International Journal of Innovative Research in Engineering
Volume 3, Issue 2 (March-April 2022), PP: 276-279.
www.theijire.com
ISSN No: 2582-8746

Fake News Detection

Anjali Singh[1], Deepshikha Singh[2], Diksha Singh[3], Diksha Sharma[4]
[1,2,3]B.Tech (pursuing), Department of Computer Science, Institute of Technology and Management (ITM) , Gida , Gorakhpur, U.P-273209,India.
[4]Associate Professor, CSE Department, Institute of Technology and Management (ITM) Gida, Gorakhpur, Uttar Pradesh-273209, India.

How to cite this paper:
Anjali Singh[1], Deepshikha Singh[2], Diksha Singh[3], Miss Diksha Sharma[4], "Fake News Detection", IJIRE-V3I02-276-279.

Abstract: *Research on Fake news detection considers previous and current methods for fake news detection in textual formats while detailing how and why news fake exists in the first place. This research paper includes various methods and concepts that is to be discussed like on Machine learning algorithms, Network Analysis approaches, and proposes a three-part method using Naïve Bayes Classifier, Support Vector Machines, and Semantic Analysis as an accurate way to identify fake news on social media. There are different social media platforms that are accessible to the users like Facebook , Instagram , Twitter, Whatsapp etc. Any user can make a post that's misleading or false and spread the false news through these online platforms. These platforms do not verify those users or their posts. In this research paper, we aim to perform binary classification of various news articles available online with the help of concepts related to Artificial Intelligence, Natural Language Processing and Machine Learning. We also aim to provide the user with the ability to classify the news as fake or real and also check the authenticity of the website that is publishing the news. Therefore, this research compares the existing approaches to build the models and with further improvements to be expected by using the combination of different machine learning techniques.*

Key Word: *Fake news, social media, Network Analysis, Naïve Bayes Classifier, Support Vector Machine, Machine learning, Text Classification, Natural Language Processing.*

I. INTRODUCTION

World is changing rapidly. No doubt we have a number of advantages of this digital world but it also has its disadvantages as well. There are different issues in this digital world. One of them is fake news. Someone can easily spread a fake news. Fake news is spreading to harm the reputation of a person or an organization or for any political purpose. It can be a propaganda against someone that can be a political party, leader or an organization. There are different online platforms and apps where the person or user can spread the fake news. This includes Facebook, Twitter etc. Different types of machine learning algorithms are available that involves the supervised, unsupervised and reinforcement machine learning algorithms. The different algorithms first have to be trained with a data set called train data set. After the training, these can be used to perform different tasks. Mostly, machine learning algorithms are used for prediction purpose or to detect and identify something that is hidden. The current project involves utilizing machine learning, its algorithms and natural language processing techniques to create a model. Many of the present automated approaches to this problem are centered around a "blacklist" of authors and sources that are known as producers of fake news. But, what about when the author is not known or when false news is published through a commonly reliable source? In these cases, it is necessary to depend simply on the content of the news article to make a decision on whether it is fake or not. By collecting examples of both the real and fake news and training a model, it should be possible to classify fake news articles with a certain degree of accuracy. The goal of this project is to find the effectiveness and limitations of language-based techniques for detection of fake news through the use of machine learning algorithms.

This literature review will answer the different research questions. The importance of machine learning to detect fake news will be proved in its literature review part. It will also be discussed how machine learning may be used for detecting the false news. Machine learning algorithms used to detect false news will be discussed in the literature review of our research.

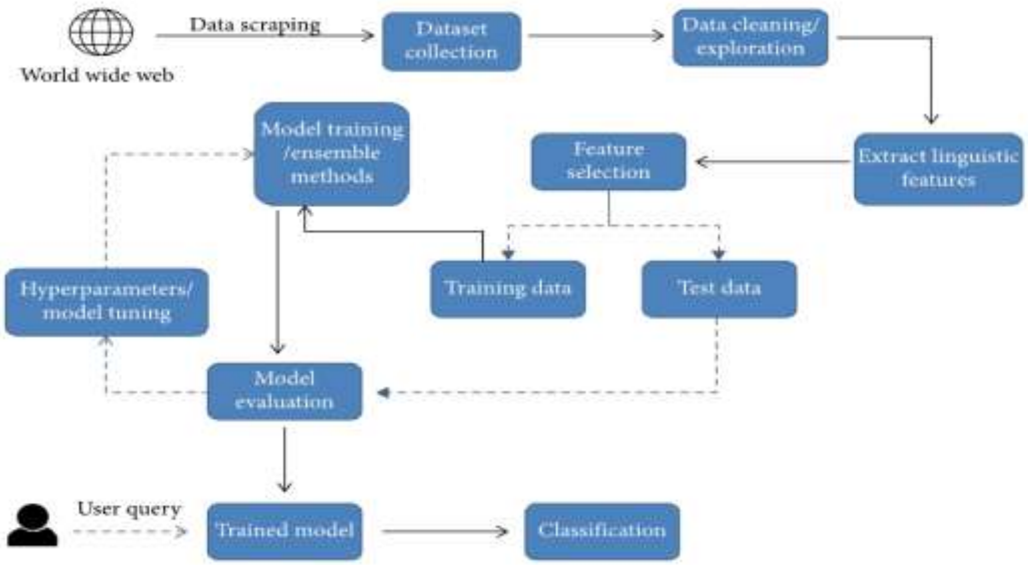
II. TECHNIQUES USED IN FAKE NEWS DETECTION

