**MINI PROJECT**

**ON**

**FAKE NEWS PREDICTION**

****

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**VASANTDADA PATIL PRATISHATHAN’S COLLEGE OF ENGINEERING AND VISUAL ARTS**

**MUMBAI UNIVERSITY**

**2022-23**

**Mini Project Report**

**ON**

**FAKE NEWS PREDICTION**

Submitted in partial fulfilment of the requirements of the requirements for the degree of Bachelor of Engineering in

Information Technology Semester-VI

By

|  |  |  |
| --- | --- | --- |
| .No | Name | ID No. |
| 1 | Kshitij Ghuge | VU4F2021003 |
| 2 | Jagannath Mestry | VU4F2021009 |
| 3 | Pranit Temkar | VU4F2021031 |
| 4 | Abhishek Palav | VU4F2021034 |

Under the Guidance

of

**Prof. MEDHA KULKARNI**



Department of Information Technology

Vasantdada Patil Pratishthan's College of Engineering & Visual Arts

UNIVERSITY OF MUMBAI

(AY 2022-23)

**CERTIFICATE**

This is to certify that the project entitled **“FAKE NEWS PREDICTION”** is bonafide work of KSHITIJ GHUGE(VU4F2021003), JAGANNATH MESTRY(VU4F2021009),PRANIT TEMKAR(VU4F2021031),ABHISHEK PALAV(VU4F2021034), submitted to University of Mumbai in partial fulfilment of requirement for the award of degree of “**Bachelor of Engineering in Information Technology-Semester-VI”.**

Prof. MEDHA KULKARNI

Guide

Dr.Pradip Mane Dr. Alam N.Shaikh

Head of Department(IT) Principal & Campus Director

**Project Report Approval for T.E.**

This project report entitled **“FAKE NEWS PREDICTION”** by Kshitij Ghuge, Jagannath Mestry, Pranit Temkar, Abhishek Palav, is approved for the degree of **Bachelor of Engineering** in **Information Technology**.

Examiners

1.

2.

Date:

Place: Mumbai – 22

Submitted by:

|  |  |  |
| --- | --- | --- |
| S No. | Name | ID No. |
| 1 | KSHITIJ GHUGE | VU4F2021003 |
| 2 | JAGANNATH MESTRY | VU4F2021009 |
| 3 | PRANIT TEMKAR | VU4F2021031 |
| 4 | ABHISHEK PALAV | VU4F2021034 |

## **ABSTRACT**

With increasing popularity in the use of social media for news consumption, the substantial widespread dissemination of fake news has the potential to adversely affect individuals as well as the society as a whole. Even in the midst of the current covid-19 pandemic, false information shared on websites such as WhatsApp, Twitter, and Facebook have the potential to cause panic and shock a large number of people in various parts of the world. These misconceptions obscure healthier habits and encourage incorrect procedures, which aid in the transmission of the virus and, as a result, result in poor physical and psychological health results for individuals. Therefore, it is a research challenge to validate the source, content and publisher of a news article for classifying it as genuine or fake. The existing systems and techniques are not efficient enough to accurately classify a given news based on its statistical rating. Machine learning plays an imperative part in categorizing news data and information, despite some limitations.

Our project not only aims on fake news detection but also on generation of real news once the fake news is detected. We propose a user-friendly webpage on which the user enters the news article statement. It is then tested by our machine learning algorithm which then classifies it as genuine or fake, after which the important words are extracted from the statement which helps to get the corresponding genuine news by scraping it from trusted sources and show it to the user. We have compared two machine learning algorithms in this which are- Passive Aggressive Classifier and Naïve Bayes algorithm. We got an accuracy of about 93.5% from Passive Aggressive Classifier and about 83.5% from Naïve Bayes algorithm.

### ACKNOWLEDGEMENT

The success and the final outcome of this project required guidance and assistance from different sources and we feel extremely fortunate to have got this all along the completion of our project. Whatever we have done is largely due to such guidance and assistance and we would not forget to thank them.

We express our sincere thanks to the Head of the Department, Department of Information Technology, Dr. Pradip Mane, for all the help and infrastructure provided to us to complete this project successfully and her valuable guidance.

We owe our profound gratitude to our project guide Prof. Medha Kulkarni, who took keen interest in our project work and guided us all along, till the completion of our project work by providing all the necessary information for developing a good system.

We are thankful to and fortunate enough to get constant encouragement, support and guidance from all the Teaching staff of the Department of Information Technology who helped us in successfully completing our project work. Also, we would like to extend our sincere regards to all the nonteaching staff of the department of Information Technology for their timely support.

**CONTENTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CHAPTER NO** | | | **NAME OF CHAPTER** | **PAGE NO.** |
| **1** |  |  | **INTRODUCTION** |  |
| 1.1 |  | INTRODUCTION DESCRIPTION | **10** |
| **2** |  |  | **LITERATURE REVIEW** |  |
| 2.1 |  | EXISTING SYSTEM | **12** |
| 2.2 |  | LIMITATION OF EXISITING SYSTEM AND RESEARCH GAP | **13** |
| 2.3 |  | PROBLEM STATEMENT AND OBJECTIVE | **14** |
| 2.4 |  | SCOPE | **15** |
| **3** |  |  | **PROPOSED SYSTEM** |  |
| 3.1 |  | ALGORITHM | **16** |
| 3.2 |  | DETAILS OF HARDWARE AND SOFTWARE | **17** |
| 3.3 |  | DESIGN DETAILS |  |
| 3.3.1 | PERT CHART/GANNT CHART | **19** |
| 3.3.2 | SYSTEM ARCHITECTURE | **20** |
| 3.3.3 | FLOW CHART | **21** |
| 3.4 |  | METHDOLOGY/PROCEDURES | **22** |
| 3.5 |  | RESULT AND DISCUSSIONS |  |
|  | 3.5.1 | RESULTS, DISCUSSION- COMPARATIVE STUDY/ ANALYSIS | **23** |
| **4** |  |  | **CONCLUSION** | **28** |
| **5** |  |  | **REFERNCES** | **29** |

**List of figures**

|  |  |  |
| --- | --- | --- |
| **Sr.No** | **Name of the figure** | **Page no.** |
| 1 | Use Case | 20 |
| 2 | System Architecture | 21 |
| 3 | Flow Chart |  |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Sr.no.** | **Name of the table** | **Page no.** |
| **1** | **Gantt Chart** | **14** |

**INTRODUCTION**

* 1. INTRODUCTION DESCRIPTION

In recent times we have come across many instances of fake news popping up in our Facebook and Twitter feeds from time to time. Not only social media sites suffer from this menace but even sites of popular newspaper publications provide its readers with fake information many a times. This is a major problem which needs to be addressed because fake news is misleading and tends to create unrest among common people. The best case of this fake news problem emerged in the 2016 US Presidential Elections where a number of political pundits claimed that the rise of significantly biased and/or untrue news influenced the election and hence may have been one of the major deciding factors for the result of the elections

Obviously, a deliberately misleading story is “fake news” but lately blathering social media discourse is changing its definition. Some now use the term to dismiss facts counter to their preferred viewpoints, which is especially seen in the political scenario and such a vaguely-defined term is ripe for cynical manipulation.

Back here in India we have been facing similar problems. Indians and the ‘Indian National Anthem’ being adjudged the best by UNESCO is probably the most common fake news going around the Internet. While this false information may have been innocuous, it captures the larger menace we face today. With the advent of new age digital and social media, fake news has pervaded all spheres of life, political and social.

