A

Project Report

on

Fake News Detection using Naive Bayes classifier

Submitted in partial fulfillment of the requirements

for the award of the degree of

Bachelor of Technology

in

Data Science

by

Kavisha Jain (21520002)

Under the Supervision of

Dr Sahil Kansal

JAGAN NATH UNIVERSITY, BAHADURGARH

BAHADURGARH-JHAJJAR ROAD, HARYANA

JAGAN NATH UNIVERSITY, BAHADURGARH
BAHADURGARH-JHAJJAR ROAD, HARYANA

CERTIFICATE

This is to certify that the project report entitled "**Fake News detection using Naive Bayes classifier**" submitted by Mr./Ms. Kavisha Jain **,Roll No: 21520002** to the Jagan Nath University, Bahadurgarh Bahadurgarh-Jhajjar Road, Haryana in partial fulfillment for the award of Degree of Bachelor of Technology in Data Science is a bonafide record of the project work carried out by them under my supervision during the year 2023-2024.

Ms. Rani (Project Guide)
Sr Faculty

Dr Sahil Kansal
Mentor-IT
Academics

JAGAN NATH UNIVERSITY, BAHADURGARH
BAHADURGARH-JHAJJAR ROAD, HARYANA

ACKNOWLEDGEMENT

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend my sincere thanks to all of them.

We are highly indebted to Dr Sahil Kansal for his guidance and constant supervision. Also, we are highly thankful to them for providing necessary information regarding the project & also for their support in completing the project.

We are extremely indebted to Dr Sahil Kansal, Department of Information Technology, and Ms. Rani, Project Coordinator, for their valuable suggestions and constant support throughout my project tenure. We would also like to express our sincere thanks to all faculty and staff members of the Department for their support in completing this project on time.

We also express gratitude towards our parents for their kind cooperation and encouragement which helped me in completion of this project. Our thanks and appreciation also go to our friends in developing the project and all the people who have willingly helped me out with their abilities.

(Kavisha Jain)

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 situations like elections ,Russia - Ukraine war,the covid-19 pandemic etc, 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, this project validates the source, content and publisher of a news article for classifying it as genuine or fake. Machine learning plays an imperative part in categorizing news data and information, despite some limitations. In this project we use Naive Bayes classification model to predict whether a text article will be labeled as real or fake.

KEYWORDS: Naïve Bayes algorithm, fake news detection,Machine Learning , text article

CONTENTS

ABBREVIATIONS

ML - Machine Learning
NLP - Natural Language Processing
TF-IDF -> Term Frequency–Inverse Document Frequency
NLTK -> Natural language toolkit

CHAPTER 1: INTRODUCTION

Machine learning is a rapidly growing field of computer science that involves the development of algorithms and models that enable computers to learn from data and improve their performance over time. It is a subset of artificial intelligence (AI) that focuses on the development of algorithms that can learn and make decisions based on data inputs.

The goal of machine learning is to enable computers to recognize patterns in data, and use those patterns to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms learn from data through a process of training, where they are fed large amounts of data and adjust their parameters based on the patterns and relationships they find in that data. Once trained, machine learning models can be used to analyze new data and make predictions or decisions based on what they have learned.

There are several types of machine learning algorithms, including supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the algorithm is trained on labeled data, meaning that it has been given specific examples of what it should be predicting or classifying. This is used in applications such as image recognition and natural language processing. In unsupervised learning, the algorithm is trained on unlabeled data and must identify patterns and relationships without any specific guidance. This is used in applications such as clustering and anomaly detection. In reinforcement learning, the algorithm learns by trial and error, receiving feedback in the form of rewards or penalties based on its actions. This is used in applications such as game playing and robotics.

Machine learning has a wide range of applications, including image recognition, natural language processing, recommendation systems, fraud detection, and predictive analytics. It is being used by companies and organizations of all sizes to gain insights from data, improve decision making, and automate processes. In addition, machine learning is being used to tackle some of the world's most challenging problems, such as climate change and disease detection.

