaussian: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

Multinomial: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc.
The classifier uses the frequency of words for the predictors.

Bernoulli: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

It is used for **Credit Scoring**. It is used in **medical data classification**. It can be used in **real-time predictions** because Naïve Bayes Classifier is an eager learner. It is used in Text classification such as **Spam filtering** and **Sentiment analysis**

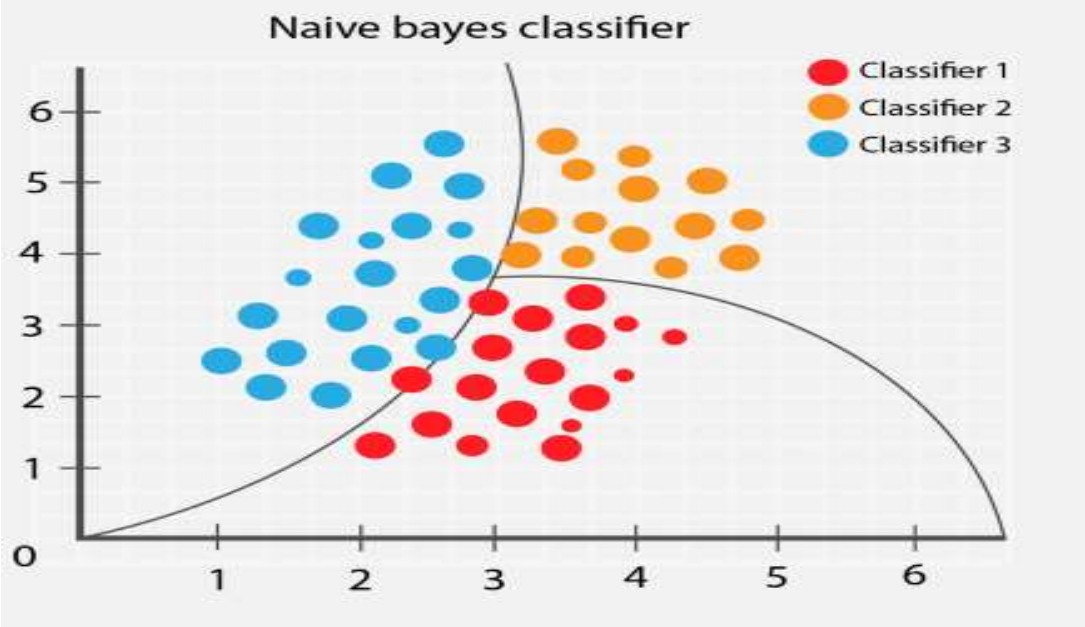

Fig. 4: Naïve bayes classifiers

FLOWCHART

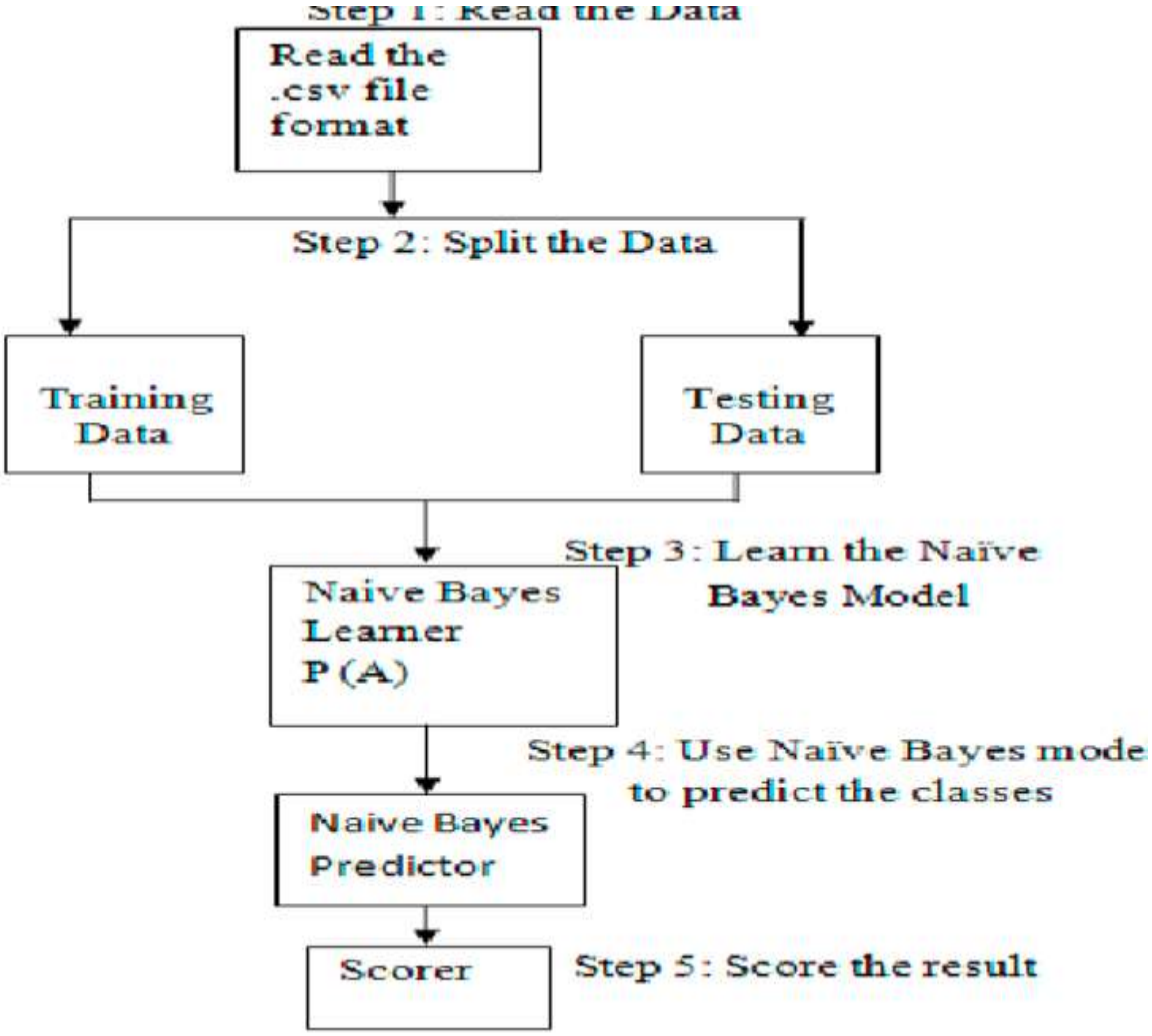

Fig. 5: Flow chart for Naïve bayes

LOGISTIC REGRESSION

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function

indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

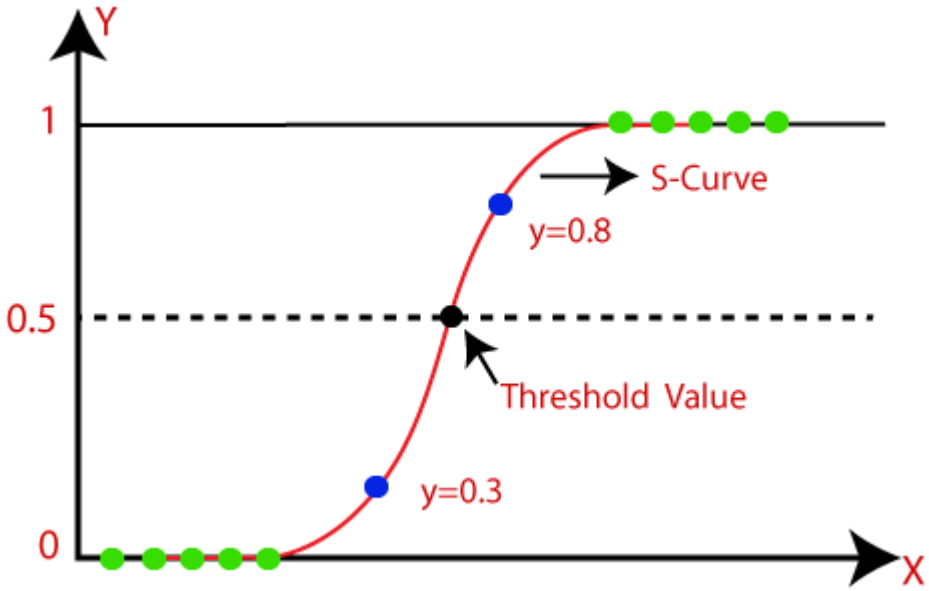