TFIDF (Term-Frequency Inverse Document Frequency) As it is known that the input set of data must be cleaned by removing or stopping ,punctuation words and common words that appear in English grammar, to convert text to word count vectors. It focuses on the occurrence of the words in the documents. It is done by assigning each word with a unique or different number. The value in each position in the vector could be filled with a count or frequency of each word in the document. TFIDF are word frequency scores that try to highlight the words. Term frequency-inverse document frequency is a text victory which transforms or changes the text into a usable vector. It combines of two concepts, Term Frequency (TF) and Document Frequency (DF). The term frequency is defined as the number of occurrences of a particular word or term in a document that are is interesting.

$$\boxed{w_{i,j} = tf_{i,j} \times \log \frac{N}{df_i}}$$

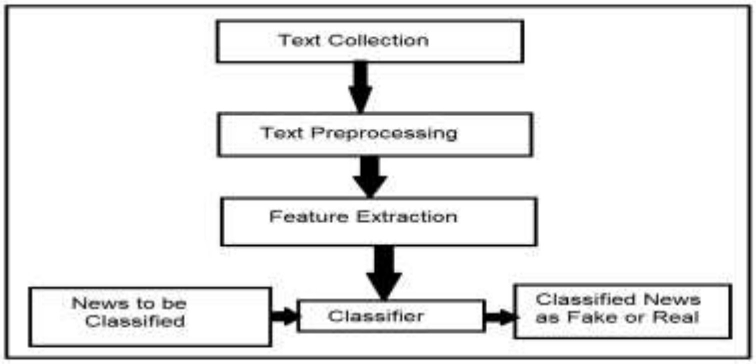
$Tf_{i,j}$ = Number of Occurrences of I in j
Df_i = Number of Documents Containing i
N= Total Number of Document

III. PROPOSED FRAMEWORK



IV. METHODOLOGY

The Fake News model detection is built using steps like Text Collection, Text Preprocessing, Feature Extraction and then finally classification using different classifiers.



- **Logistic Regression**

Logistic regression is a process of modeling the probability of a discrete outcome when an input variable is given. The most common logistic regression models a binary outcome; something which can take two values such as true or false, yes or no, and so on. Multinomial logistic regression can model those scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where one is trying to determine that if a new sample fits best into a category or not. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique.

- **Decision Tree Classification**

Decision Tree is a Supervised learning technique which can be used for both of the classification and Regression problems, but mostly it is preferred for solving the Classification problems. It is a tree-structured classifier, where the internal nodes denotes or represents the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

- **Gradient Boosting Classifier**

Gradient Boosting is defined as a popular boosting algorithm. In gradient boosting classifier, each predictor corrects the error of it's predecessor. In contrast to Adaboost, the weights of the training instances are not modified instead of that, each predictor is trained using the residual errors of predecessor as levels.

- **Random Forest Classifier**

Random Forest is a trademark term for an ensemble of decision trees. In Random Forest classifier, we have a collection of decision trees (so it is known as "Forest"). To classify a new object that is based on attributes, each tree gives a classification and we say that the tree "votes" for that class. The forest chooses the classification which is having the most votes (over all the trees in the forest). The random forest is a classification algorithm that consists of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. Random forest, like its name indicates, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class that has most of the votes becomes our model's prediction. The reason that the random forest model works so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

- **Passive Aggressive Classifier Algorithm**

Passive-Aggressive classifier algorithms are generally used for large-scale learning. It is one of the few 'online-learning algorithms'. In online machine learning algorithms, the input data comes in a sequential order and the machine learning model is updated step-by-step, as opposed to batch learning, where the entire training dataset is used at once. This is very useful in those situations where there is a large amount of data and it is computationally infeasible to train the entire dataset because of the sheer size of the data. We can simply say that an online-learning algorithm will get a training example, update the classifier, and then throw away the example.

V. RESULT

The accuracy is achieved by algorithm on four datasets. The maximum accuracy achieved on Fake News Dataset is 99%. Linear SVM achieved an accuracy of 98%.The average accuracy attained by ensemble learners is 97.67%. Where as corresponding average for individual Learners is 95.25%. The exact difference between individual learner and ensemble learners is 2.42%.The scope of this project is to cover the Dataset for Fake News Detection which is labeled by Fake or true news. The results of analysis of dataset using the algorithms that have been depicted using the confusion matrix.
The algorithms are as:-

