Realtime News Analysis using Natural Language Processing

Aakansha Ramesh
Department of Information Technology, SIES Graduate School of Technology
Navi Mumbai, India
aakansharamesh19@siesgst.ac.in

Gauri Thube
Department of Information Technology, SIES Graduate School of Technology
Navi Mumbai, India
thubegauri19@siesgst.ac.in

Swaranjali Jadhav
Department of Information Technology, SIES Graduate School of Technology
Navi Mumbai, India
swaranjalimanik19@siesgst.ac.in

***Abstract*— Fake news has become a huge issue in present times; it is both a societal and technical issue. It is difficult for companies to identify as the term covers different meanings like satire, false tales, factual errors, misleading headlines, and propaganda. Social media plays a crucial role in spreading fake news. It is propagated over different social networking sites and creates confusion, biases, and induces fear among people. Different approaches have been implemented over the years. Despite all the different trials, fake news remains a crucial challenge. The application will analyze the news and classify it into real/fake and clickbait/non-clickbait. In addition to analyzing the news into different categories, it summarizes and includes a dynamic feature that fetches live news.**

Keywords—Fake news, Clickbait, Natural Language Processing, Passive Aggressive Classifier, Naïve Bayes, CountVectorizer, Porter Stemmer.

I. INTRODUCTION

The Internet has made modern life very easy. Transmission and information sharing are possible from anywhere across the globe. People can access surplus information, and this saves both time and energy. They do not have to wade through tons of books or libraries anymore. This has transformed the lives of humans, but at the same time now there is a huge risk of fake and malicious news. Anyone can publish content on the web today irrespective of credibility. Fake news has a large market, as it entices people and uses sensational news headlines to attract attention. People are easily deceived and circulate misinformation quickly. This kind of news causes lasting damage to the minds of people. Social media sites like Twitter, Facebook, and WhatsApp play a significant role in propagating counterfeit information. According to the Economic Times, about 1 in every 2 Indians are victims of misinformation from Facebook and WhatsApp. Scientists believe that Machine Learning and Artificial Intelligence can address the problem of fake news. This is due to the recent influx of classification problems (like image processing, and speech detection) that can be solved using Machine Learning and Artificial Intelligence.

A recent survey conducted in 2019 by Tandoc Jr et al. [1], elaborated on the strength a social media influencer has on their followers, especially detailing on its ability to control children belonging to the age group of 13 to 17 years. A single tweet or post has the ability to completely either demolish or clear-out the stock of an organization, as explained in the article published by Naeem et al. [2]. A post is sensationalized even more if it is found to have originated from a member of a political party; it is automatically considered as unpretentious and earnest by a major proportion of the population. Thus, calling for a need for applications that can help in the segregation of such reports.

However, there fails to exist any system that can perform the task of fake news analyzer, news summarization as well as an interactive system that will display all the hottest news in a single interface. To prevent and control the extent to which such dubious articles are published we have created a system that can aid in making people much more informed about the progress of the world.

From Figure 1 it is evident that over the span of six years a gradual increase in the production of fulsome news is evident. The year 2019 indicated towards the highest number of fake news published by various media platforms especially in Hindi language, the most used language in India. Thus, making it even more important to introduce a system that will help readers formulate correct judgment in whatever they read.

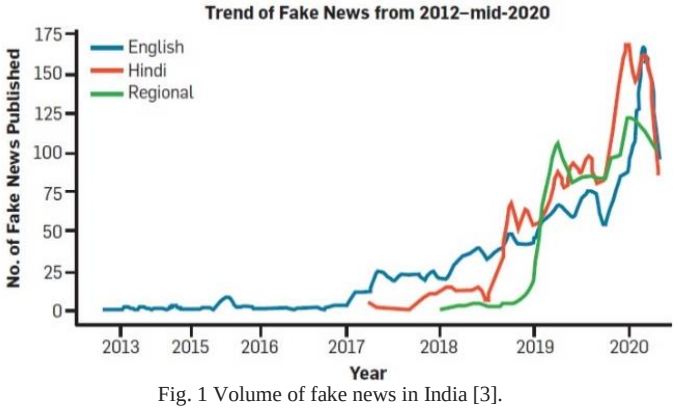

Fig. 1 Volume of fake news in India [3].