As the amount of data being generated continues to grow, and the demand for intelligent decision-making tools increases, machine learning is expected to continue to play an increasingly important role in shaping the future of technology and innovation.

Machine learning can be a powerful tool in combating fake news by identifying and flagging false or misleading information before it is widely disseminated. There are several ways in which machine learning can be used to address the problem of fake news:

Content analysis: Machine learning algorithms can be trained to analyze the content of news articles, social media posts, and other online content to identify patterns and markers of false or misleading information. For example, algorithms can be trained to detect sensationalist language, exaggerated claims, or biased reporting.

Source evaluation: Machine learning algorithms can also be trained to evaluate the credibility and reliability of news sources, using a variety of indicators such as the reputation of the publisher, the expertise of the author, and the consistency of the information with other reputable sources.

Network analysis: Machine learning algorithms can be used to analyze the networks of individuals and organizations that spread fake news, identifying key influencers and patterns of dissemination. This can help to identify sources of fake news and prevent it from spreading further.

User profiling: Machine learning algorithms can also be used to analyze the behavior of individual users, identifying patterns of engagement with fake news and other false information. This can help to identify users who are likely to be susceptible to fake news and target them with educational interventions or counter-messaging.

While machine learning is not a perfect solution to the problem of fake news, it has the potential to significantly improve our ability to identify and counter false or misleading information. However, it is important to note that machine learning algorithms are only as good as the data they are trained on, and must be continuously updated and refined to keep up with the evolving tactics of those who spread fake news.

NLP stands for Natural Language Processing. It is a field of study that focuses on the interaction between human language and computers. Specifically, it involves developing algorithms and computational models that can analyze, understand, and generate natural language text or speech.

NLP involves a wide range of tasks, including language translation, sentiment analysis, speech recognition, text classification, question-answering, and more. The goal of NLP is to enable machines to process and understand human language in a way that is similar to how humans do, which can have many practical applications, such as improving customer service chatbots, developing better language learning tools, and automating the analysis of large amounts of textual data.At its core, NLP involves using computational techniques to extract meaning from natural language data. This involves breaking down language into its component parts, such as words, sentences, and phrases, and analyzing these components in order to infer meaning.

One of the key challenges in NLP is dealing with the complexity and ambiguity of human language. Human language is incredibly varied and context-dependent, and the same words or phrases can have different meanings depending on the situation. For example, the phrase "I saw her duck" could refer to either an animal or a physical action, depending on the context.

To overcome these challenges, NLP researchers have developed a wide range of techniques and algorithms. Some of the most common techniques used in NLP include:

Tokenization: This involves breaking text up into individual words or tokens, which can then be analyzed individually.

Part-of-speech tagging: This involves identifying the parts of speech of individual words, such as whether they are nouns, verbs, adjectives, etc

Named entity recognition: This involves identifying named entities such as people, places, and organizations within text.

Sentiment analysis: This involves analyzing the overall sentiment or emotion expressed in a piece of text.

Machine translation: This involves automatically translating text from one language to another.

Text summarization: This involves automatically generating a summary of a longer piece of text.

These techniques can be used in a wide range of applications, from chatbots and virtual assistants to language learning tools and social media analysis. However, despite the significant progress that has been made in the field, NLP is still a rapidly evolving field, with many open research questions and challenges yet to be fully addressed.Some of the most popular machine learning algorithms used in NLP include neural networks, decision trees, and support vector machines.

By using machine learning algorithms to automatically learn patterns and relationships in natural language data, NLP researchers and practitioners can build more sophisticated and effective language processing systems. Machine learning is a critical component of many NLP applications, such as speech recognition, machine translation, and text classification.Machine learning involves developing algorithms that can automatically learn patterns and relationships in data, without being explicitly programmed. In the context of NLP, this means training algorithms on large datasets of natural language text or speech, in order to learn how to analyze, understand, or generate language.