Fabricated and manipulated content are gaining steam, leading to the possibility of potential violence and impacting society. The rise of digital and social media as powerful platforms has only magnified the effect of fake and false news. Umpteen number of news/information portals are being set up as there are few entry barriers unlike in the traditional media. In addition, growing polarisation of society on ideological lines has made the job of spreading fake news easier. Content that denigrates leaders/groups of the opposite ideology based on falsehoods, deepens communal polarisation or incites hatred have gained traction in the last few years. In the past, communal violence in India was a localised affair. Today, it is extensively being fed and triggered through the Internet. Provocative content, inaccurate information, doctored videos, and pictures are being disseminated through various online and mobile platforms. The platforms serve like nodal agencies distributing unverified information.

While media researchers around the world are investigating the fake news scene, little credible information is available on the creators and the intention behind it. But if what has been found is true, it is purely a way to make advertising money through click baits, enticing people to click and continue reading, and is organised by political or other social groups. In India, numerous sites are set up to peddle fake news with click bait headlines. They are also very well organised with multiple linked pages on social media platforms that are used to make the content go viral.

The advent of fake news is not new or recent, only its potential to reach people has amplified due to online platforms and applications that are free. the sheer expanse of the Internet and the anonymity it grants makes it difficult to track down people. Unlike mainstream media that falls under comprehensive regulation, online platforms have scope for wrongdoing due to the lack of binding rules, and the ability to keep owners and editors private like in the case of fake news sites. In the absence of such crucial information, there is no understanding of the liability and the credibility of the information that is being hosted on their respective sites. This is the main strength of the creators of fake news, the ability to remain anonymous in the guise of a media outlet. Most digital media outlets do not have basic information regarding editors, publishers or the physical address of the registered entity. We could do well to begin with some basic regulation for digital media outlets like compulsory and online registration of details.

Hence, we can conclude that detecting fake news is the need of the hour and here we have tried to create our own approach towards tackling the problem and have proposed a proper solution extensively documented below.

**LITERATURE REVIEW**

* 1. EXISTING SYSTEM

Through the inserted data, the algorithm will first learn to distinguish between bogus and authentic news. After understanding the distinction, the system will learn to make judgments based on the data presented. Fake news tracker programmes monitor the collection, analysis, and visualisation of fake news. The bogus database displays no news channel names, but the genuine dataset displays individual headquarters for each station.

Manipulating the concept of dataset fraudulent channels are exploiting an unregistered news portal. As a result, using the original dataset, one may compare and explicitly identify them.The data analysis also involves a number of dangers. The proper usage of data assessment in relation to references must be considered. During data analysis, there are some assessment elements that Python does not recognise, which creates the data clarity difficulty.

2.2 LIMITATION EXISITING SYSTEM AND RESEARCH GAP

A major challenge in fake news detection is to detect it in the early phase. Another challenge in fake news detection is the unavailability or the shortage of labelled data for training the detection models. We propose a novel fake news detection framework that can address these challenges.

The existing systems in fake news detection project have made significant progress in recent years, but there are still some limitations and research gaps that need to be addressed. Here are some of the key limitations and research gaps:

Limited data availability: One of the main challenges in developing effective fake news detection systems is the limited availability of reliable labeled datasets. Many of the existing datasets are small, biased, or outdated, which can make it difficult to train accurate and robust models.

Over-reliance on linguistic features: Many of the existing fake news detection systems rely primarily on linguistic features, such as sentence structure, word choice, and sentiment analysis. While these features can be informative, they may not be sufficient for detecting more sophisticated forms of fake news, such as deepfakes or manipulated images.

Lack of transparency and interpretability: Many of the existing fake news detection systems are based on complex machine learning models that can be difficult to interpret or explain. This can make it difficult to understand how the system is making decisions, and can limit its usefulness in real-world settings where transparency and accountability are important.

Limited coverage of languages and regions: Most of the existing fake news detection systems have been developed for English language content and may not generalize well to other languages or regions. This can be a significant limitation for detecting fake news in global contexts, where multiple languages and cultures are involved.

Limited focus on user behavior: Many of the existing fake news detection systems focus on the content of the news article, rather than the behavior of the users who consume and share the news. However, user behavior can be an important predictor of whether a news article is fake or not, and more research is needed to incorporate this information into detection systems.

2.3 PROBLEM STATEMENT AND OBJECTIVE

PROBLEM STATEMENT:

Often sensational news is created and spread through social media to achieve intended end. On the other hand, it may also involve narration of a true fact however being deliberately exaggerated. This may also affect the affect the importance of serious news media. The problem is to identify the authenticity of the news and online content. Equally important problem is to identify the bots involved in spreading false news

OBJECTIVE:

* To detect the fake news on social media.
* To identify and reduce the spreading of fake news.
* To determine the difference between real and fake news.
* To classify the news articles fake and real using deep learning algorithms.
* To increase the accuracy of fake news detection, using the hybrid model based (PassiveAggressiveaClassifier and NLP )model.
* To do comparative analysis of both models and choose the best model.
  1. SCOPE
* Massive amounts of data gave birth to AI systems that are already producing human-like synthetic texts, powering a new scale of disinformation operation .
* Based on Natural Language Processing (NLP) techniques, several lifelike text-generating systems have proliferated and they are becoming smarter every day.
* This year, OpenAI announced the launch of GPT-3, a tool to produce text that is so real, that in some cases it’s nearly impossible to distinguish from human writing. GPT-3 can also figure out how concepts relate to each other, and discern context.
* Tools like this one can be used to generate misinformation, spam, phishing, abuse of legal and governmental processes, and even fake academic essays.

**PROPOSED SYSTEM**

* 1. ALGORITHM

MACHINE LEARNING ALGORITHMS

* Passive Aggressive Classifier:

Passive Aggressive Classifiers are an online learning algorithm family that functions in the same way as a perceptron since they do not need a learning rate. Such a classifier remains passive when the classification outcome is correct, however it turns aggressive as soon as it comes across an incorrect outcome in the event of a miscalculation, after which it updates and modifies the unwanted outcome. In this project, such a classifier can help detect fake news and then fetch and generate relevant, genuine news to the user in the process from trusted news sources, thus fulfilling its purpose of making the much-needed modifications that corrects the loss. Due to its simplicity in terms of implementation as well as its quality to be used for incremental large-scale learning, it plays an imperative role in classifier 34 training stage after a dataset has been through a test-train split procedure to approximate and improve the efficiency of the machine learning model used in this project.

* TF-IDF Vectorizers

TF-IDF stands for term frequency-inverse document frequency and it is a measure, used in the fields of [information retrieval (IR)](https://en.wikipedia.org/wiki/Information_retrieval) and machine learning, that can quantify the importance or relevance of string representations (words, phrases, lemmas, etc)  in a document amongst a collection of documents (also known as a corpus).

Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document.

Inverse document frequency looks at how common (or uncommon) a word is amongst the corpus. IDF is calculated as follows where **t** is the term (word) we are looking to measure the commonness of and **N** is the number of documents (d) in the corpus (D).. The denominator is simply the number of documents in which the term, **t**, appears in.