II. NEED AND OBJECTIVE

Recent technological advancement has made social media one of the most crucial parts of people's lives. Social media is one of the most famous tools for sharing ideas, information, and news on any topic. There are also daily reports, where data transmission is most essential. This era has become the era of information sharing. But this fast delivery of the news cycle has only aggravated the problem of misinformation. There is a dearth of trustable information and real news sites. To overcome this problem, we have developed an integrated system for various aspects of Natural Language Processing (NLP) to apply Machine Learning techniques to detect fake news and better predict fake articles. Many organizations tend to create news articles with clickbait headlines to increase their revenue. This in turn is inconvenient for users as they are directed to unreliable news sources. To tackle this problem this system classifies the news article into clickbait or non-clickbait. Machine learning text classification has been used as it improves the level of security that is needed in social media daily based networking.

Unlike the earlier times as explained by Veloso et al. [4], when newspapers dominated the market; people nowadays use mobile phones to read the news over the Internet. There are different social media sites and news websites that publish news on the World Wide Web. The question arises about the authenticity of different news articles, especially on social networking sites where there is no check on fake news. Instead of actions being taken on fake news they are further boosted and even go viral on most social networking sites. A news article released by Peter Dizikes [5], discusses how researchers from

the Massachusetts Institute of Technology concluded that on Twitter, fake news travels 6 times faster than real news and also receives 100 times more engagement. Believing in rumors and misinformation has dire consequences for the lives of people. Organized crimes and lynchings are also on the rise due to the lack of credible and real information sources.

The need of today's times is to stop the rumors, especially in developing countries like India, and focus on the correct, authenticated news articles. The main objective is to detect fake news, which is a classic text classification problem and can be solved using Machine Learning and text processing techniques. A model needs to be built that can differentiate between "Real" news and "Fake" news and differentiate between "Clickbait" and "Non-clickbait" news. The proposed system helps to find the authenticity of the news.

Hence, this project aims to develop a Real Time News Analysis System that makes use of Natural Language Processing (NLP) techniques and Machine Learning algorithms for analyzing and classifying the news into fake/real and clickbait/non-clickbait.

III. LITERATURE SURVEY

The ideology of fake news is interlinked with several parameters such as clickbait, misleading title or headlines, satire, fabricated content, sometimes even half-truths are implemented to render a large audience. Such tactics are implemented to either generate profits through streams of click though or even for nefarious activities such as creating political wars among the citizens of a nation.

A study conducted by Castro [6] discussed deeply and thoroughly on how supposedly credible and trustworthy media franchises adopt strategic maneuvers during sensitive times such as elections to sensationalize a political issue typically in a negative fashion. We will be discussing existing projects that in a similar manner propagate malicious and factually incorrect news clipping.

Chen, Honglin et al. [7] conducted a study on the role social media and news channels played in influencing the public during the Covid-19 pandemic in China. They collected several articles through the CNKI (China National Knowledge Infrastructure Database) database. In accordance with their project, they primarily made use of the Jieba toolkit, a Python based module that performed indexing and stop word removal. This mainly consisted of lexical analysis of words, symbols, and numbers in Chinese language. Furthermore, they also implemented a term categorization hierarchical structure that like Thesaurus would aid in extracting indistinguishable words by means of synonyms and similar class categories. This methodology of textual analysis was used as the basis for creating visualization neural networks that helped to create links and relationships of words. They analyzed their newspaper using CNN(Convolutional Neural Networks) and LSTM(Long Short-Term Memory Term).

Thene Grundmann et al. [8] study on climate change over the course of the past decades emphasis on the importance of collecting an appropriate database for accurate evaluation of the data available. According to this research paper, it was strongly recommended to collect large volumes of data, especially from several different references. This was done to ensure that data collected from different sources would provide a wide range of views and ensure that the database is not biased in any manner. Additionally, their system utilized several Linguistic tools for efficient functioning of their system. Various kinds of corpora were integrated in their project. They referred to data collected from conferences, old news articles; social media platforms like Twitter, Facebook, etc., journals were also one of their most trusted document sources. Onto these collected corpora they performed annotation, abstraction using extractive method and finally analysis of the model was performed to evaluate how well their model could classify the data samples into their associated labels correctly [9].

The paper also discussed in detail the correct method of summarization according to the need of the system. Summarization is broadly classified as abstractive and extractive. Abstractive summarization deals with summarization by considering information beyond the scope of what is provided. Here, we deal with inferences and conclusions made from data outside what has been provided. This means that they don't typically tend to be factually accurate in nature. Such a method is useful in scenarios that deal with perhaps survival of literary books or documents of a similar kind.

