IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

FAKE NEWS INDENTIFIER THROUGH MACHINE LEARNING

[1] Mrs. Chetana Patil, [2]Jigyas Subedi, [3]Sujan Bhattarai, [4]Bhoomika L, [5]Chandan Indalkar,

[1]Asst. Professor, Computer Science and Engineering
[1]Impact College of Engineering and Applied Sciences, Bengaluru, India.

Abstract: Technology is changing the way that we live our lives. With the advances in technology social media attains a lot of attention of the people around us. With this fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. Users can get infected easily by this online fake news, which has brought about tremendous effects on the society. An important goal is to stop the spread of rumors and focus on the correct, authenticated news articles. This paper focuses on techniques such as natural language processing (NLP) techniques for text analytics and train deep learning models for detecting fake news based on news title or news content. Using this technique, text pre-processing such as regular expression, tokenization, lemmatization and stop words removal are used before vectorizing them into N-gram vectors or sequence vectors using terms frequency inverse document frequency (TF-IDF) or one-hot encoding respectively. With the help of Machine learning and natural language processing, author tried to aggregate the news and later determine whether the news is real or fake using deep neural networks. Experimental evaluation yields the best performance using Term Frequency-Inverted Document Frequency (TF-IDF) as feature extraction technique.

IndexTerms **- Fake News, social media, Natural Language Processing, Deep, Neural Network**

I.INTRODUCTION

The phenomenon of fake news is a significant problem in today's digital world. It refers to the spread of news, data, reports, and information that is wholly or partially false. The impact of fake news is not limited to individuals but also affects society as a whole. It is like a wildfire that can influence many people every day. Fake news poses a threat to a country's security, economy, prosperity, and even to individuals. Often people are not aware of how fake news can impact the matter surrounding them or how to handle it when it happens.

With billions of articles created every day on the internet, people can unintentionally become a part of the problem by spreading fake news without realizing it. A seemingly simple action of sharing a news article can become a serious issue if there is no control gate to prevent the spread of fake news stories. Therefore, it is crucial to develop and implement methods to detect and combat fake news effectively. By doing so, we can help ensure that the news and information we receive are accurate and reliable, ultimately contributing to a more informed and responsible society.

Problem Definition

Most of the smart phone users prefer to read the news via social media over internet. Lots of Falsified news is being circulated on social media. The need of an hour is to stop the rumors and focus on the correct, authenticated news articles. The question is how to authenticate the difference. The main objective with this project is to classify the news into fake or real. With the help of Deep Learning and Neural Networks we can classify it.

Motivation

In Today's world, anybody can post the content over the internet. It is harmful for the society to believe on the rumors and pretend to be a news. The fake news can lead to mob lynching, riots and negative or wrong information. Fake news detection is made to stop the rumors that are being spread through the various plat- forms whether it be social media or messaging platforms Fake news tries to stop such false information thereby protecting the society from this kind of violence. Satirical Cues are prevalent in false news they can help in detection.

Goals and Objective

Most of the smart phone users prefer to read the news via social media over internet. Lots of Falsified news is being circulated on social media. The need of an hour is to stop the rumors and focus on the correct, authenticated news articles. The question is how to authenticate the difference. The main objective with this project is to classify the news into fake or real. With the help of Deep Learning and Neural Networks we can classify it.

II.NEED OF THE STUDY.

With the advances in technology social media attains a lot of attention of the people around us. With this fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. Users can get infected easily by this online fake news, which has brought about tremendous effects on the society as well as nation. An important goal is to stop the spread of rumors and focus on the correct, authenticated news articles.

III.RESEARCH METHODOLOGY

The methodology section outline the plan and method that how the study is conducted. This includes Universe of the study, sample of the study, Data and Sources of Data, study's variables and analytical framework. The detail share as follows;

3.1 Population and Sample

For a fake news detection, population is considered as the total number of news articles available on the internet or a specific news platform. However, since analyzing the entire population of news articles is not feasible, a sample is selected for analysis. The sample should be representative of the population and should be large enough to provide accurate results. We collected a sample of news articles from various news platforms in India, including both real and fake news articles. The sample size may depend on factors such as the availability of data and computational resources. And then we use this sample to train and test their machine learning model for fake news detection.