MOTIVATION

The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in human history before. With the current usage of social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading with no relevance to reality.

Automated classification of a text article as misinformation or disinformation is a challenging task. Even an expert in a particular domain has to explore multiple aspects before giving a verdict on the truthfulness of an article. In this work, we propose to use a machine learning ensemble approach for automated classification of news articles. Our study explores different textual properties that can be used to distinguish fake contents from real.Fake news refers to false or misleading information presented as if it were true, often disseminated through traditional or social media platforms with the intention to deceive or mislead people. It can take many forms, including fabricated stories, manipulated images or videos, or biased reporting.

Fake news can be created and spread by individuals, organisations , or even governments to promote a particular agenda or to gain attention and clicks. It can also be the result of a lack of journalistic integrity or fact-checking, where inaccurate or misleading information is reported as true.
Fake news can have serious consequences, such as causing harm to individuals or groups, influencing political opinions, or even destabilising entire societies. In some cases, fake news can lead to violence or exacerbate existing conflicts.

To combat fake news, it is essential to critically evaluate the sources of information, fact-check claims before sharing them, and support reputable journalism. It is also important to be aware of our own biases and to seek out diverse perspectives to ensure that we have a more complete understanding of complex issues.

Fake news can be used as a tool for propaganda, to spread misinformation and manipulate public opinion. It can also be used to generate profits through clickbait headlines and sensationalized stories that attract more readers or viewers.

In recent years, the term "fake news" has become a political buzzword, often
Fake news is a big problem for several reasons. Firstly, fake news can spread quickly and easily through social media and other online platforms. This is because people tend to share news stories that they find interesting or that confirm their existing beliefs, without taking the time to verify their accuracy. As a result, fake news can quickly gain traction and reach a wide audience.
Secondly, fake news can have serious consequences. It can influence people's opinions and decisions, leading to actions that can be harmful to themselves or others. For example, during the COVID-19 pandemic, the spread of fake news about unproven treatments or conspiracy theories has led some people to refuse vaccination or to ignore public health guidelines, potentially putting themselves and others at risk.
Fake news can also have political implications, as it can be used to influence public opinion and sway elections. In recent years, there have been numerous cases of fake news being used to spread propaganda or to manipulate public opinion in favor of a particular political candidate or party.

Furthermore, fake news can undermine trust in the media and other institutions. If people cannot trust the news sources they rely on for information, they may become more susceptible to misinformation and propaganda.
The problem of fake news is compounded by the fact that it can be difficult to distinguish between real and fake news stories. This is especially true for people who lack media literacy skills or who are not accustomed to critically evaluating news sources.

To combat the problem of fake news, it is important to raise awareness about the issue and to encourage people to be more skeptical of news stories they encounter online. It is also important to develop effective techniques for detecting and flagging fake news stories, such as those based on machine learning algorithms. Additionally, media organizations can help to combat fake news by fact-checking stories and providing accurate, balanced reporting.

CHAPTER 2: LITERATURE REVIEW

In this paper, Shuo yang et al [1] examine the matter of Unsupervised discovery of fake news on social media by utilizing the users‘ reckless social media engagement details. They used current event truths and users' integrity as dormant random factors, and they used users' social media engagements to recognise their views on the validity of current events. They suggest a method for unsupervised learning. This system employs a probabilistic graphical paradigm to model current case truths and, as a result, the users' reputation. To solve the inference dilemma, an effective Gibbs sampling technique is proposed. Their experiment results show that their proposed algorithm outperforms the unsupervised standards.

Kai Shu et al [2] examines two facets of the issue of false news identification: -

Characterization- This aspect introduces the fundamental concepts of fake news in both traditional and social media.

Detection- The current detection methods, including feature extraction and model construction, are examined from a data mining perspective.