3.2 DETAILS OF HARDWARE AND SOFTWARE

## **Software requirement:-**

## **PROGRAMMING LANGUAGE**

## **In this project, Python version 3.5 has been implemented. Python programming language offers creation of solutions to machine learning problems with code that is readable and intuitive, its simplicity also enables developers to develop robust, reliable projects. Python is also platform independent which enables the developers to deploy and utilize the code or frameworks on different systems with little to no changes. Python is also supported by a variety of platforms, some of which includes Windows, macOS and Linux.**

LIBRARIES

The libraries that have been implemented in this project are as follows:

* Sci kit-learn: It is a Python library and it plays an imperative role for implementing machine learning concepts using Python programming language. It contains functions and tools for machine learning as well as for statistical modelling which includes clustering, regression, classification and dimensionality reduction.
* NumPy: It is a Python library used to enable computational power to a python program. It contains N-dimensional arrays, matrix data structures and functions to work with arrays. It is a vital component to integrate variety of datasets into the project.
* Pandas: It is a package that provides developers with efficient, high-speed data analysis tools used to work with structured data which can be n-dimensional or tabular.
* Matplotlib: It is a Python library that consists of a set of functions that can be implemented to visualize and plot data.
* Seaborn: is a visualization library that is based on Matplotlib which is used to implement an interface to create interactive visualization and graphics.
* NLTK: Natural Language Toolkit is a python suite that contains functions and text processing packages such as stemming and tokenization in order to enable a python program to utilize natural language data. In this project, tokenizers such as RegexpTokenizer and 28 WordpunctTokenizer are implemented to extract tokens or key pieces of text by using regular expressions and by separating punctuation from string of words or sentences. Porter‘s stemmer algorithm has been implemented for the process of stemming used to reduce words into its root form to filter any unnecessary piece of text. This algorithm implements data mining and information retrieval techniques. When a news article entered by the user is detected and classified as fake, then RAKE which stands for Rapid Automatic Keyword Extraction is implemented which is a keyword search algorithm that determines key words or terms that occurs concurrently in different collection of documents based on which, relevant genuine news can be fetched and displayed to the user.
* Beautiful Soup: This is a Python library that is used to extract data from HTML and XML formats. In this project, such a library can help extract relevant news content from websites of trusted news sources once a news article input inserted by the user is classified as fake by the machine learning model.
*  Pickle: It is a python module that is used to convert an object structure into bytestream so that it can be stored in a file and then it can be converted back into bytestream. In other words, it can serialize and deserialize an object structure

PLATFORM (IDE)

In this project, VS Code IDE which is developed by Microsoft  has been implemented. VS Code also

supports various libraries and consists of an interactive console.

Google Colab which is a free IDE has also been implemented in this project. makes it easy for

developers to share code through their google drive account, it also comes preinstalled with plenty of

frequently used modules and has an user-friendly interface.

## 

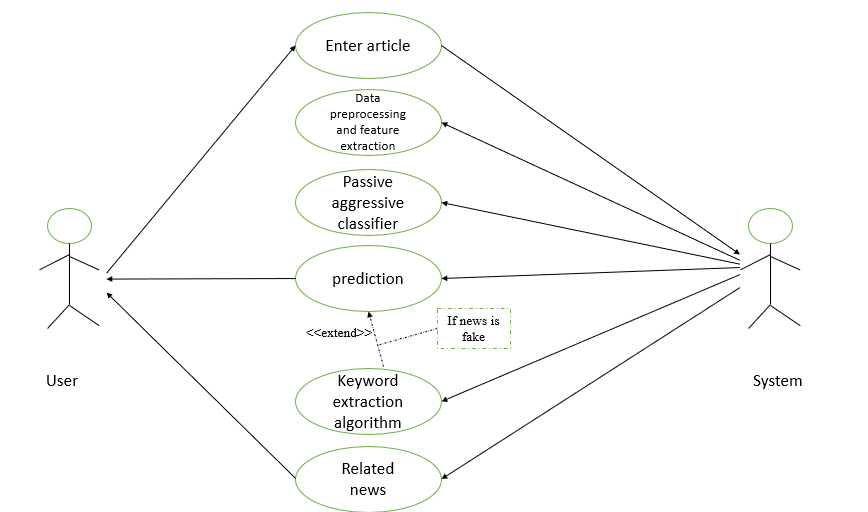
## **Hardware requirement:-**

* Processor : Intel Pentium 4 with 1.7 GHz speed or more.
* Ram : 512 MB
* Hard disk : At least 5 Gb

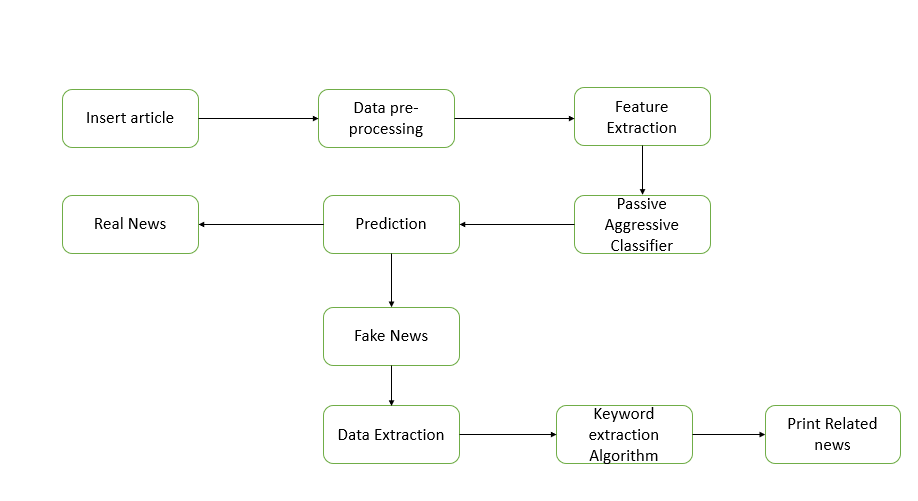
.