3.2 Data and Sources of Data

In this study, the proposed system aims to detect fake news in the Indian context by generating its own dataset, since existing datasets are predominantly based on Western media news. We are collecting data by scraping prominent media houses and notoriously known fake news sites in India. By capturing the unique nuances of Indian news reporting, including the use of local languages and cultural references, the proposed system can address the significant difference in reporting styles between Western and Indian media, which is crucial for enhancing the accuracy of fake news detection in India. This approach is expected to improve the performance of the proposed system compared to existing implementations, and contribute to the ongoing effort to tackle fake news in India.

3.3 Theoretical framework

Machine learning is a subfield of artificial intelligence that involves the use of statistical techniques to enable computer systems to learn and improve performance on a specific task without being explicitly programmed. One application of machine learning is in fake news detection, where algorithms are trained to identify fake or misleading news articles.

Machine learning models can be trained on a dataset of known fake and real news articles to learn the patterns and features that distinguish between them. Common techniques used in machine learning for fake news detection include natural language processing (NLP) and deep learning.

On the other hand, Fake news refers to false or misleading information presented as if it were true, often spread through social media and other digital platforms. The spread of fake news has become a major concern, as it can impact public opinion and even influence political outcomes. Machine learning has emerged as a promising solution for detecting fake news. This is because machine learning algorithms can learn to identify patterns in data and make predictions based on those patterns. In the case of fake news detection, machine learning algorithms can be trained on large datasets of news articles and other content to learn the characteristics of fake news.

Some of the key features that machine learning algorithms can be trained to recognize in fake news include sensationalist language, lack of reliable sources, and inconsistencies in the story. Machine learning can also analyze the social media networks where fake news is often spread to identify patterns of behavior that suggest a story is fake. There are several different approaches to using machine learning for fake news detection. One common approach is to use supervised learning algorithms, where the algorithm is trained on a labeled dataset of fake and real news articles. The algorithm can then be used to classify new articles as either fake or real. Another approach is to use unsupervised learning algorithms, where the algorithm is trained on an unlabeled dataset of news articles and uses clustering and other techniques to identify patterns that are indicative of fake news.

Overview of project module:

i Dataset Scraping

•The Core difference between existing implementations and proposed system is that existing systems are trained purely on Kaggle Datasets which are predominantly based on news by Western Media.

•There is a stark difference between the style of reporting in case of Western Media and Indian Media.

•Therefore, as the first module of our system, we are generating our own dataset by scrapping the prominent media houses and notoriously known fake news sites.

ii Dataset Filtration Module

•Can also be called as Data Preprocessing Module.

•Dataset scrapped is raw data and not necessarily clean in nature, therefore our deep learning system will not be able to interpret anything from it.

•Therefore, this module of the project has the objective of making the raw data into proper data which can be readily fed into the DL Model.

•Based on our literature review, we have identified the NLP Operations required to be performed on the raw data and in this phase, we will be implementing that.

iii Neural Network Module

•Once we have processed the dataset, we can pass that to our model. In this module, we will develop the Deep Learning System as proposed by referring the literature survey.
•From our survey, we know that we can get better results, by using both the news title and news content.
•N-Gram Vectorization and Sequence Vectors are two pro- posed methods which will be tested and the better one will be selected.

iv Frontend Module
•Users cannot directly interact with the DL Model.
•Though theoretically it is possible but not user-friendly to do so.
•Therefore, we need a User Interface so that users can use the developed model easily, without any hassle.
•The UI would accept the news title and news content from user and display the predicted output to the user.

v Integration Module
•Module 2 and 3 can be considered as backend and Module 4 as frontend.
•However, React Frontend cannot directly interact with the Tensorflow backend.
•Therefore, we need a separate module responsible for all communications between the frontend and backend.
•This module will pass the obtained input to the deep learning model and then send the processed output back to the frontend.
•Since the backend is predominantly Python, Flask is a viable option for this module.