Fig. 6: Logistic regression

Logistic Function

The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

We know the equation of the straight line can be written as:

$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}$$; 0 for y= 0, and infinity for y=1

But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Logistic Regression can be classified into three types:

Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"

Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

4. RESULTS AND DISCUSSION

There are basically four feature vectors that was extracted from our text dataset, which are Count Vector, Word Level Vectors , N-gram vectors, Character level vectors. Firstly, Naïve Bayes Model was tested on each of the feature vectors. It gives 73% accuracy on count vector feature, 75% on Word Level TF-IDF, N - gram vector, and character vector as well. Then Logistic regression model was performed. This time the performance was slightly optimized than before, predicting 74% and 76% respectively on count and word level vectors. Whereas 75% and 76% was the results of Logistic Regression stage. Thirdly Support Vector Machine was performed for observing any further enhancement over the previous result so far. But no enhancements were observed at this stage as testing accuracy from SVM is 74% in all the four mentioned features vectors in our study.

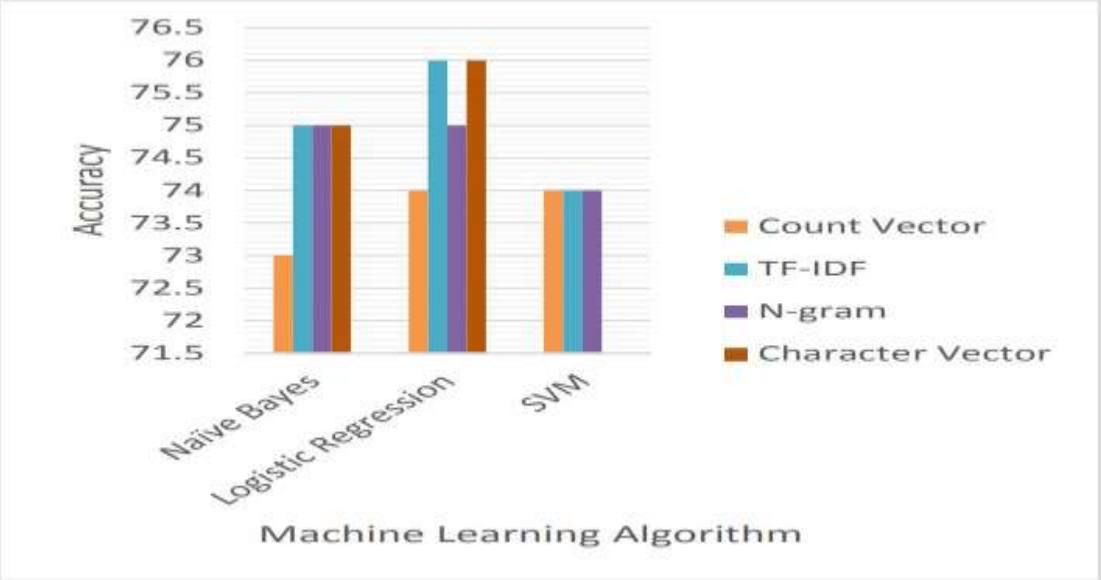

Fig. 7: Accuracy for all four feature vectors

After performing cross validation. There is 62.47% for Logistic Regression, 84.56% for Naïve Bayes and 89.34% for SVM in Count Vector feature. And for TF-IDF, there is 69.47% for Logistic Regression, 89.06% for Naïve Bayes and 89.34% for SVM