They described fake news and characterized it by evaluating various theories and properties in both traditional and social media. They continue to systematically describe the issue of detecting fake news and summarize the strategies of doing so. They discussed about the datasets and measurement criteria that are currently used in existing methods.

Yuta Yanagi et al [3] proposes a fake news detector that can create fake social contexts(comments), with the aim of detecting fake news early on in its spread when few social contexts are available It's been trained on a series of news articles and their social situations. They also trained a classify model using news posts, real-posted comments, and generated comments. They compared the quality of produced comments for articles with actual comments and those generated by the classifying model to determine the detector's effectiveness.

Limitation- According to their study, the words "!", "?", "false," "breaking," and other similar phrases are essential signals of fake news.

[4] The task of classifying news manually needs in-depth data of the domain and experience to spot anomalies within the text. During this analysis, we tend to mention the matter of classifying faux news articles victimization by machine learning model and NLP techniques. The info we tend to utilize in our work is Kaggle dataset and contains news articles folks Election results. The first aim is to spot patterns in text that differentiate faux articles from true news. We tend to extract totally different matter options through NLP techniques associated with the feature set as an input to the models. The Multinominal Naïve Bayes classifier was trained and parameter-tuned to get optimum accuracy

CHAPTER 3: PROBLEM FORMULATION

Problem statement

Problem Statement: Fake News Detection using Naive Bayes Classifier

The spread of fake news has become a significant concern in today's digital age. False information presented as legitimate news can mislead individuals, influence public opinion, and disrupt democratic processes. Therefore, there is a pressing need to develop effective techniques to detect and combat fake news.

The problem at hand is to design a system that can accurately identify fake news articles from a given dataset. The objective is to employ the Naive Bayes classifier, a machine learning algorithm, to classify news articles as genuine or fake based on a set of predefined features. By leveraging the principles of probability theory and the independence assumption, the Naive Bayes classifier can provide a probabilistic assessment of the likelihood of an article being fake.

The project aims to address the following challenges:

Dataset preparation: Gathering a comprehensive and well-labeled dataset of news articles that encompass both genuine and fake examples. The dataset should cover various topics and domains to ensure a diverse range of news articles.

Feature selection: Identifying informative features that can help distinguish between genuine and fake news articles. These features can include textual characteristics, linguistic patterns, metadata, and contextual information.

Model training: Building and training a Naive Bayes classifier using the prepared dataset and selected features. This involves encoding the articles into a suitable representation (e.g., bag-of-words) and estimating the conditional probabilities based on the occurrence of features in genuine and fake articles.

Performance evaluation: Assessing the performance of the Naive Bayes classifier using appropriate evaluation metrics such as accuracy, precision, recall, and F1 score. The goal is to achieve high accuracy in correctly identifying fake news articles while minimizing false positives and false negatives.

Handling limitations: Understanding and addressing the limitations of the Naive Bayes classifier in the context of fake news detection. Exploring ways to mitigate the impact of assumptions like feature independence and the handling of out-of-vocabulary words.

By developing an effective fake news detection system using the Naive Bayes classifier, we can contribute to the broader goal of combating misinformation and promoting media literacy. The system can be used as a tool to assist individuals, fact-checkers, and social media platforms in identifying and flagging potentially misleading or false news articles, thereby fostering a more informed and responsible society.

CHAPTER 4: PROPOSED WORK

4.1 PROGRAMMING LANGUAGE

Python programming language is an open-source programming language and since it is free, its use is extensive and has an active community development and support.

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.

One of the major reasons for implementing Python programming language is its extensive collection of libraries and frameworks. In this project, Pandas, NumPy, Seaborn are a handful of examples of libraries that have enabled developers to create the system quickly and effectively.

4.2 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 a 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 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.

4.3 PLATFORM (IDE)

In this project ,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. (even supports Anaconda IDE.)