3.4ALGORITHMS

3.4.1 TF-IDF

In the TF-IDF method, a document term matrix is generated and each column represents a single unique word. The difference in the TF-IDF method is that each cell doesn't indicate the term frequency, but the cell value represents a weighting that highlights the importance of that particular word to the document. TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a statistical measure that is widely used in information retrieval and natural language processing to evaluate the relevance of a term in a document. TF-IDF assigns a weight to each term in a document based on how frequently it appears in the document and how rare it is across the entire collection of documents. The goal of TF-IDF is to identify terms that are both frequent in a document and unique to that document, as these are the terms that are most likely to capture the essence of the document's content. TF-IDF is commonly used in machine learning algorithms for tasks such as text classification, information retrieval, and sentiment analysis.The formula to calculate TF-IDF is:

$$W_{x,y} = tf_{x,y^*} log\left(\frac{N}{df_x}\right)$$

$W_{x,y}$= Word x within document y

$tf_{x,y}$= frequency of x in y

df_x= number of documents containing x

N=total number of documents

3.4.2 N-gram Vectorization

N-gram vectorization is a technique used to represent text data in a numerical format that can be used by machine learning algorithms. It involves breaking down the text into sequences of contiguous words, called n-grams, and representing them as vectors. For example, let's say we have the sentence "The quick brown fox jumps over the lazy dog." To create 2-gram vectors, we would first break the sentence down into pairs of adjacent words, like this: "The quick", "quick brown", "brown fox", "fox jumps", "jumps over", "over the", "the lazy", and "lazy dog." Each of these pairs is then represented as a vector, where the value of each element in the vector corresponds to the frequency of that n-gram in the text.
So, for the sentence above, the 2-gram vector representation would be:
[the quick] = 1
[quick brown] = 1
[brown fox] = 1
[fox jumps] = 1
[jumps over] = 1
[over the] = 1
[the lazy] = 1
[lazy dog] = 1

This vector representation can then be used as input for machine learning algorithms, such as those used for fake news detection. By comparing the n-gram vectors of a given text to those of known fake or real news articles, the algorithm can classify the input text as either fake or real.be near zero).

3.4.3 Deep Learning Model

With all the pre-processed new titles and content in vectors form, Keras neural network models with some dense layers are built and trained using TensorFlow framework to perform classification task of detecting the fake news. In the Keras neural network model, the layers 1, 3 and 5 are using RELU as the activation function with 32 nodes, 16 nodes and 8 nodes respectively.
In between these 3 layers, 2 dropout layers of 40% dropout rate are also added to the network. The purpose of adding dropout layers is used to avoid from overfitting by dropping some node and therefore, generalize better.

Then, the last layer of the Keras neural network model, also known as output layer, is added with sigmoid activation function so to ensure the output of the network is in binary format because there are only 2 possible outcomes from this project such that "0" refers to real news and "1" refers to fake news.
The most commonly used machine learning classification algorithms such as Naive Bayes , Random forest, SVM, Decision trees are not able to perform well and are not able to get at least 90% of accuracy. But here the Keras NN model with all the preprocessed media titles and/or media content in vector format, achieves that result and thus, Keras NN can be used for detecting fake news with some dense layers.
In the proposed model the Keras is used and the layers 1,3,5 are Dense Layers with 64 nodes, 16 nodes and 2 nodes respectively and Rectified Linear Unit as the activation function. There is addition of two dropout layers of 40% dropout rate to the proposed neural model. This 40% dropout is done to avoid the overfitting of the model and to be able to generalize the dataset better as experimentally we observed that 20% dropout was not giving optimal results.

3.5 System Implementation plan

For the implementation plan of our proposed project on fake news detection, we will follow a phased approach. The first phase will involve the data collection process, where we will scrape data from prominent media houses and fake news sites in India. We will then clean and preprocess the data to ensure that it is usable for training our machine learning models.

In the second phase, we will implement the machine learning algorithms to train the models to detect fake news. We will explore different algorithms, including logistic regression, random forest, and support vector machines, to identify the best approach for our project. We will also experiment with different vectorization techniques, such as TF-IDF and N-gram vectorization, to identify the best approach for our dataset.

The third phase will involve testing and evaluation of our models. We will split our data into training and testing sets to evaluate the performance of our models. We will use metrics such as accuracy, precision, recall, and F1-score to evaluate the performance of our models.

In the final phase, we will deploy our models to a web-based application that will enable users to input news articles and get real-time feedback on the authenticity of the news. We will also develop a user interface that is user-friendly and easy to navigate. We will ensure that our application is scalable and can handle a large volume of users simultaneously. We will also ensure that our application is secure and protected against malicious attacks.