3.3 DESIGN DETAILS

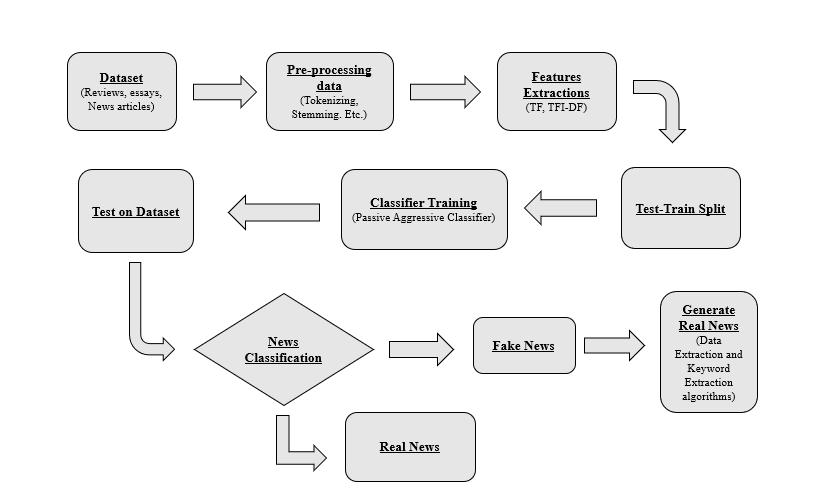
3.3.1 USE CASE DIAGRAM



3.3.2 SYSTEM ARCHITECTURE



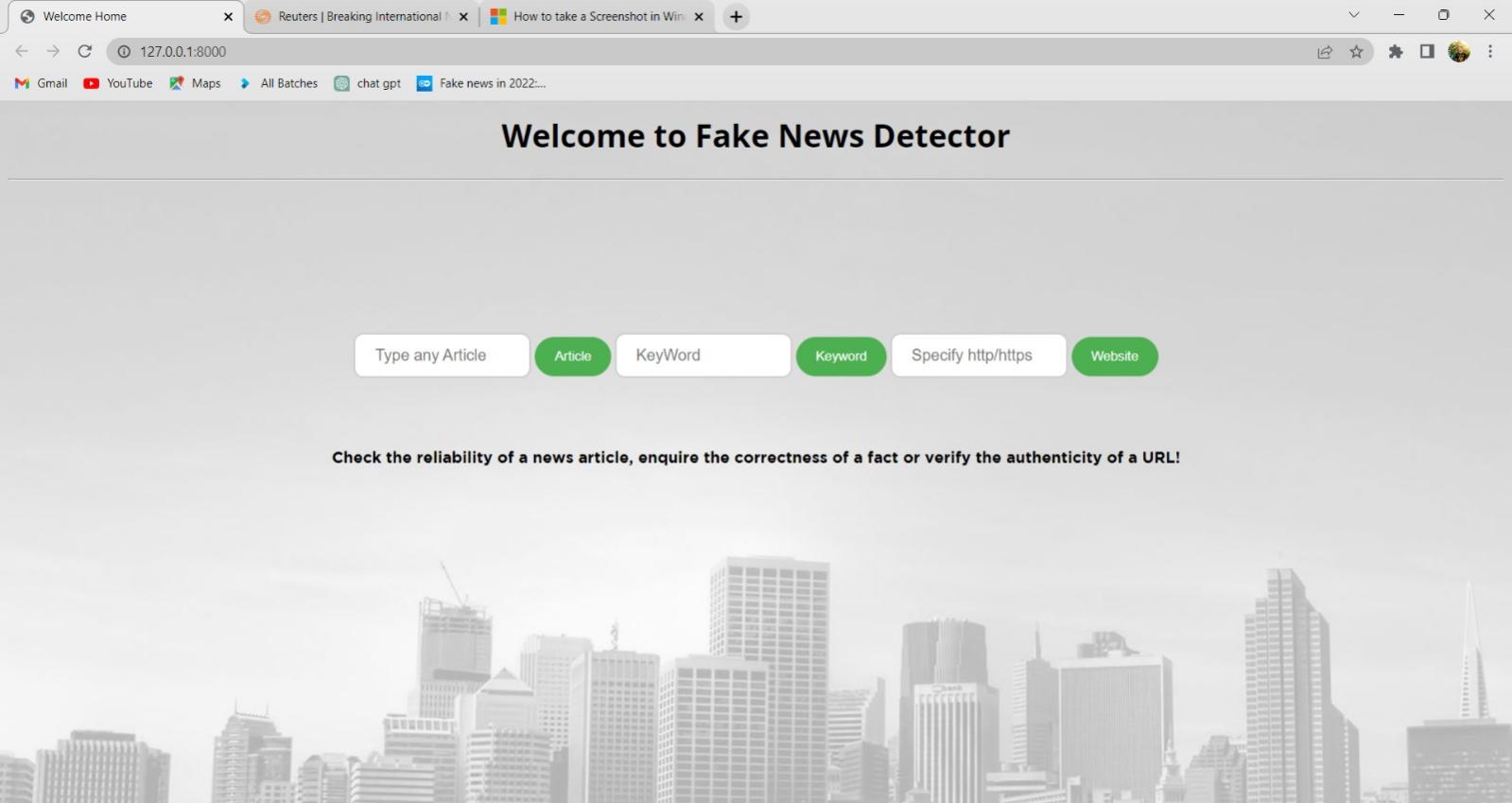
3.3.3 FLOW CHART

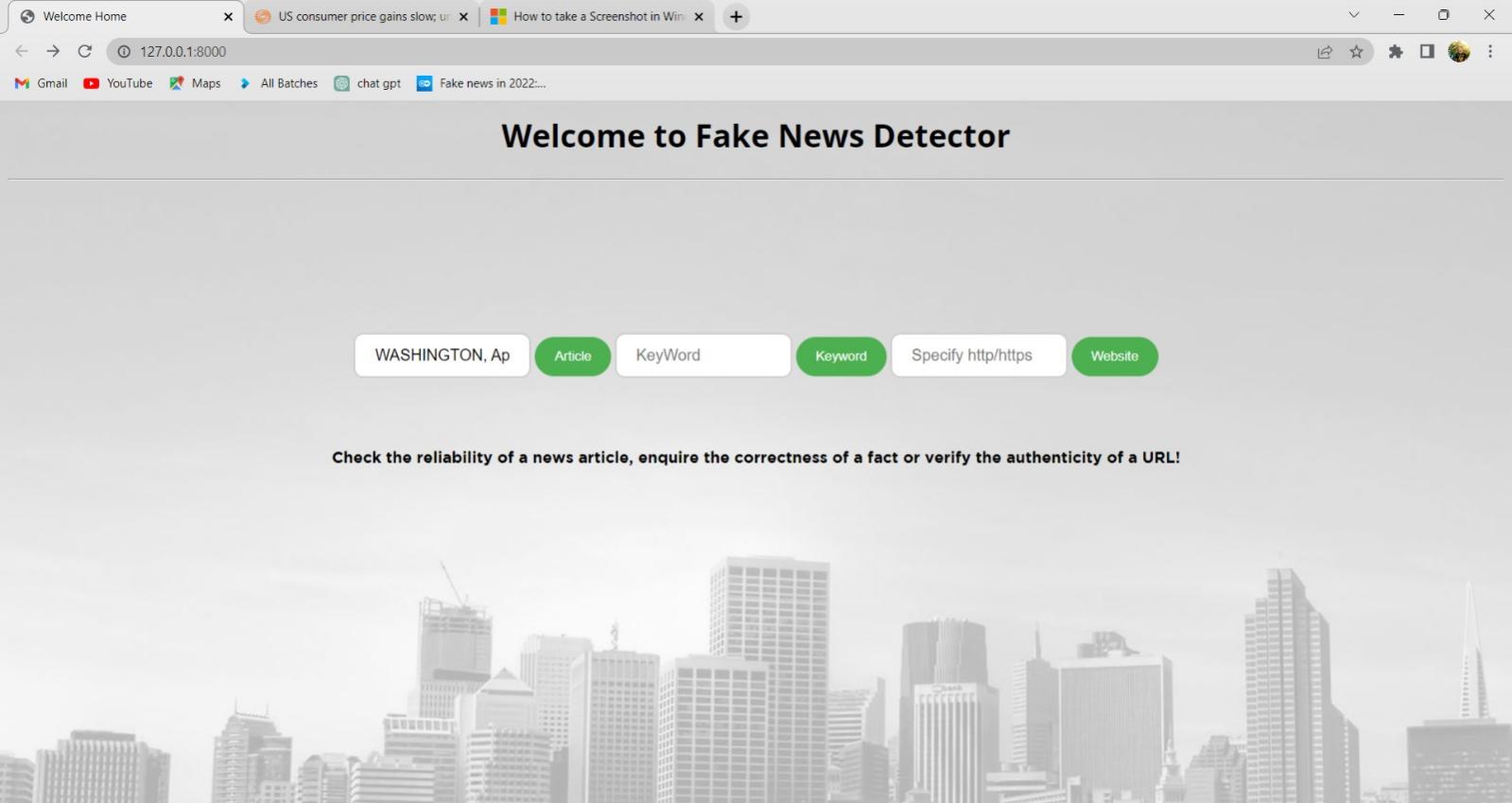


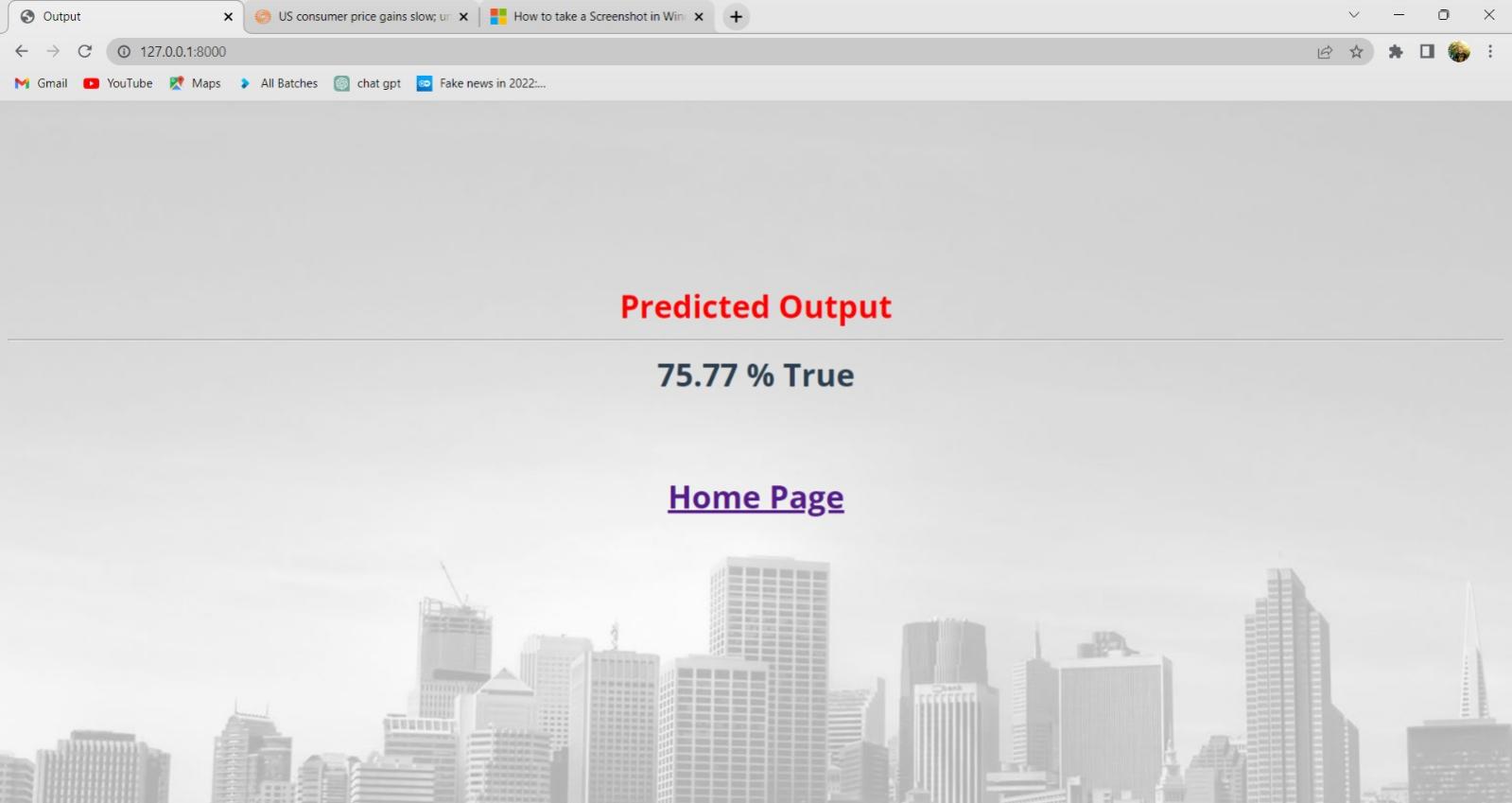
3.4 METHODOLGY /PROCEDURES

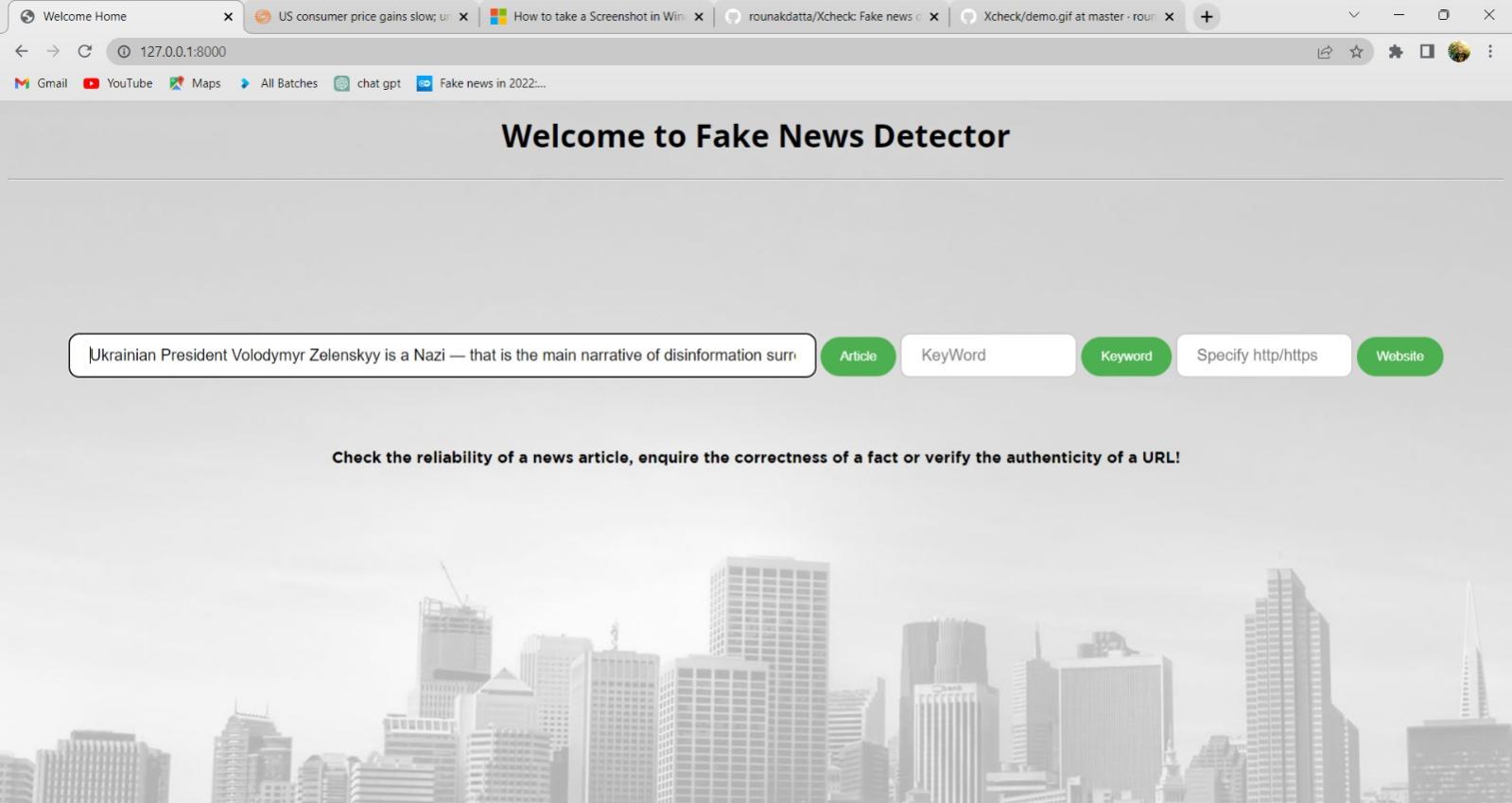
* The problem can be broken down into 3 statements :- Use NLP to check the authenticity of a news article.
* If the user has a query about the authenticity of a search query then we he/she can directly search on our platform and using our custom algorithm we output a confidence score.
* Check the authenticity of a news source. These sections have been produced as search fields to take inputs in 3 different forms in our implementation of the problem statement.
* 1 ST Search Field: The dataset available to us is split into the test and training sets using the train \_test \_split () function of Sci kit-learn.
* To get an idea if the words and tokens in the articles have a significant impact on whether the news in it is actually fake or real, we use Count Vectorizer and Tfi df Vectorizer.
* For the tf idf\_ vectorizer the threshold is set to 0.7 which means that the words which appear more than 70% of the time in the articles are not counted in.
* The stopwords parameter is used for avoiding the various stop words used commonly in the English language.