CHAPTER 5: SYSTEM DESIGN

4.1 SYSTEM OVERVIEW

We have moved from receiving news from old, traditional means such as radio, newspapers, and TV news to more widespread, dynamic outlets which can be attributed to the growth of the internet, thanks to the rapid rise of social media and the protean technological advances in recent years. Thus, we are living in a time when knowledge is readily available and increasing exponentially. However, such conveniences that have been brought on by whole host of social media networks have also added multiple layers of intricacies and complexities which have made it more complicated for a news consumer to differentiate between genuine and fake news, and such dissemination of news followed by sharing and forwarding of such news articles without cross-verification have contributed to rise in prevalence of falsification of news that can not only have grave consequences in the events of the real world but also risks the credibility of social media.

We suggest a model in this project that makes use of machine learning algorithms and various feature extraction methods to identify fake news by cross-referencing it with other reliable news sources, as well as producing and displaying real news from reliable sources in the form of a website. To achieve a perfect result, we strive to achieve maximum accuracy in fake news detection and real news generation in this project.

These are the steps followed:

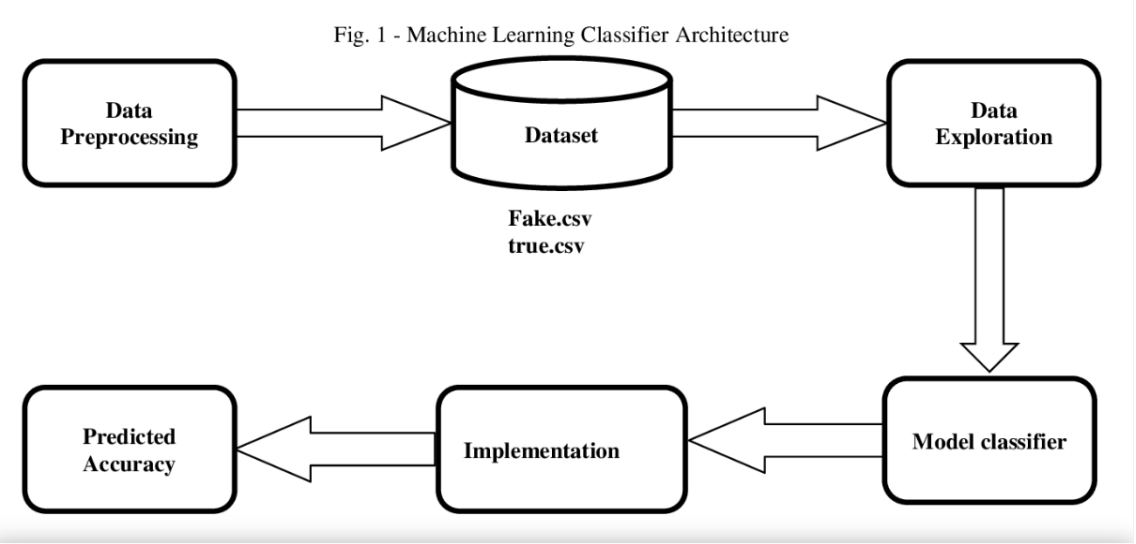
A model is proposed to check whether a given stance of information or news article is true or false.

Basically, the title content and domain name are checked.

Once we know that a piece of information is not real, it will give genuine news from trusted sites so the dissemination of false information can be stopped.

.

Fig 4. Architecture Diagram



4.2 FAKE NEWS DETECTION

4.2.1. DATA COLLECTION

In the proposed system, the data is collected keeping in mind the current covid situation. So, we have collected the dataset which were publicly available on Kaggle. We went through various datasets and at last came up with dataset with maximum number of records.

4.2.2. PRE-PROCESSING

In the pre-processing step, the data is cleaned such that the unwanted and unnecessary information can be removed and only the relevant details will be kept. In this project we have used Stemming and stopwords. There are different methods used in pre-processing. Some of the methods are mentioned below-

TEXT NORMALIZATION:

Text normalization is a process of transforming text into a single canonical form. Normalizing text before storing or processing it allows for separation of required data from the rest so that the system can send consistent data as an input to the other steps of the algorithm.

