International Journal of Innovative Research in Engineering
Volume 3, Issue 3 (May-June 2022), PP: 14-18
www.theijire.com
ISSN No: 2582-8746

Fake News Detection using Machine Learning

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, Uttar Pradesh, 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], Diksha Sharma4, "Fake News Detection using Machine Learning", IJIRE-V3I03-14-18.

Abstract: *A lot of research is already Focused on detecting the Fake news. A lot of things read online, especially in the social media feeds, which may appear to be true, often is not. Fake news is news, stories or mislead or deceive readers. Usually, the Fake news on social media and various other media is spreading and creates serious concern, these stories are created to either influence people's views, or cause confusion and can often be a profitable business for online publishers. The spread of fake news in today's digital world has effected beyond a specific group. Mixing both believable and unbelievable information on social media has made the difficulty of truth. That is, the truth will be hardly classified. This paper comes up with the applications of NLP (Natural Language Processing) techniques for the process will result in feature extraction and vectorization using a skit learn we build tfidf vectorization 'fake news', which is the misguiding news that is being published through unknown sources usually through social media due to its ability to cause a lot of social with destructive impacts.*

Keyword: *NLP, logistic regression, naïve bayes classifier, SVM, Fake News*

I. INTRODUCTION

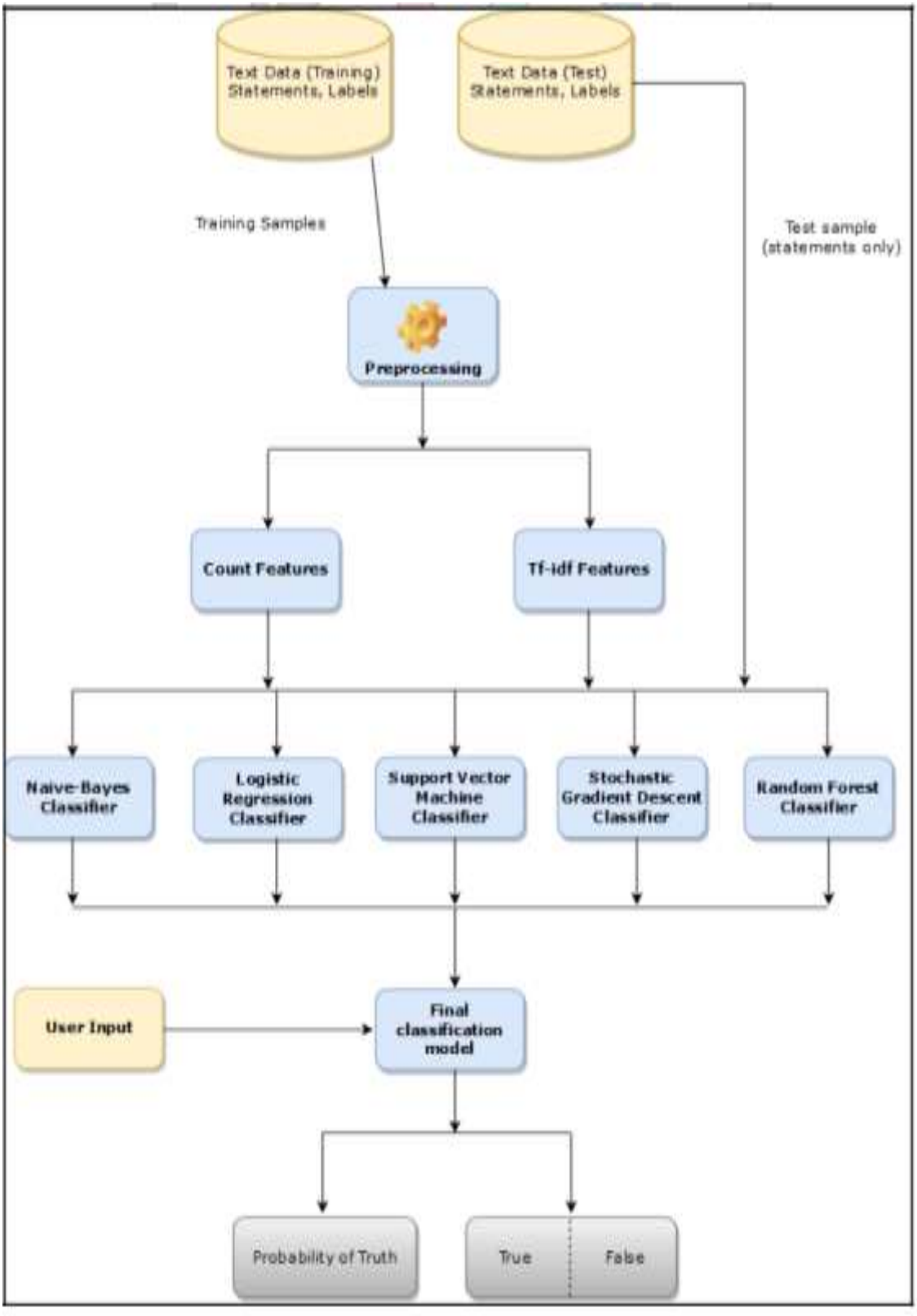
These days, fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content.