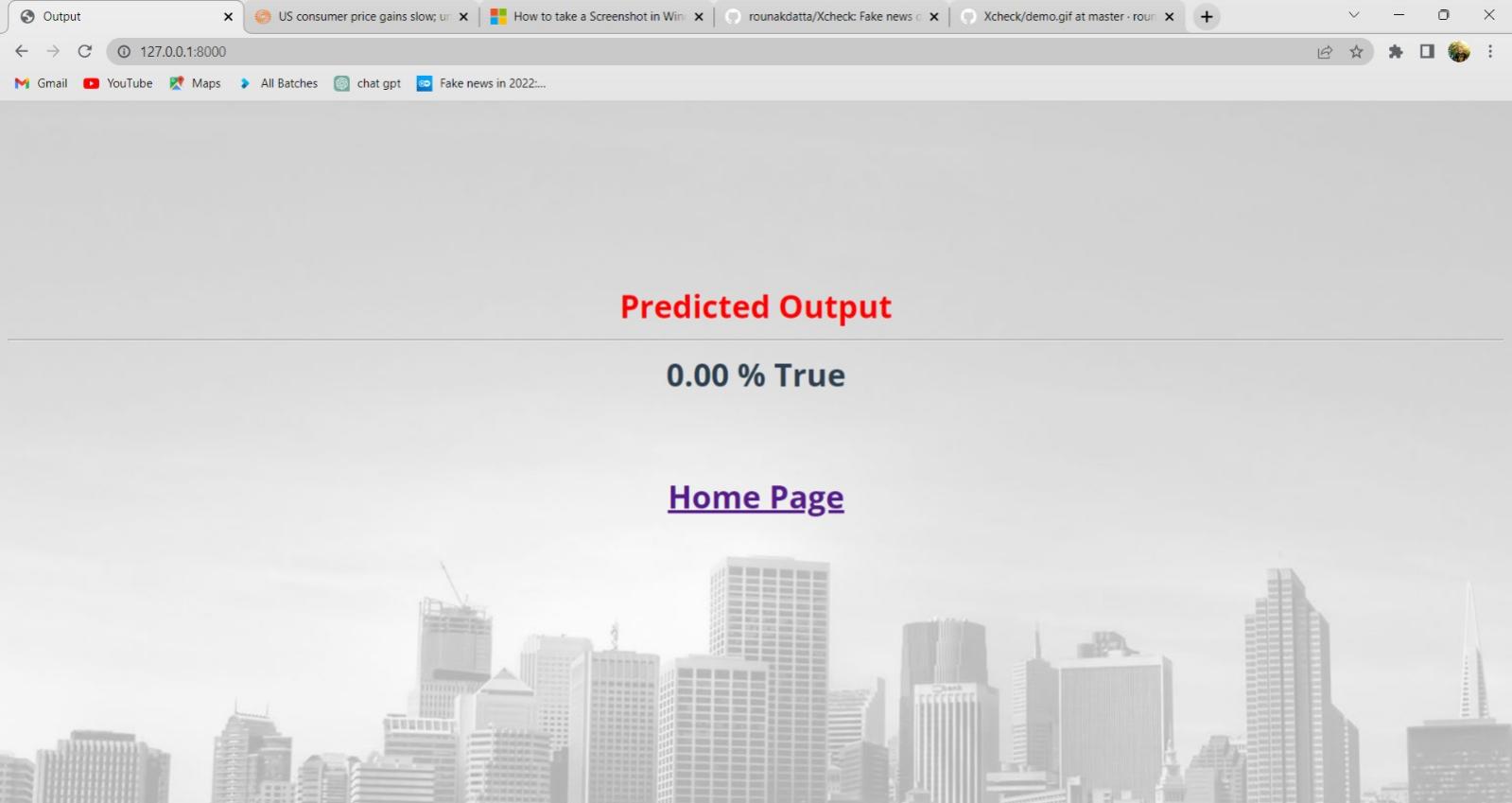
1. **RESULT AND DISCUSSIONS**

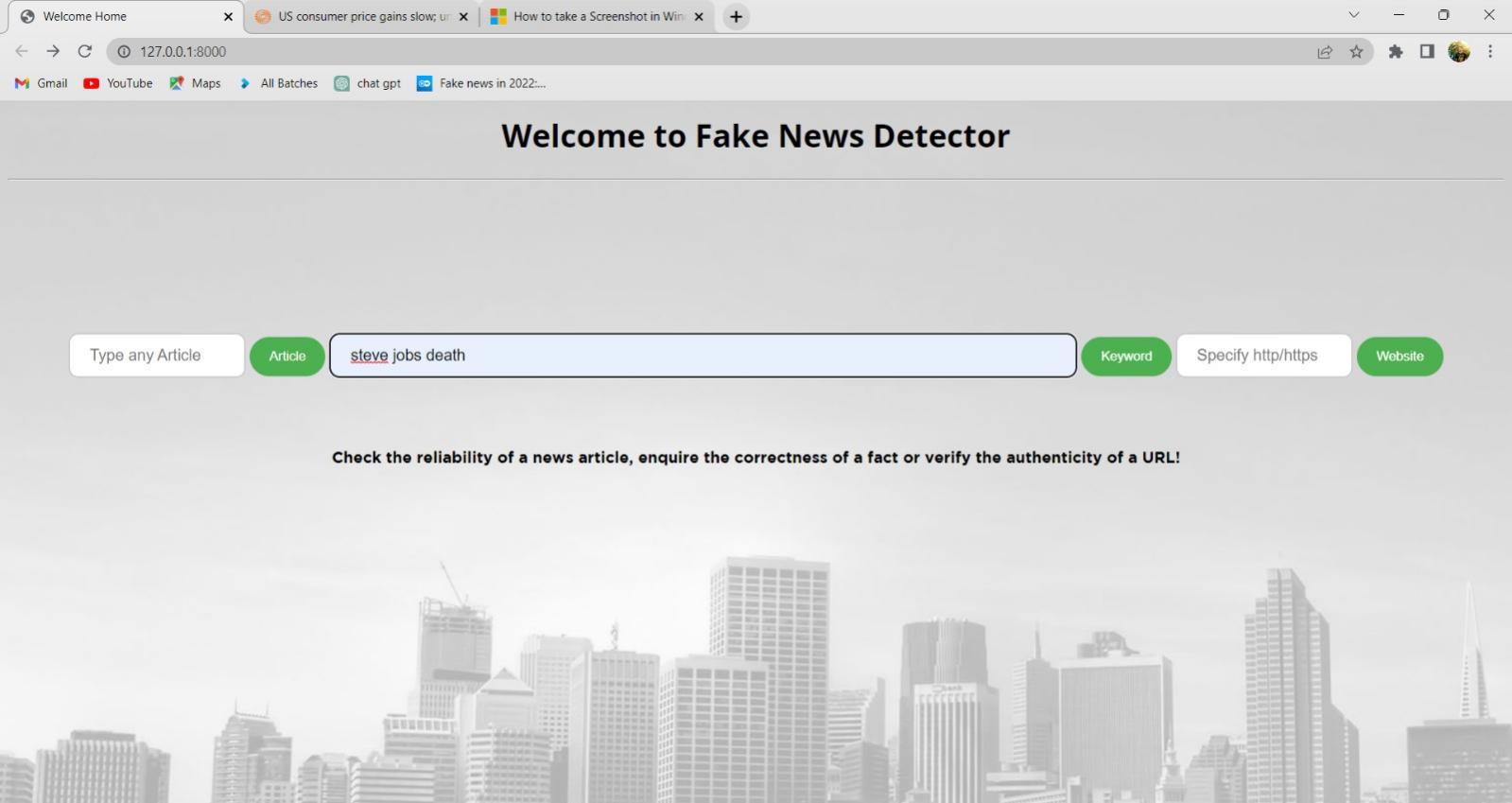
****

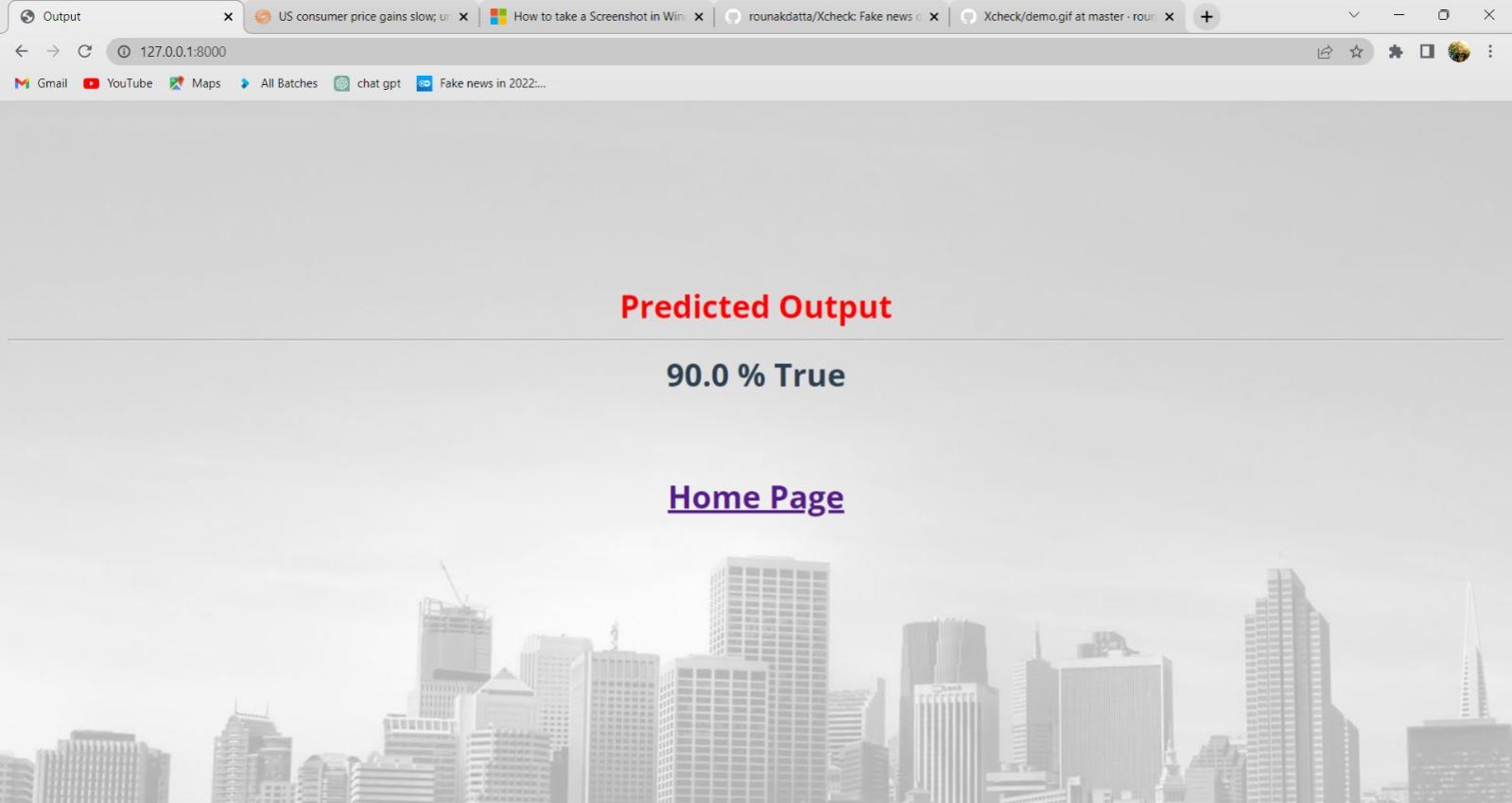
****

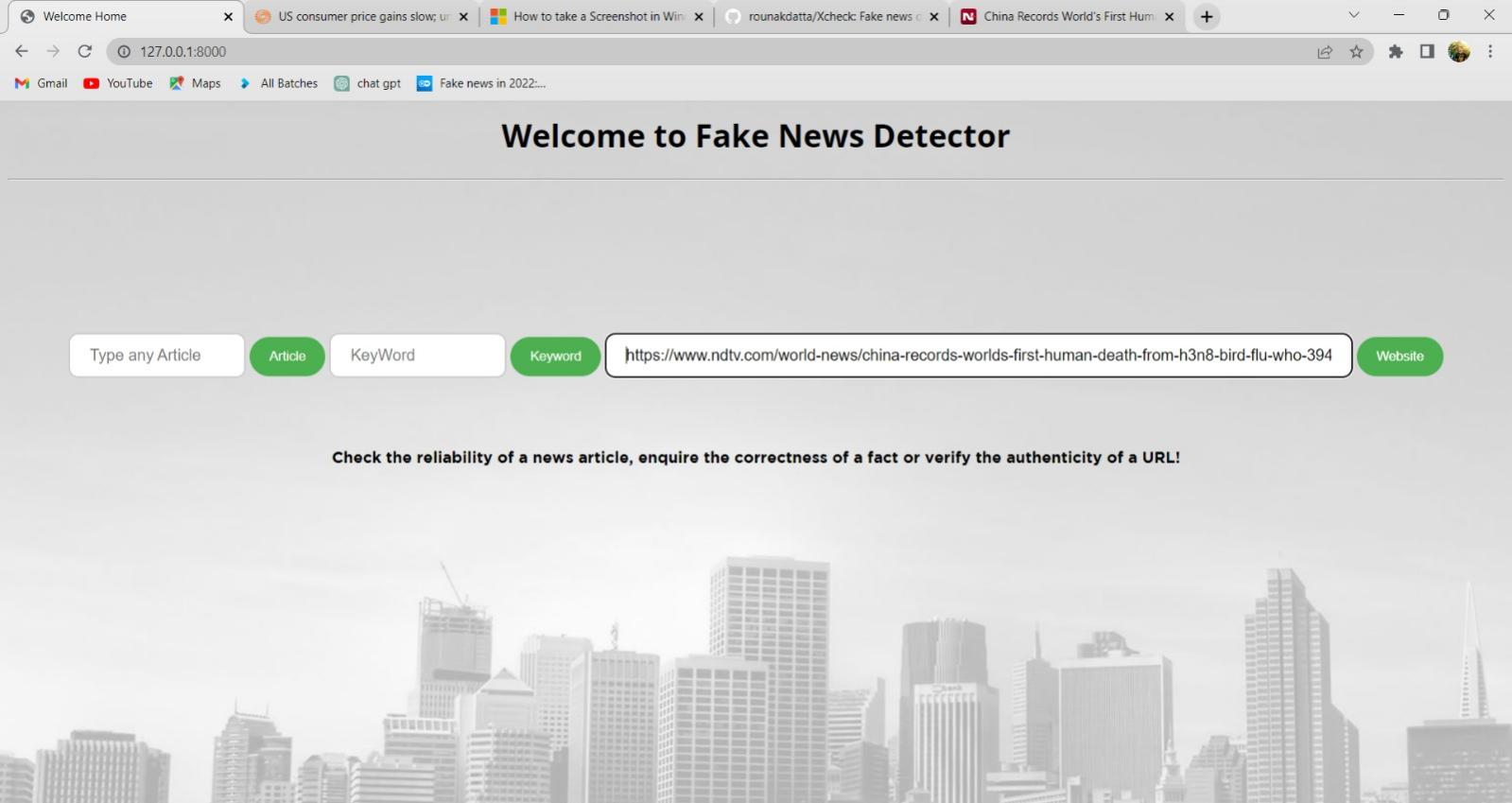
****

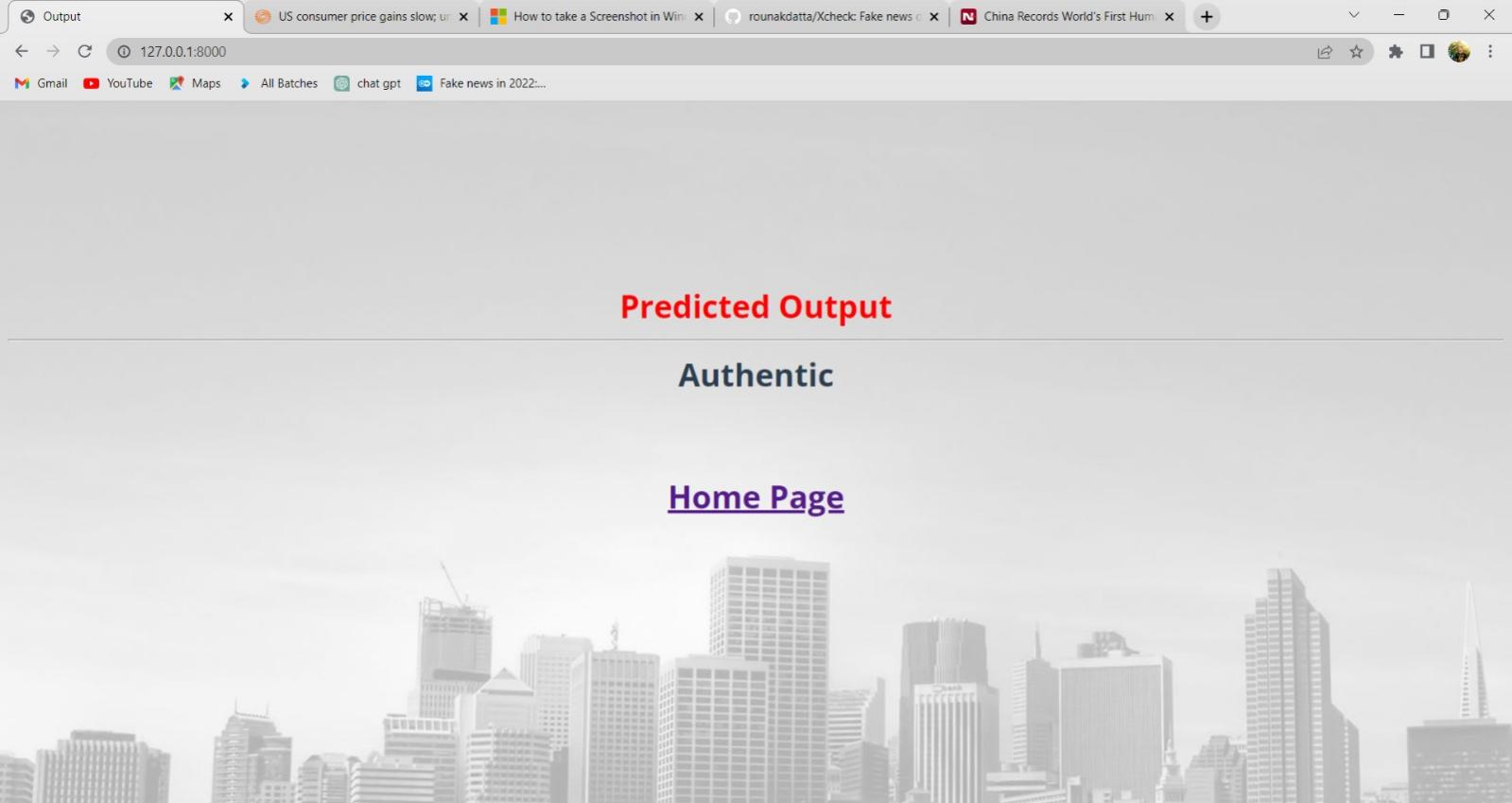
****

****

****

****

****

****

**5. CONCLUSION**

With the increased use of social media for news consumption and in prevalence, the widespread distribution of false news has the potential to harm both individuals and society as a whole. Even in the midst of the current covid-19 pandemic, false information on platforms like WhatsApp, Twitter and Facebook can cause panic and have a shocking impact not just on an individual but to a society as a whole. The objective is to detect the fake news through latest technologies and algorithms like Passive aggressive classifier. We used fake news detection where the user will enter the text and this text will go through our various models and at last give a prediction whether it is true or false. Further, our real news generation will check and validate the news and give us some news from trusted sites.