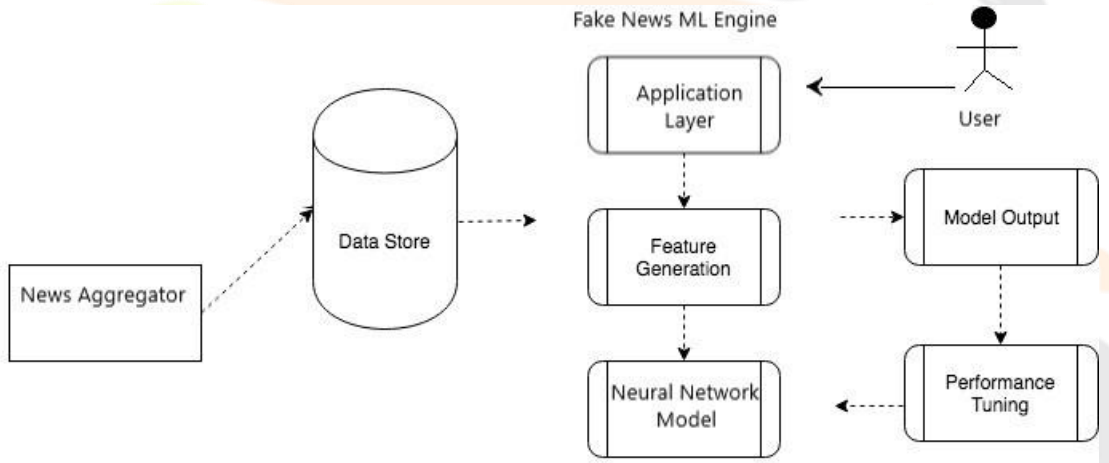

Figure1: System Implementation

3.5.1 Data Collection and Cleaning

With the help of Beautiful-Soup and Scrappy, we managed to build a novel dataset consisting of more than 10000 articles with a balance between no. of fake articles and no. of true articles and then performed Data Cleaning on them to remove non-unicode characters and other irrelevant data.

Head	Body	Label
Four more arrested in Rinku Sharma murder case	Delhi Police on Sunday arrested four more per...	0
Unnao case | Police adds poisoning charge; sa...	The police on Saturday added a charge relate...	0
The man who follows India's rivers on foot	Siddharth Agarwal has been following India's ...	0
Realtor takes tips from movie to burgle frien...	A realtor, who had sustained huge losses, all...	0
Voting underway for polls to 6 municipal corp...	Elections to six municipal corporations in G...	0
...	...	...
Bollywood Celebrities Mistakenly Take Covid Va...	The second wave of COVID-19 seems deadlier tha...	1
WHO Successfully Convinces Corona Virus To Ent...	In a major victory against the coronavirus, th...	1
Government Likely To Request Punjab Youth To A...	WorldÃ¢Â□Â□s largest COVID-19 Vaccine drive by...	1
Brown Mom Gets Four COVID-19 Vaccines Packed A...	India started the worldÃ¢Â□Â□s biggest Covid-1...	1
Rahul Gandhi Too Profound To Be Understood By ...	After failing to make people understand that R...	1

Figure2: Sample of Scrapped Data

3.5.2 Data Exploration

In the data exploration phase of the project, the cleaned data was used to generate word clouds for both real and fake news data. The word counts were calculated from a pool of 161,225 unique words for real news and 150,501 unique words for fake news. By analyzing these word counts, the top 20 words were identified that contribute significantly to the fakeness or realness of an article. This analysis helped to gain insights into the language used in real and fake news articles, and to identify the key features that could be used to train a machine learning model for fake news detection.