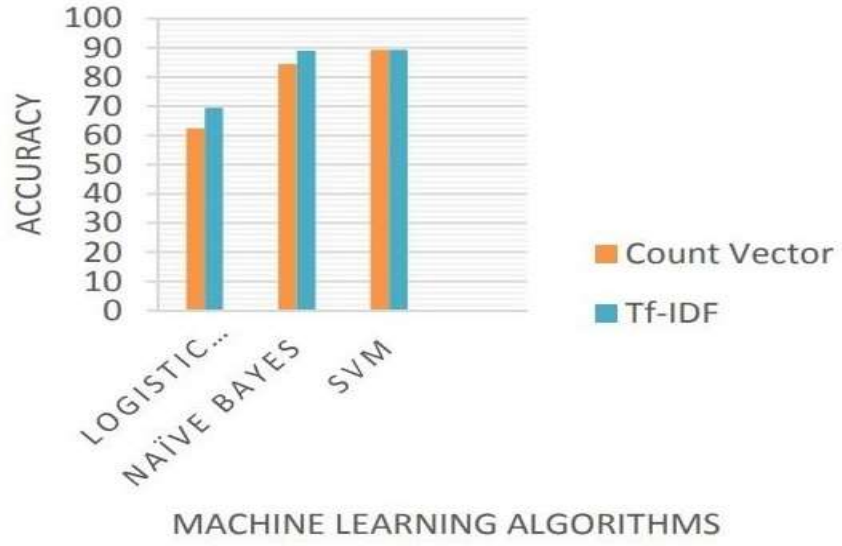

Fig. 8: Accuracy of ML algorithms on tf-idf and counter vector

SVM Shows the highest accuracy with TFIDF feature among all the classifier we have implemented in Chile Earthquake Dataset 2010. Logistic regression attempts to predict outcomes based on a set of independent variables, but logistic models are vulnerable to overconfidence. It requires that each data point be independent of all other data points. In our dataset, the feature word length depends on the news statement. If observations are related to one another, then the model will tend to overweight the significance of those observations.

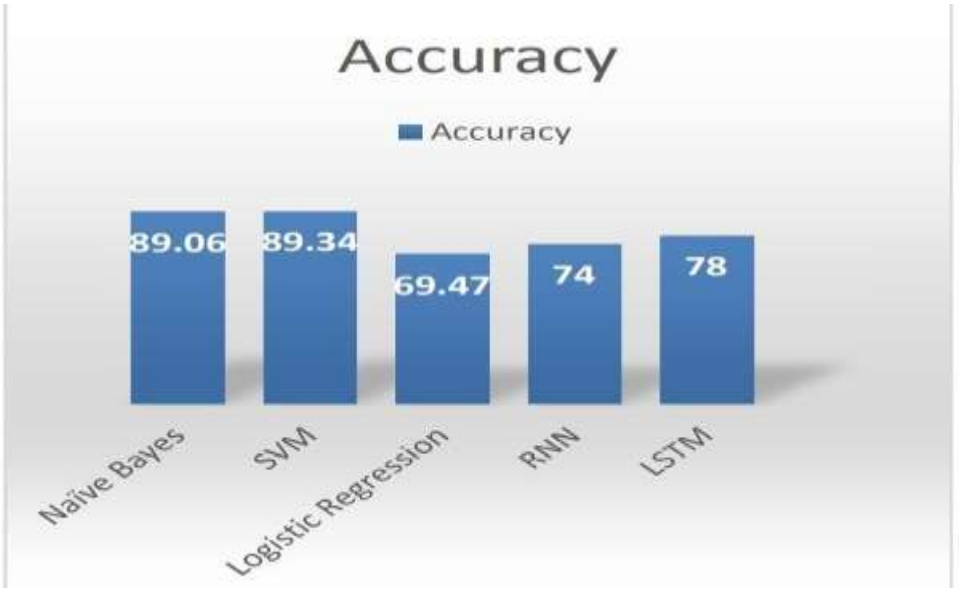

Fig. 9: Overall Accuracy Comparison of ML and DL Algorithms

All three algorithms have same level of precision. SVM outperforms the other algorithms considering recall. F1-score of both Naïve Bayes and SVM is 0.94 which is the highest.