Our proposed model consists of two components, one where the detection takes place and the other where its correction takes place, if the news is found out to be false corresponding correct news is given as output. We determine the accuracy of these models and discuss about their limitations. In our project, the user can enter the text. Various machine learning algorithms are performed and we found out that Passive aggressive classifier gives a better accuracy as compared to Naïve Bayes. Further, the data is extracted and then real news generation is done using the keyword extraction algorithm. On the basis of our analysis, we can successfully remove the fake news if any.

**6.REFERENCES**

[1] Yang, S., Shu, K., Wang, S., Gu, R., Wu, F. and Liu, H., 2019, July. Unsupervised fake news detection on social media: A generative approach. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 5644-5651).

[2] Kai Shu , Amy Sliva , Suhang Wang , Jiliang Tang , and Huan Liu,2017 september. Fake News Detection on Social Media: A Data Mining Perspective

[3] Yanagi, Y., Orihara, R., Sei, Y., Tahara, Y. and Ohsuga, A., 2020, July. Fake News Detection with Generated Comments for News Articles. In 2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES) (pp. 85-90). IEEE.

[4] Zhang, J., Dong, B. and Philip, S.Y., 2020, April. Fakedetector: Effective fake news detection with deep diffusive neural network. In 2020 IEEE 36th International Conference on Data Engineering (ICDE) (pp. 1826-1829). IEEE.

[5] Thota, A., Tilak, P., Ahluwalia, S. and Lohia, N., 2018. Fake news detection: A deep learning approach. SMU Data Science Review, 1(3), p.10.

[6] Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H., 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. Big Data, 8(3), pp.171-188.

[7] Groza, A., 2020. Detecting fake news for the new coronavirus by reasoning on the Covid-19 ontology. arXiv preprint arXiv:2004.12330.