Extractive summarization on the other hand, is strongly recommended for applications that are very technical in nature. Articles that deal with information regarding the actual ongoing of the world. And any deviation from mentioned data could result in conclusions that are extremely misleading or even bogus in nature. In today's world it has been noted that there is a huge shift from the traditional news perusal activities. Thus, resulting in a scenario where a large majority of our population use social media platforms as a basis for retrieval of news regarding the trends of the world. Around millions of terabytes of data are generated only from Twitter daily. Making it extremely difficult to differentiate fake news from real. As rightly stated by Pennycook Gordon et al. [10] the psychology involving the creation of hoax news is indeed very intricate in nature.

Numerous posts that seem to appear from the so-called genuine websites or sources are in fact fulsome. Technology has made it very easy for us to impersonate others or even publish content that is of an imposturous nature. Additionally, catchy headlines are used, where the actual content largely varies from the published headlines. And since people are unable to dedicate time towards reading an entire article, they end up formulating their opinions only based on the titles such [11]. Tao Jiang, J. Li et al. [12] developed a model that could help in separating fake news from the rest of the reliable news with very high accuracy. They were able to achieve this by combining various Machine Learning and Natural Language Processing methods. During the fitting of their model, they utilized the concepts of tokenization and Term Frequency-Inverse Document Frequency [13][14]. This helped them achieve an accurate representation of all the linguistic terms present in their database. Additionally, they merged this elaborate pre-processing method with Multimodal Naïve Bayes, Passive Aggressive Classifier and Neural Networks. Thus, combining the algorithms used by several individual systems into a single system [13]. All of which gave the model to learn and identify fake news from a large volume of available data.

The aforementioned pre-processing gave them the ability to segregate clickbait from the rest of the news by identifying certain features. New articles that tend to be redundant with numbers or flashy words such as shocking, unbelievable, astonishing, and so on. This in turn helped them achieve a

steeping accuracy of 93%.

IV. PROPOSED SYSTEM

Figure 2 explains the general workflow of the system. This system architecture is a visual representation of all processes in sequential order. The process starts when the user enters the URL of any article and then the system will process the URL by applying various NLP techniques like stop word removal, clitics removal, stemming and word tagging. Web scraping is used to fetch the data and after applying feature extraction a summarized view of the article in form of title and headline is generated. After a summarized view is generated, Machine Learning algorithms such as Naive Bayes and Passive Aggressive Classifier are used to classify the news article into real/fake and clickbait/non-clickbait.

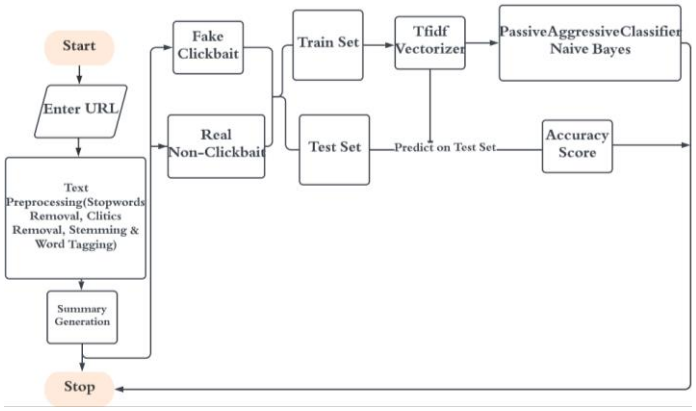

Fig 2 Internal Working of system

V. IMPLEMENTATION

A. Data Analysis

1) Clickbait and Non-Clickbait

The Clickbait data has been collected from various sources like Buzzfeed, ViralNova, and PoliticianInsider to name a few; and was compiled into a single data frame. Whereas, Non-Clickbait data was extracted from NY Times, The Guardian, Bloomberg, Reuters and so on, and was loaded separately, as evident in Table 1.

TABLE I: Datasets Extracted for Analysis

Sr no	Datasets		
	Sources	***Dimension***	***Description***
1.	Clickbait and Non clickbait sources	52,000 headlines	The headlines dated from 2007-2020. Various clickbait sources used in dataset are Buzzfeed, Upworthy, ThatScoop, ViralStories. And non-clickbait sources are from NY times, the Washington post, the Guardian Boomberg and Reuters.
2.	Kaggle Dataset	30,000 data	The data was collected from 2007-2016
3.	Twitter APIs and online publications	22,000 headlines	Headlines were web scraped from Internet.