The Fake news is creating different issues from irony articles to a news and plan government propaganda in some outlets. Fake news and real news in the media are growing problems in our society. Fake news contains misleading story is "fake news" but lately social media's changing its definition. Some of them are use the term to dismiss the facts counter to their preferred viewpoints. The importance of deception within American political discourse was the subject of weighty attention, particularly following the American president election. The term 'fake news' became common for the issue, which generally arises through the social media particularly to describe factually incorrect and misleading articles or information which published mostly for the purpose of making money through page views. In this paper, it is produce a model that can accurately predict the possibility that a given article is fake news. Facebook has been at the attraction of much review following media attention. They have already implemented a feature to flag fake news on the site when a user sees it; they have also said publicly they are working on to distinguish these articles in an automated way. Certainly, it is complicated task. A given algorithm must be politically unbiased – since fake news exists on both ends of the spectrum – and also give equal balance to legitimate news sources on either end of the spectrum. In addition, the question of legitimacy is a difficult one. However, in order to solve this problem, it is necessary to have an understanding on what Fake News is. It is needed to look into how the techniques in the fields of machine learning, natural language processing help us to detect fake news. The main purpose of this system is to detect the fake news, which is a classic text classification problem with a straight forward proposition. It is needed to build a model that can differentiate between "Real" news and "Fake" news.It maintains lie about a certain statistics in a country.

II. METHODOLOGY

The main reason for utilizing Natural Language Processing is to consider one or more specialization of system or an algorithm. This method comes with the utilizing the applications of NLP (Natural Language Processing) techniques for detecting the 'fake news' or real, that is deceive news stories that comes from the non-reputable sources. In this paper a model is built based on the count the TFIDF vectorizer or a TFIDF matrix (i.e. word appears relatives to how often they are used in other articles in your dataset) can help. Since this problem is a kind of text classification, and implementing a Naive Bayes classifier and Logistic Regression will be better as for text-based processing. The actual goal is to reduce the time gap between a news release and detection developing a model which was the text transformation (count vectorizer vs TFIDF vectorizer) and choosing which type of text to use (headlines vs full text). Now the next step is to extract the most optimal features for count vectorizer or tfidf-vectorizer, this is done by using a knn K-Nearest Neighbor algorithm is the most used words, and/or phrases, lower casing not only removing the stop words which are common words such as "the", "when", and "there" and only using those words that appear at least a given number of times in a given text dataset.

III. SYSTEM ARCHITECTURE



Initially, the training data is extracted i.e taken out and the TFIDF matrix & count vectorizer are generated after preprocessing. This preprocessed data and the test data are sent to the different classifier methods. A final classification model is selected. In this research paper, logistic regression classification model is used. The input data is transferred to the final classification model which gives out result of the test data i.e., "True" or "False" and also gives the probability of truth as output.