STOP WORD REMOVAL

A Stop Word is a commonly used word in any natural language such as “a, an , the, for, is,was, which, are, were, from, do, with, and, so, very, that, this, no, yourselves etc....”. These Stop Words will have a very high frequency and so these should be eliminated while calculating the term frequency so that the other important things are given priority. Stopword removal is such a Pre-processing step which removes these stop words and thereby helping in the further steps and also reducing some processing time because the size of the document decreases tremendously.
Consider a Sentence
“This is a sample sentence, showing off the stop word removal”.
Output after Stop word removal is:
[“sample”, “sentence”, “showing”, “stop”, “word”, “removal”]
Note: Though Stop words refer to the most commonly used words in a particular language, there is no single universal list of stop words, different tools uses different stop words.

STEMMING:
Stemming is a pre-processing step in Text Mining applications as well as a very common requirement of Natural Language processing functions. In fact it is very important in most of the Information Retrieval systems. The main purpose of stemming is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.
Eg: A stemmer for English should identify the strings "cats", "catlike", "catty" as based on the root "cat".
RULES OF SUFFIX STRIPPING STEMMERS:
1.If the word ends in 'ed', remove the 'ed'.
2.If the word ends in 'ing', remove the 'ing'.
3.If the word ends in 'ly', remove the 'ly'.
RULES OF SUFFIX SUBSTITUTION STEMMERS:
1.If the word ends in ‘ies’ substitute ‘ies’ with ‘y’.
Generally this stemmer is used because of some word like families etc

Tokenization: - Tokenization refers to splitting of text or words into small tokens. For example, in a paragraph, a line is a token. Similarly, in a line a word is a token. Tokenization is important because, by studying the words in a document, the meaning of the text can be easily deduced. There are different types of tokenization present such as word tokenization, line tokenization, regular expression tokenization etc.

FEATURE EXTRACTION

In Feature extraction, after identifying the key feature from the document, the data is reduced so that it can be cleaned and further be tested on various machine learning algorithms. There are various feature extraction methods. In this project, we have used the TFIDF vectorizer.

TFIDF vectorizer

TFIDF vectorizer is an abbreviation for Term Frequency and Inverse Document Frequency. It checks how significant a word is in the whole document. The term frequency function determines how often a term appears in the text. The inverse document frequency determines whether a word is uncommon or common across a document.

The TFIDF will thus check the authenticity. So, if a word occurs frequently in many documents like

‗what', ‗if' etc., they have the chances that they are fake, while the words that appear often in one text but not in all others have a good chance of being true.

4.2.4 MACHINE LEARNING ALGORITHMS

Naïve Bayes Algorithm:

Naïve Bayes Algorithm is a family of classification algorithms which works on the principle of Bayes Theorem. Therefore, it is also known as a collection of probabilistic classifiers and can be implemented in various classification tasks. In such an algorithm, all pairs of features which are classified are independent of each other. Some of its applications include filtering spam, sentiment prediction and classification of documents. Naïve Bayes holds great significance in this project when it comes to classifying a news article as real news or fake news since it is highly scalable, efficient and can be used to produce real-time predictions while handling continuous as well as discrete data.