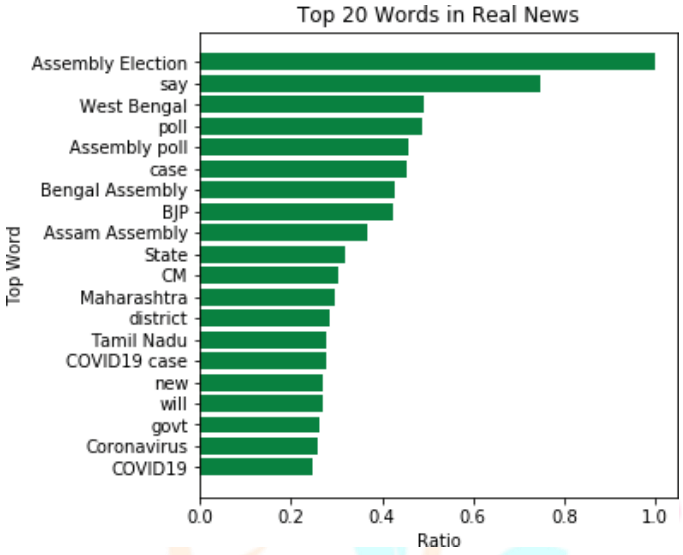

Figure3: Word Count of Real News Data

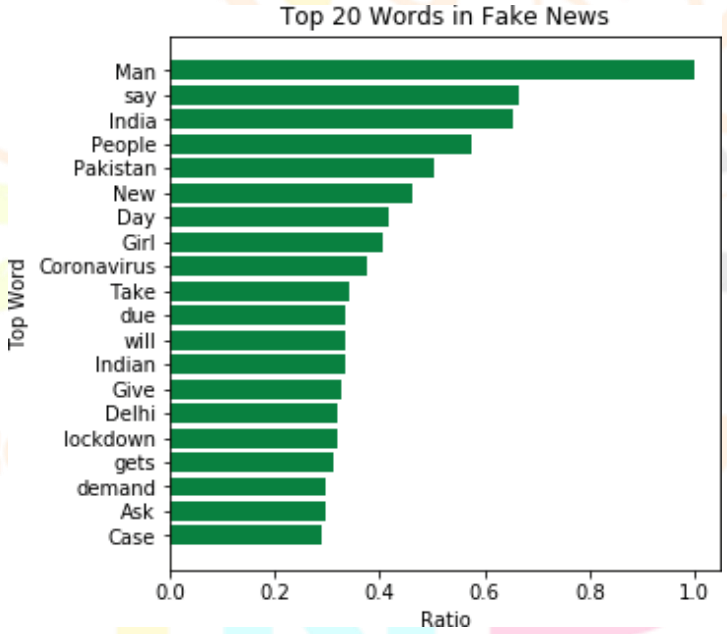

Figure4: Word Count of Fake News Data

3.5.3 Data Pre-processing

Data pre-processing is an essential step in any machine learning project, including fake news detection. In this step, the raw data is transformed and cleaned to prepare it for analysis. The data pre-processing step for fake news detection involves several tasks, such as removing irrelevant information, handling missing values, and converting the data into a format that can be analyzed by machine learning algorithms.

One common pre-processing task for fake news detection is text cleaning. This involves removing stop words, punctuation, and other non-essential information from the text. Additionally, text normalization techniques like stemming or lemmatization can be applied to reduce words to their base form and avoid duplicating similar words.

Another pre-processing task is vectorization, which involves converting text data into numerical format that machine learning algorithms can understand. Techniques like TF-IDF or N-gram vectorization can be applied to represent the text data in a way that preserves the meaning and context of the words.

Overall, data pre-processing is an important step to ensure that the data is ready for analysis and that the machine learning algorithms can accurately detect fake news. In this project we use both vectorization and NLP techniques in order to get the clean and accurate data.

3.5.4 Model Training

We then trained using these matrices on our constructed Neural Network architecture as discussed in the preceding section. Training Output is shown in Figure below.