After the extraction of data from various sources Pre-Processing operations was performed. This included the conversion of words to lowercase, removal of punctuation, stop words, and other symbols. The main aim was to introduce uniformity.

Exploratory Data Analysis was also performed, and this includes Tokenization of the data. It is the process of tokenizing or splitting string or text into a list of tokens. It is used in Natural Language Processing to split paragraphs and sentences into smaller units. This makes it easier for the system to assign meaning to the terms. And this in turn helps the model to accurately classify Clickbait terms with that of non-Clickbait.

The final step was the modeling of data. The dataset was divided into train-test split in the ratio 80:20. Term-Frequency Inverse Document Frequency (TF-IDF) is applied to the headline to determine its relevance. TF-IDF was used to assign numbers to text so that they are represented in a meaningful manner. It was observed that most of the stopwords were assigned the value zero, whereas terms that had more impact in terms of information they provided were given values ranging closer to 1. Thus, the model assigned more significance to Non-Clickbait terms in contrary to fake or misleading words.

We analyzed the class distribution of various engineering features and their relevance on each class. On average, it is observed that Clickbait headlines are slightly longer than non-Clickbait headlines.

2) Real and Fake

The very first step is to load the data onto which processing, and analysis is to be performed. For Real or fake classification Train.csv is loaded. This dataset has 25 thousand records.

The concept of Regular Expression was then applied to identify certain patterns in the given dataset and get a better understanding of the data.

The NLTK (Natural Language Toolkit) library was imported to apply the concepts of stopword removal and stemming, as explained by Loper et al. [17]. Stopwords are terms that most frequently occur across the span of the corpus, as the frequency of terms increases the amount of usefulness or information capabilities significantly decreases. Thus, making this removal stage particularly important. Stemming on the other hand, is the extraction of the stem or root word from a given term. This increases the discoverability of a word. The implementation of this stage was done by using Porter Stemmer algorithm, where the suffix of the word is eliminated.

B. User Interface

The very first page is the home page of the website with an easy to navigate navigation bar. The navigation bar has different options: Home, About and Light/Dark mode.

To increase the ease with which the end-user interacts with the given system, that is usability an additional functionality of Dark mode has been provided. The system has both dark and light modes to be used according to user preference.

There is an option to enter the URL of a news article to classify it into Real or Rake and Clickbait or Non-Clickbait, Figure 3. This is the stage wherein the user is directly interacting with the system.

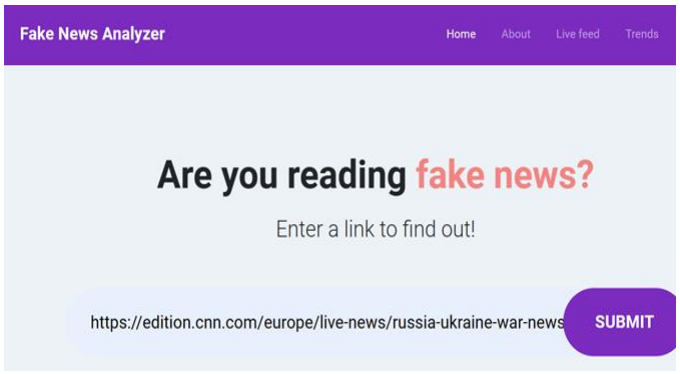

Fig 3 User URL input

Figure 4 shows the result of a news article whose URL was entered. The system makes use of NLP techniques to fetch and summarize the article and uses Machine Learning Algorithms to classify it into real and non-clickbait. The summarization method that has been used is extractive summarization rather than abstractive summarization. As stated by several researchers [18][19][20], extractive summarization is preferred as, in the case of abstractive summarization, inferences are made beyond the scope of the information provided to it. Making it ineffectual in scenarios where information provided to users is only factual in nature, as it is needed by the system.