	Precision	**Recall**	**F1-Score**
Naïve Bayes	0.89	0.99	0.94
Logistic Regression	0.89	0.75	0.81
SVM	0.89	1.0	0.94

Table 1- Table for Precision, recall, f1-score calculation on ml algorithms

5. CONCLUSION

In conclusion, fake news are false stories in order create propaganda or influence the public on a political or sociological issue with the assistance of social media, internet, and established news sources. Fake news serves various purposes to society such as influencing and persuade people through the means of fake material and damaging a person, event, idea, concept, or people in general. As shown, the three main causes are; the social media and internet impact, the rise of unreliable news sources and fall of authoritative news outlets, and lastly the anonymity behind the creation of fake news and misinformation.

Early detection of fake news is of primary importance for public and the society, for redesigning the 'information ecosystem in the 21st century' which would eventually lead to the creation of a system and culture having values that promote truth. In this study, there is a computerized model for checking the verification of news extracted from Twitter which gives general answers for information accumulation and expository demonstration towards fake news recognition. After having an idea from the supervised models, a deep learning-based model is proposed to identify fake news. The accuracy metric presumably would be altogether improved by methods for utilizing progressively complex model. It is worth noting, that even with the given dataset, only part of the information was used. The study demonstrated that even the very basic algorithms on fields like AI and Machine Learning may find a decent outcome on such a critical issue as the spread of fake news issues worldwide.

Such a model could be important to a huge number of social media users by expanding their own credibility decisions. The dataset in this examination is relied upon to be utilized for arrangements which utilized machine learning based statistical calculations, for example, Support Vector Machines (SVM), Naive Bayes (NB), Recurrent Neural Network (RNN), Logistic Regression (LR), Long Short Term Memory (LSTM). In this investigation, SVM performs best for characterization technique.

6. REFERENCES

[1] Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita, "Fake News Detection Using Machine Learning approaches: A systematic Review", IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8 ©2019 IEEE.

[2] Adrian M. P. Bra¸soveanu, R˘azvan Andonie," Semantic Fake News Detection: A Machine Learning Perspective", c Springer Nature Switzerland AG 2019 I. Rojas et al. (Eds.): IWANN 2019, LNCS 11506, pp. 656–667, 2019. https://doi.org/10.1007/978-3-030-20521-8_54.

[3] Marina Danchovsky Ibrishimova and Kin Fun Li,,'' A Machine Learning Approach to Fake News Detection Using Knowledge Verification and Natural Language Processing", © Springer Nature Switzerland AG 2020 L. Barolli et al. (Eds.): INCoS 2019, AISC 1035, pp. 223–234, 2020. https://doi.org/10.1007/978-3-030-29035-1_22.

[4] Arvinder Pal Singh Bali(&) , Mexson Fernandes, Sourabh Choubey, and Mahima Goel," Comparative Performance of Machine Learning Algorithms for Fake News Detection", © Springer Nature Singapore Pte Ltd. 2019 M. Singh et al. (Eds.): ICACDS 2019, CCIS 1046, pp. 420–430, 2019. https://doi.org/10.1007/978-981-13-9942-8_40.

[5] Adrian M. P. Bra¸soveanu,·R ˘azvan Andonie," Integrating Machine Learning Techniques in Semantic Fake News Detection", Accepted: 3 October 2020 © Springer Science+Business Media, LLC, part of Springer Nature 2020 https://doi.org/10.1007/s11063-02-x.

[6] Abdullah-All-Tanvir, Ehesas Mia Mahir, Ehesas Mia Mahir, Mohammad Rezwanul Huq," Detecting Fake News using Machine Learning and Deep Learning Algorithms ",, 2019 7th International Conference on Smart Computing & Communications (ICSCC), 978-1-7281-1557-3/19/$31.00 ©2019 IEEE 2019.