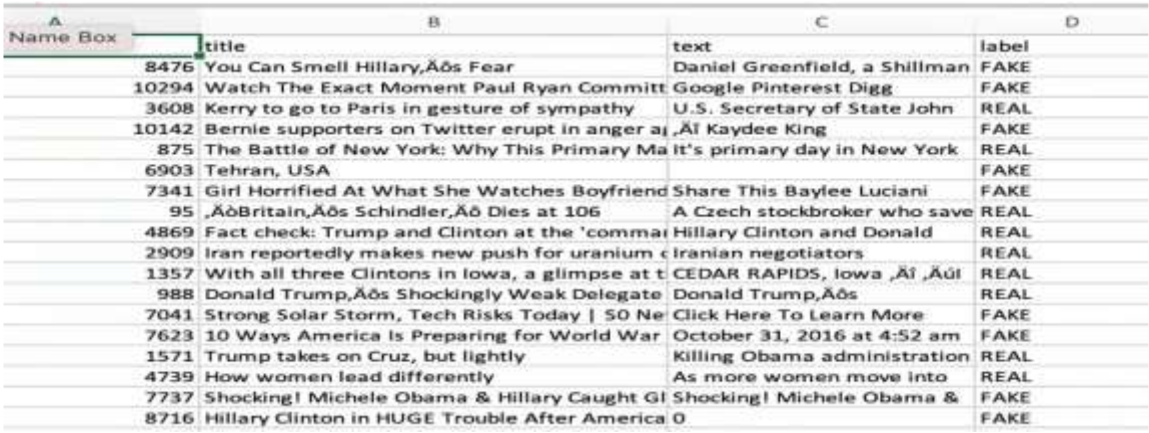
1. Naïve Bayes
2. K-Nearest Neighbors(KNN)
3. SVM

VI. CONCLUSION

We are attempting to use machine learning algorithm and ensemble technology to work a platform that identifies take news on the basis of Patterns in the text . We primarily use LIWC tool to extract different textual feature from article and use the feature as input to models. Attempts are made to achieve maximum accuracy. It is easier way and faster to use a pertained neural network, and obtain the results on our dataset. More complex and efficient methods that could be applied to this dataset by using the entire text or extracting other features. One of the great things about it is to that extremely difficult to trained state of neural network.

Sample Dataset

The size of the data set is 77964,it means 7796 rows and 4 column 70% of the data is used for Training the Machine learning Model 30% of the data is used to test the model. Accuracy score of the model is 95.19%



A	B	C	D
Name Box	title	text	label
8476	You Can Smell Hillary‚Äôs Fear	Daniel Greenfield, a Shillman	FAKE
10294	Watch The Exact Moment Paul Ryan Committ	Google Pinterest Digg	FAKE
3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John	REAL
10142	Bernie supporters on Twitter erupt in anger a	‚Äî Kaydee King	FAKE
875	The Battle of New York: Why This Primary Ma	It's primary day in New York	REAL
6903	Tehran, USA		FAKE
7341	Girl Horrified At What She Watches Boyfriend	Share This Baylee Luciani	FAKE
95	‚ÄòBritain‚Äôs Schindler‚Äô Dies at 106	A Czech stockbroker who save	REAL
4869	Fact check: Trump and Clinton at the 'comma	Hillary Clinton and Donald	REAL
2909	Iran reportedly makes new push for uranium c	Iranian negotiators	REAL
1357	With all three Clintons in Iowa, a glimpse at t	CEDAR RAPIDS, Iowa ‚Äî ‚Äúl	REAL
988	Donald Trump‚Äôs Shockingly Weak Delegate	Donald Trump‚Äôs	REAL
7041	Strong Solar Storm, Tech Risks Today | S0 Ne	Click Here To Learn More	FAKE
7623	10 Ways America Is Preparing for World War	October 31, 2016 at 4:52 am	FAKE
1571	Trump takes on Cruz, but lightly	Killing Obama administration	REAL
4739	How women lead differently	As more women move into	REAL
7737	Shocking! Michele Obama & Hillary Caught Gl	Shocking! Michele Obama &	FAKE
8716	Hillary Clinton in HUGE Trouble After America	0	FAKE

References:

[1]. Antani, S., Rodney Long, L., Thoma, G.R. 2004. Based Image Retrieval for Large Biomedical Image Archives, MEDINFO 2004, 829-33

[2]. A test bed on region-based image retrieval using multiple segmentation algorithms with MPEG-7 experiment Model T::he Schema Reference System.

[3] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.

[4].Color Texture Analysis-based Image Retrieval www2.wmin.ac.uk/oreillyp/Posters/Poster _AM.Hoang.ppt

[5]. Chun, J., Stockman, G. (2001). Sub band image segmentation using VQ for content-based image processing, Proceedings of the ninth ACM international conference on Multimedia September 30 - October 05, 2001 Ottawa, Canada, 486-8

[6.]H. G Kaganami and Z. Beij, "Region Based Detection vs Edge Detection", IEEE Transactions on Intelligent information hiding and multimedia signal processing. 2009

[7] Fake news websites. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Fake_news_website. Accessed Feb. 6, 2017

[8] Cade Metz. (2016, Dec. 16). The bittersweet sweepstakes to build an AI that destroys fake news.

[9] Conroy, N., Rubin, V. and Chen, Y. (2015). "Automatic deception detection: Methods for finding fake news" at Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.