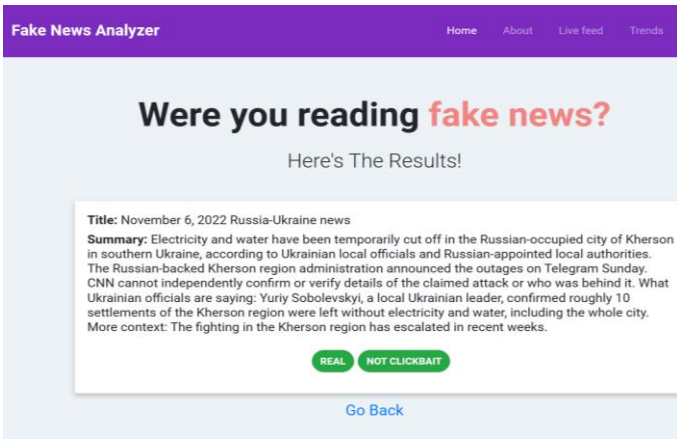

Fig 4 News Classification as Real and Non-Clickbait

We have also included the About page of the system which gives brief information about the need and use of the News Analyzer application.

To make this application more interactive in nature, a page titled the 'Live Feed' has been included, as visible in figure 5. This provides the latest news in real time along with labels 'Fake or Real' and 'Clickbait and Non-Clickbait' so that the users are well aware of the credibility of the news they are pursuing through. Live news feed works in the same concept as Google news, it will update the news daily as articles are fetched from News API and is thus labeled with the aforementioned tags.

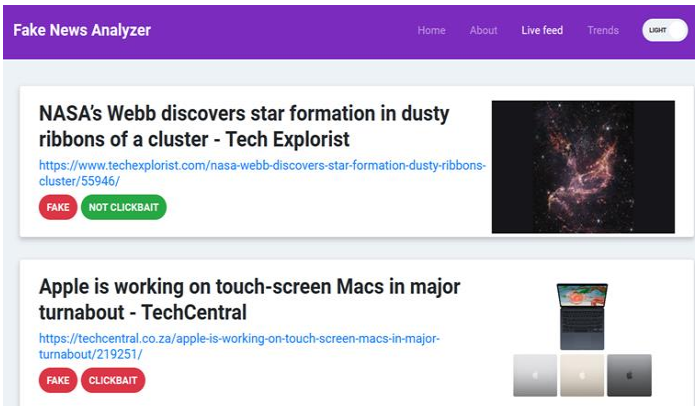

Fig 5 Live Feed Page of system

Figure 6 is the Trends Page that contains the statistics of our website, wherein Google charts drawing feature in JavaScript is used to display the results in the form of a pie diagram. The displayed results have the classified outputs that will be saved in the MonogoDB database. The database stores the count of the URL submissions based on the results like 'Fake' or 'Real' and uses this count to display it in the form of 3D charts, in the form of the number of real articles compared to fake and number of clickbait articles compared to non-clickbait. This feature makes the system entirely dynamic as it will update on a daily basis and will store the count of results in MonogoDB database so that statistics are accordingly updated.

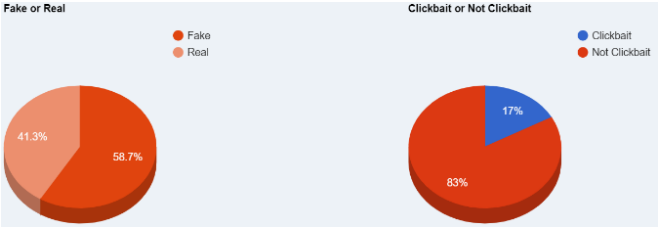

Fig 6 Trends Page of system

VI. Results And Discussions

For the Clickbait and Non-Clickbait model we have created a word cloud. There exist two-word clouds each of which corresponds to the tendency occurrences of certain words and numbers that may result in a Clickbait headline. The size of the word is proportional to the frequency of that word. Larger size of word is a direct indication of its increased usage across the article [21].

Figure 7 depicts that news articles that tend to have numerical values or words like "things" and "people" are most likely to be a Clickbait article.

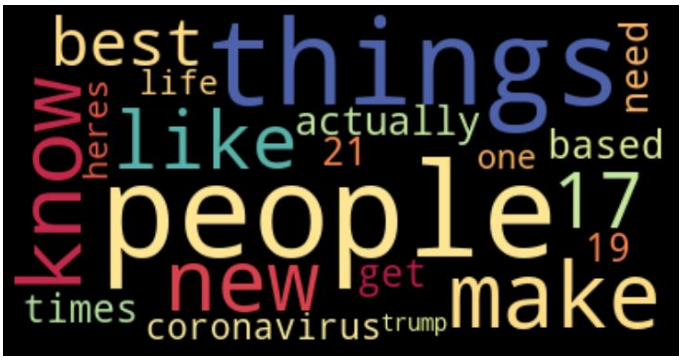

Fig 7 Clickbait Word cloud

Similar is the case with non-clickbait articles as well. News reports that have terms such as “us”, “new” or “police” tend to be Non-Clickbait in nature.