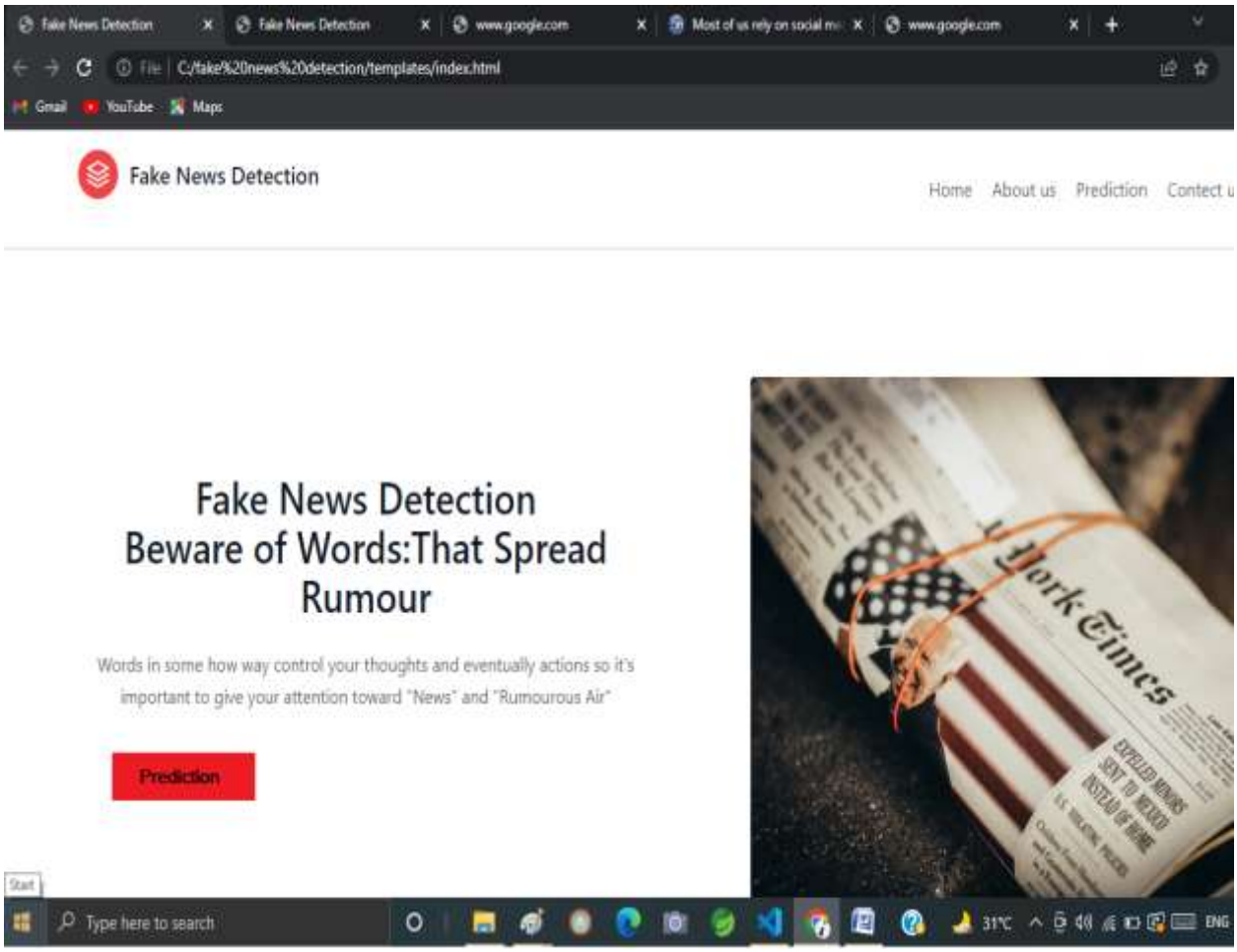
IV.SAMPLE DATASET

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%

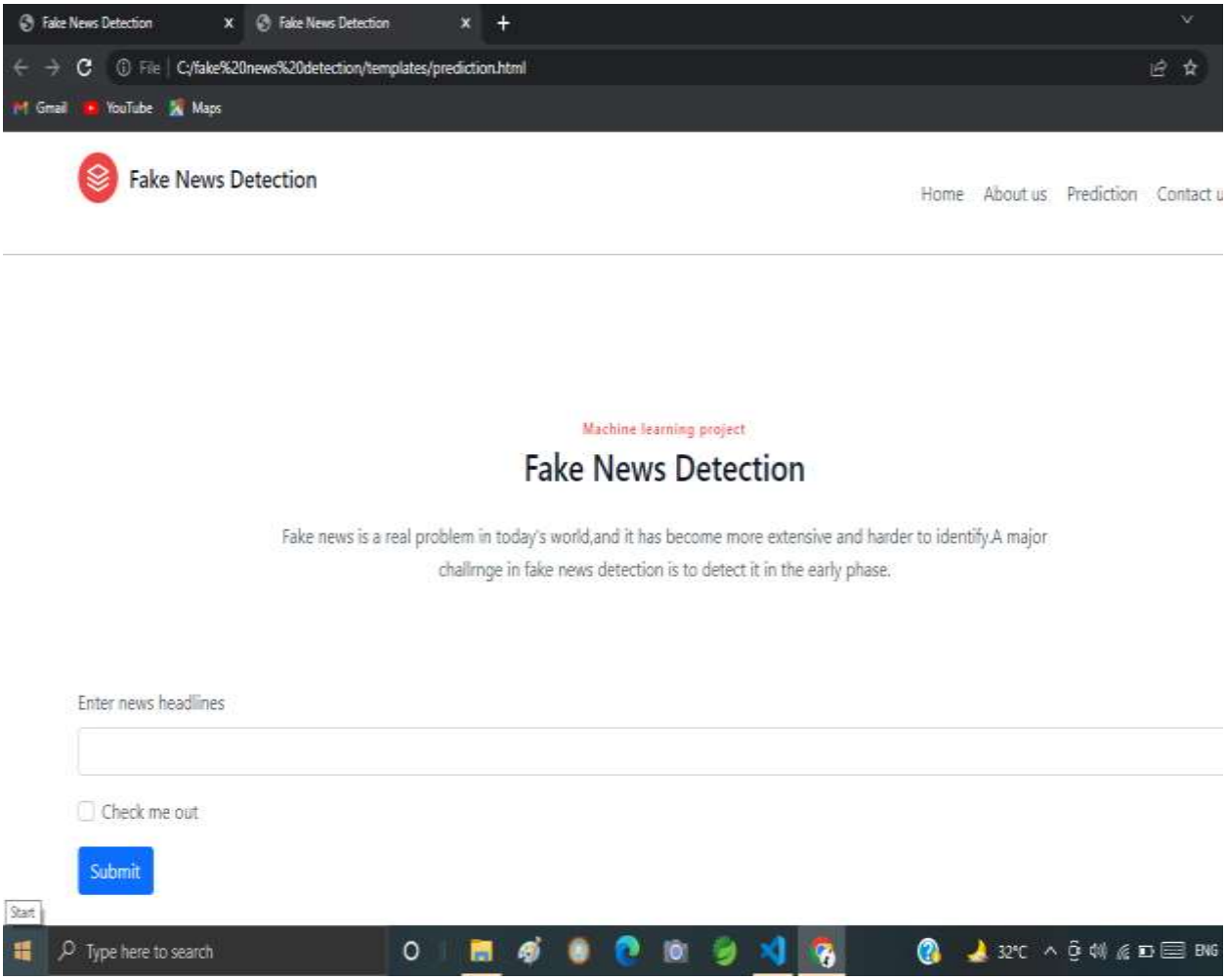
Name Box	title	text	label
8476	You Can Smell Hillary,Äôs Fear	Daniel Greenfield, a Shillman	FAKE
10294	Watch The Exact Moment Paul Ryan Committ	Google Pinterest Digg	FAKE
3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John	REAL
10142	Bernie supporters on Twitter erupt in anger 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

V. RESULT

This is our front page click on prediction then new page are open:



The news statement which is to be tested is entered in the textbox that is present under the caption "Enter news headlines" as shown below and click in check box then submit:



VI.DISCUSSION

Internet is one of the great sources of information for its users (Donepudi, 2020). There are different social media platforms that includes Facebook or Twitter that helps the people to connect with other people. Different kind of news are also shared on these platforms. People nowadays prefer to access the news from these platforms because these are easy to use and easy to access platforms. Another advantage to the people is that these platforms provide options of comments, reacts etc. These advantages attract people to use these platforms (Donepudi et al., 2020b). But as like their advantages, these platforms are also used as the best source by the cyber criminals. These persons can spread the fake news through these platforms. There is also a feature of sharing the post or news on these platforms and this feature also proves helpful for spreading such fake news. People start believing in such news as well as shares the news with other peoples. Researchers in (Zubiaga et al., 2018) said that it is difficult to control the false news from spreading on these social media platforms. Anyone can be registered on these platforms and can start spreading news. A person can create a page as a source of news and can spread the fake news. These platforms do not verify the person whether he is really reputable publisher. In this way, anyone can spread news against a person or an organization. These fake news can also harm a society or a political party. The report shows that it is easy to change people opinions by spreading fake news (Levin, 2017). Therefore, there is a need for detecting these fake news from spreading so that the reputation of a person, political party or an organization can be saved.

VII. CONCLUSION AND FUTURE SCOPE

In this research paper, it is been addressed the task of automatic identification of fake news. We also introduced two new fake news datasets, one is obtained through crowd sourcing and another one is obtained from the web covering celebrities. We developed classification models that depends on a combination of lexical, syntactic, and semantic information, as well features which are representing text readability properties. Our best performing models had achieved accuracies that are comparable to human ability to spot fake content.

This paper is based on the crowd sourcing dataset and the web covering dataset. These are the static datasets. Through these, we can only test the data which is present in the predefined training data sets. The research paper gives the appropriate and correct result for the test data which is present in the training datasets. Thus, the future scope of the paper is connecting this methodology to the internet news which gives results even for the test data that is not present in the training data sets. We can either change to any other better classifier to classify the data other than naïve bayes algorithm and logistic regression.

References

[1] Majed Alrubaian, Muhammad Al-Qurishi, A Credibility Analysis System for Assessing Information on Twitter, IEEE Transactions on Dependable and Secure Computing, 1-14. DOI : http://dx.doi.org/ 10.1109/TDSC.2016.2602338

[2] Lakshmisri Surya, "HOW GOVERNMENT CAN USE AI AND ML TO IDENTIFY SPREADING INFECTIOUS DISEASES", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.6, Issue 1, pp.899-902, March 2018, Available at :http://www.ijcrt.org/papers/IJCRT1133873.pdf

[3] Manish Gupta, Peixiang Zhao, Jiawei Han, 2012. Evaluating Event Credibility on Twitter, Proceedings of the 2012 SIAM International Conference on Data Mining, 153-164, DOI: http://epubs.siam.org/doi/abs/10.1137/ 1.9781611972825.14

[4] Krzysztof Lorek, Jacek Suehiro-Wici´nski, Micha l Jankowski-Lorek, Amit Gupta, Automated credibility assessment on twitter, Computer Science, 2015, Vol.16(2), 157-168, DOI: http://dx.doi.org/10.7494/csci. 2015.16.2.157

[5] Ballouli, Rim El, Wassim El-Hajj, Ahmad Ghandour, Shady Elbassuoni, Hazem M. Hajj and Khaled Bashir Shaban. 2017. "CAT: Credibility Analysis of Arabic Content on Twitter." WANLP@EACL (2017).

[6] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.

[7] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 2012, pp. 171–175.

[8] Ravi Teja Yarlagadda, "INTERNET OF THINGS & ARTIFICIAL INTELLIGENCE IN MODERN SOCIETY", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.6, Issue 2, pp.374-381, April 2018, Available at :http://www.ijcrt.org/papers/IJCRT1133934.pdf

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