As observed, the developed architecture initially has a mediocre performance with obtaining around 60% accuracy and with successive epochs, the performance improves to finally end up with a 97.5% accuracy with 98.2% F-1 score.

```
X_tfidf_train_title

array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

Figure5: Matrix obtained post N-Gram Vectorization

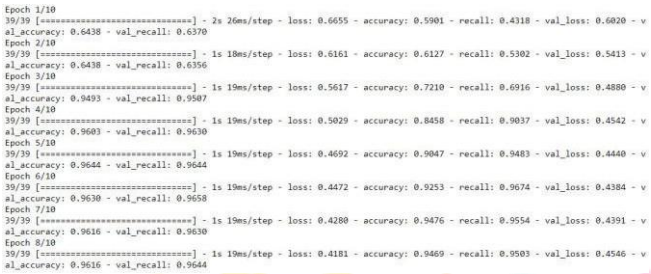

Figure6: Example Output of Training Model

IV.RESULTS AND DISCUSSION

4.1 Confusion Matrix

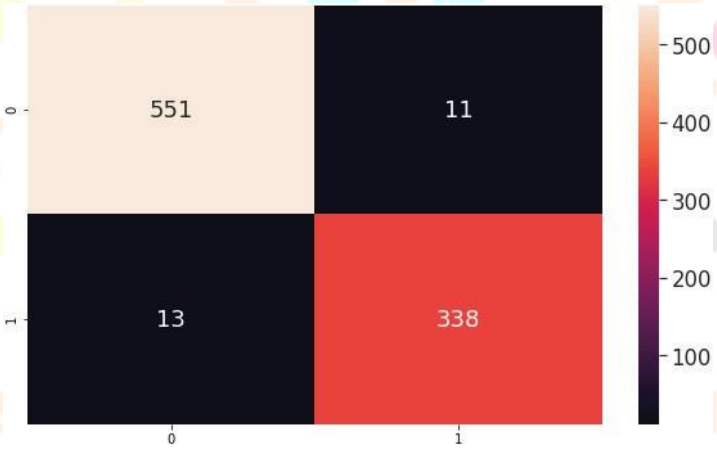

Figure7: Confusion Matrix

In the context of fake news detection, the confusion matrix measures the number of correctly classified fake and real news articles by a model. The true positives (TP) represent the number of fake news articles that were correctly classified as fake. The true negatives (TN) represent the number of real news articles that were correctly classified as real. The false positives (FP) represent the number of real news articles that were incorrectly classified as fake. The false negatives (FN) represent the number of fake news articles that were incorrectly classified as real. Here, the confusion matrix shows that the model has performed well with only 24 misclassifications out of 913 articles. This means that the model correctly classified most of the articles as either fake or real. Such a good performance on unseen data is a positive indication of the model's effectiveness in detecting fake news. The confusion matrix can also be used to calculate other performance metrics such as precision, recall, and F1 score, which provide a more detailed evaluation of the model's performance.

4.2 Model Evaluation

We also trained using same data on other ML Algorithms for comparison purposes. The results for the same are tabulate in the figure below:

Table1: Model Evaluation

No	Algorithm	Accuracy	F-1 Score
1.	NLP+SVM	91.09	94.7
2.	NLP+DT	93.8	95.6
3.	NLP+LR	89.67	91.24
4.	**NLP+NN**	**97.5**	**98.2**

As seen, we have obtained way better results using the deep learning architecture implemented in this paper. Thus, we conclude that our model performs better than existing models on novel dataset.

4.3 Result

This shows the implementation of a fake news detection model in a frontend application. The frontend has a homepage, which is represented by figure8. Once the user inputs the news article, the backend model processes it and returns a prediction result. There are two possible prediction results, either true or false. Figure9 represents the prediction result when the news article is true, while figure10 represents the prediction result when the news article is false.

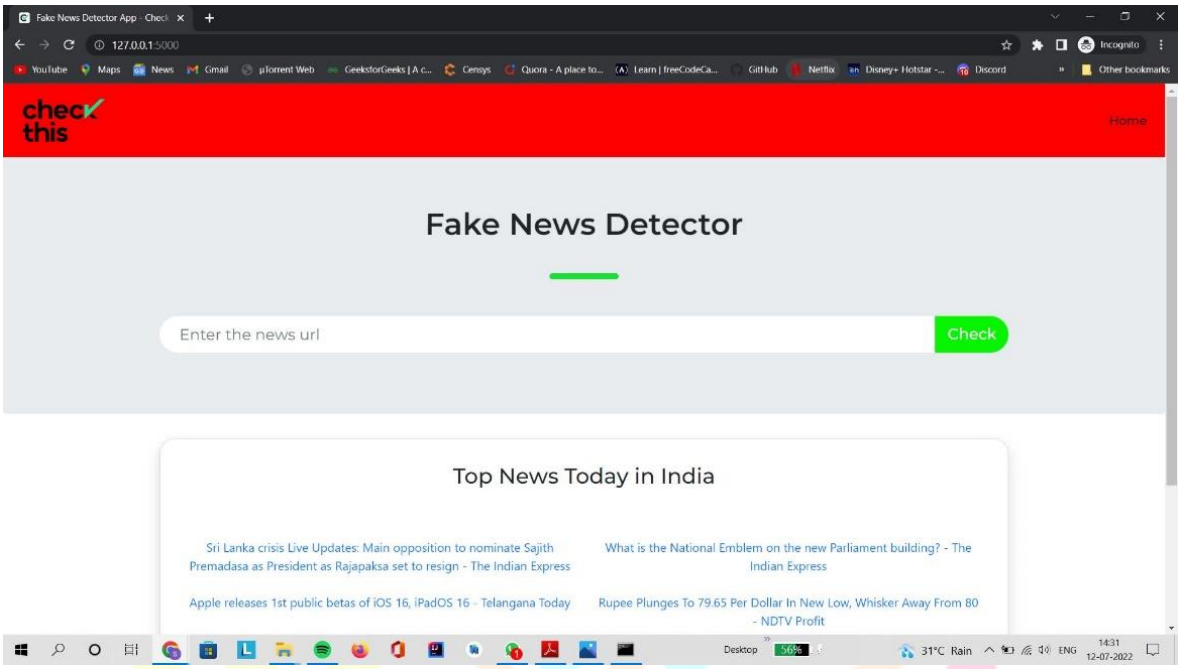

Figure8: Home page

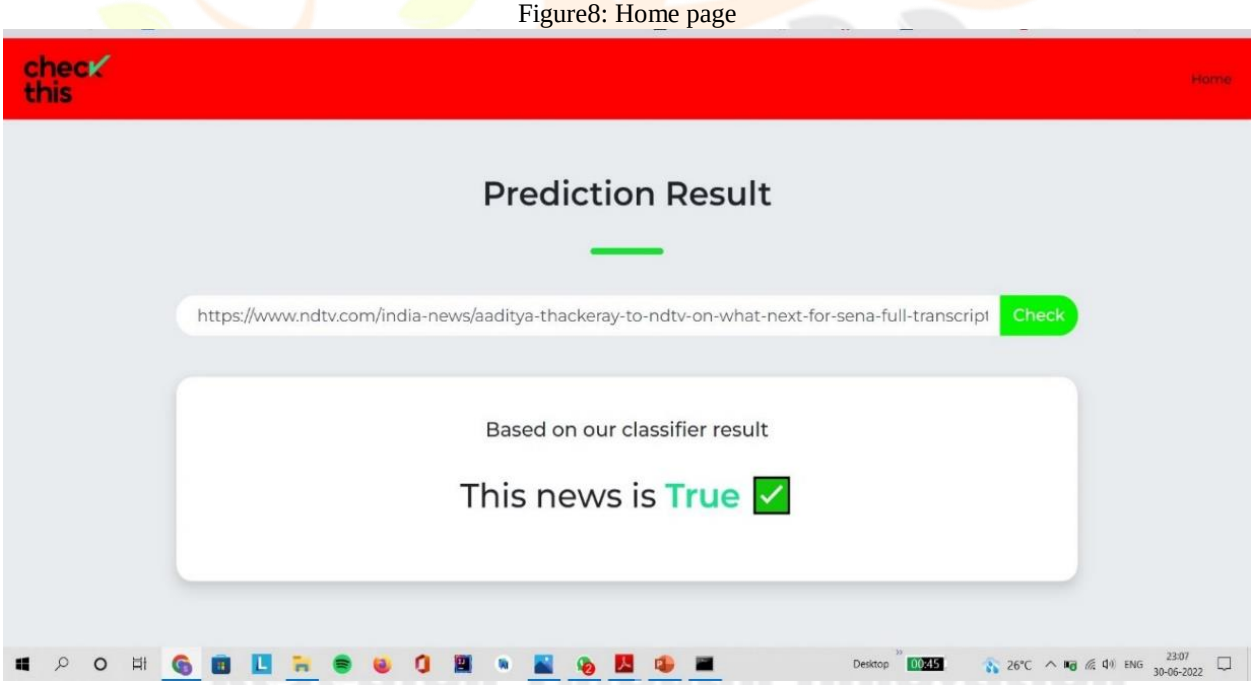

Figure9: True Prediction

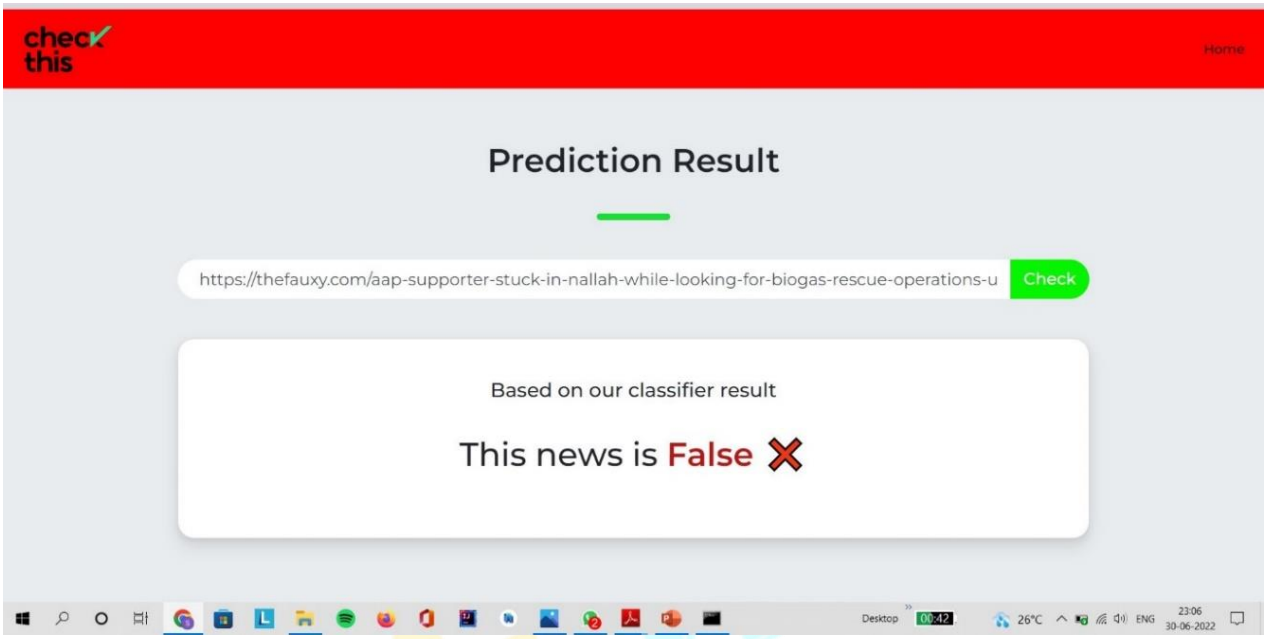

Figure10: False Prediction

Acknowledgment

The satisfaction that accompanies the successful completion of the Project would be incomplete without the mention of the people who made it possible, without whose constant guide and encouragement would have made our efforts go in vain. We consider ourselves privileged to express our gratitude and respect towards all those who guided us through the completion of this Project on, "Fake News Identifier through Machine Learning".

We consider ourselves proud to be part of Impact College of Engineering and Applied Sciences, Bangalore, the institute which stood by us through our endeavors. We would like to thank our guide Mrs. Chetana Patil, Assistant Professor, Department of Computer Science and Engineering, for the support and encouragement provided throughout the process.