We have visually represented the relation of certain stopwords with their associated frequency. These stopwords are the Clickbait terms that have been identified through analysis performed on the database. As it is obvious from the figure that the words that are shorter in length are most likely to be associated with Clickbait news. There are governing and dependent words, and it is observed that Clickbait words have more similarity distance. Thereby, this results in more dependency in case of Clickbait words as compared to Non-Clickbait words.

For instance, a news article published by moneycontrol on ‘Trade setup for Thursday: Top 15 things to know before the opening bell’ was labelled as Fake and Clickbait. Proving the efficiency of our system, as it is evident that the news has several occurrences of the word ‘things’ and is also redundant with numerous numerical values in figure 8.

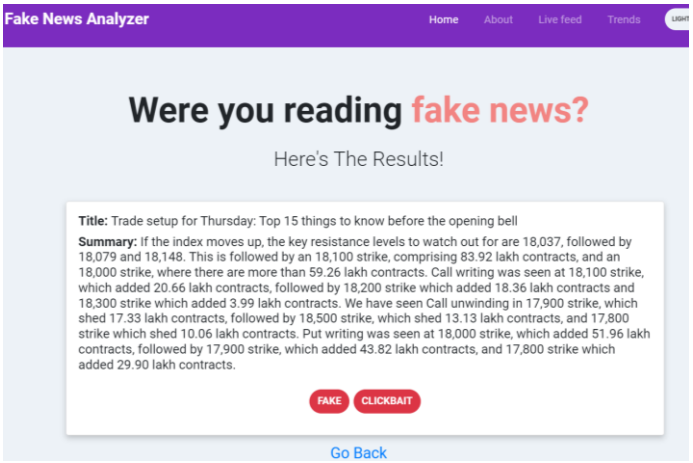

Fig 8 Demonstration of Clickbait news

A Confusion matrix for Clickbait and Non-clickbait model was picturized. We achieved an accuracy of 93% by the Naive Bayes. Additionally, we had a Recall of 94%. On contrary, Multinominal Naïve Bayes algorithm which was also used has significantly lower accuracy at 81.8%. The pivotal reason being that Multinominal Naïve Bayes focuses on terms related to a given instance. For example, as demonstrated in figure 20 this algorithm is most likely to give more weightage to words related to finance even if the article may contain equally important and Non-Clickbait terms related to domain.

Similarly, the Confusion matrix of the Real or Fake model was also developed. Here, we had an accuracy of 91.9%. The primary factor that resulted in such high accuracy is the fact that Cross Validation was performed in all the models. Here, we have evaluated the accuracy of each of our various models and compared them, to identify the best classifier.

VII. Conclusion

Fake news sharing is one of the popular research problems in recent technology based on lack of security and trust in terms of the truth of shared news in social media. Natural Language Processing has revolutionized different sectors and has a tangible impact on the detection of fake news. However, most existing systems still have issues and fail to meet user expectations. This system was designed to predict if a news article or a headline is Fake or Real and Clickbait or Non-Clickbait using NLP techniques and Machine Learning. For the purpose of hyperparameter tuning in Clickbait or Non-Clickbait news, algorithms such as Naive Bayes, SVM(Support Vector Machine) and Logistic Regression were used. Here an accuracy ranging between 90 and 93% was observed. To minimize the false negative, we have prioritized recall. Thus, making Naive Bayes most suitable for the application with an accuracy of 93% and recall of 94%. On the other hand, for classification of Real or Fake news Multinomial Naive Bayes and Passive Aggressive classifiers were used, accounting for an accuracy of 88% and 91.9% respectively. Since Passive Aggressive classifier is an online learning algorithm, we are training the model through gradual feeding of data in a sequential manner. Hence, it is most suited for the Real and Fake news.