We are grateful to Dr. Dhananjaya V, Professor and Head, Department of Computer Science and Engineering, for the support and encouragement provided throughout the process. We would like to express our gratitude to the Management and Principal, Dr. JALUMEDI BABU, Impact College of Engineering and Applied Sciences, for providing us the infrastructure and congenial environment to nourish our knowledge. We also express our gratitude and sincere thanks to all the teaching and non-teaching Staffs of Department of Computer Science and Engineering.

REFERENCES

[1]. Falsified News Detection using Deep Learning Approach Author: S. Pawar, G. Patel, P. Pawar, S. Khedkar, B. More Published On: 2021 Asian Conference on Innovation in Technology (ASIANCON) Pune, India. Aug 28-29, 2021

[2]. Big Data and quality data for fake news and misinformation detection Author: F. T. Asr and M. Taboada Published On: May 23, 2019 at Digital Object Identifier Organization

[3]. Fake News Detection using Deep Learning Author: S. H. Kong, L. M. Tan, K. H. Gan and N. H. Samsudin, Published On: 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)

[4]. Study of Fake News Detection using Machine Learning and Deep Learning Classification Methods Author: Keshav Nath. Aman Ahuja, Priyansh Soni, Rahul Katarya, Anjum. Published On: 2021 6th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), August 27th & 28th 2021

[5]. A Multi-semantics Classification Method Based on Deep Learning for Incredible Messages on Social media Author: L. Wu, Y. Rao, H. Yu, Y. Wang and N. Ambreen, Chinese Journal of Electronics Vol.28, No.4, July 2019

[6]. www.researchgate.net

[7]. https://www.arm.com

[8]. www.wikipedia.com

[9]. www.ieee.org

[10]. www.doi.org