References

[1] Tandoc Jr, Edson C. "The facts of fake news: A research review." Sociology Compass 13, no. 9 (2019): e12724

[2] Naeem, Salman Bin, Rubina Bhatti, and Aqsa Khan. "An exploration of how fake news is taking over social media and putting public health at risk." Health Information & Libraries Journal 38, no. 2 (2021): 143-149.

[3] Shivangi Singhal, Rishabh Kaushal, Rajiv Ratn Shah, Ponnurangam Kumaraguru, ”Fake News in India: Scale, Diversity, Solution, and Opportunities”, Communications of the ACM, November 2022, Vol. 65 No. 11, Pages 80-81 10.1145/3550493.

[4] Veloso, Bráulio M., Renato M. Assunção, Anderson A. Ferreira, and Nivio Ziviani. "In Search of a Stochastic Model for the E-News Reader." ACM Transactions on Knowledge Discovery from Data (TKDD) 13, no. 6 (2019): 1-27.

[5] Peter Dizikes, “Study: On Twitter, false news travels faster than true stories”, MIT News Office, March 8, 2018.

[6] Castro, Laia, Jesper Strömbäck, Frank Esser, Peter Van Aelst, Claes de Vreese, Toril Aalberg, Ana S. Cardenal et al. "Navigating high-choice European political information environments: A comparative analysis of news user profiles and political knowledge." The International Journal of Press/Politics 27, no. 4 (2022): 827-859.

[7] Grundmann, Reiner. "Using large text news archives for the analysis of climate change discourse: some methodological observations." Journal of Risk Research 25, no. 3 (2022): 395-406.

[8] Chen, Honglin, Xia Huang, and Zhiyong Li. "A content analysis of Chinese news coverage on COVID-19 and tourism." Current Issues in Tourism 25, no. 2 (2022): 198-205.

[9] Wu, Jeff, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. "Recursively summarizing books with human feedback." arXiv preprint arXiv:2109.10862 (2021).

[10] Pennycook, Gordon, and David G. Rand. "The psychology of fake news." Trends in cognitive sciences 25, no. 5 (2021): 388-402.

[11] Bryanov, Kirill, and Victoria Vziatysheva. "Determinants of individuals’ belief in fake news: A scoping review determinants of belief in fake news." PLoS one 16, no. 6 (2021): e0253717.

[12] Jiang, T. A. O., Jian Ping Li, Amin Ul Haq, Abdus Saboor, and Amjad Ali. "A novel stacking approach for accurate detection of fake news." IEEE Access 9 (2021): 22626-22639.

[13] Azam, Nouman, and JingTao Yao. "Comparison of term frequency and document frequency-based feature selection metrics in text categorization." Expert Systems with Applications 39, no. 5 (2012): 4760-4768.

[14] Wu, Ho Chung, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. "Interpreting tf-idf term weights as making relevance decisions." ACM Transactions on Information Systems (TOIS) 26, no. 3 (2008): 1-37.

[15] Akande, O., G. Egwuonwu, and W. Ajayi. "COVID-19 Fake News Detection Using Naïve Bayes Classifier."

[16] Gupta, Saloni, and Priyanka Meel. "Fake news detection using passive-aggressive classifier." In Inventive Communication and Computational Technologies, pp. 155-164. Springer, Singapore, 2021

[17] Loper, Edward, and Steven Bird. "Nltk: The natural language toolkit." arXiv preprint cs/0205028 (2002).

[18] JUGRAN, SWARANJALI, ASHISH KUMAR, BHUPENDRA SINGH TYAGI, and VIVEK ANAND. "Extractive automatic text summarization using SpaCy in Python & NLP." In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 582-585. IEEE, 2021.

[19] Awasthi, Ishitva, Kuntal Gupta, Prabjot Singh Bhogal, Sahejpreet Singh Anand, and Piyush Kumar Soni. "Natural Language Processing (NLP) Based Text Summarization-ASurvey." In 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 1310-. IEEE, 2021.

[20] Alomari, Ayham, Norisma Idris, Aznul Qalid Md Sabri, and Izzat Alsmadi. "Deep reinforcement and transfer learning for abstractive text summarization: A review." Computer Speech & Language 71 (2022): 101276.

[21] Lohmann, Steffen, Florian Heimerl, Fabian Bopp, Michael Burch, and Thomas Ertl. "Concentri cloud: Word cloud visualization for multiple text documents." In 2015 19th International Conference on Information Visualisation, pp. 114-120. IEEE, 2015.