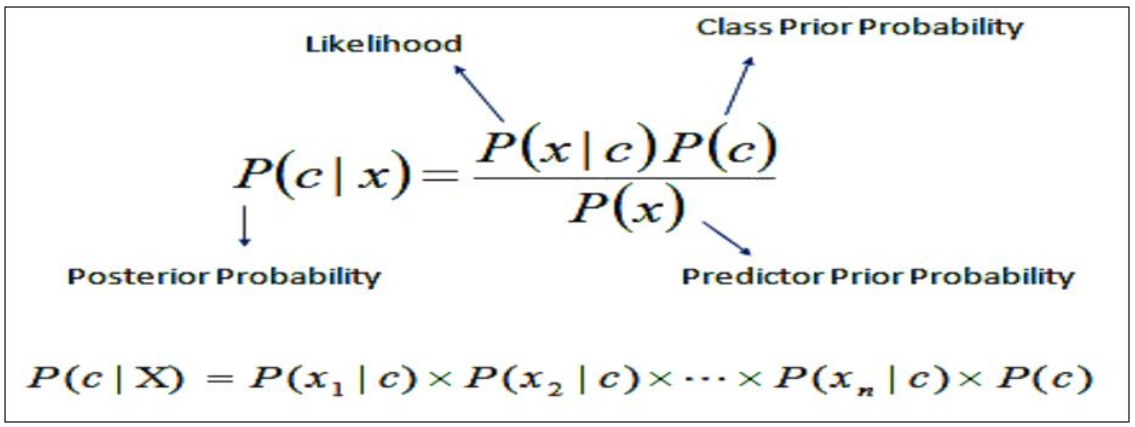
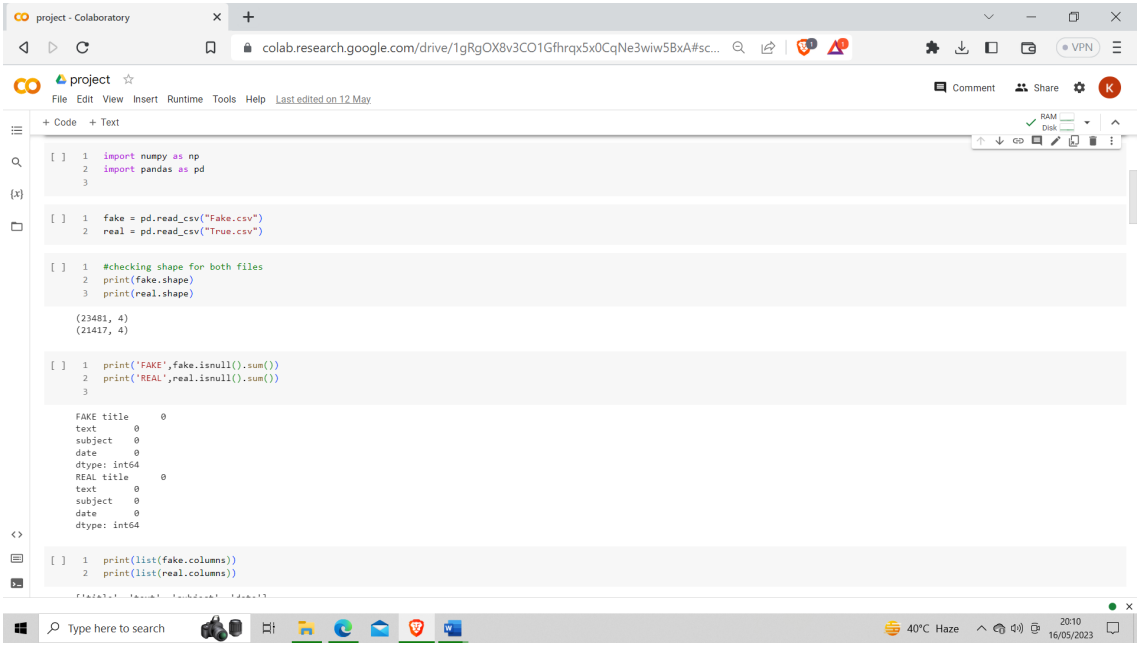


Fig 5. Naïve Bayes algorithm

4.2.5 OUTPUT PREDICTION

After the machine learning algorithms are done, the output will be predicted i.e., whether the news is real or fake. If the news is real the user will know the result and can know about the fact and be aware, but if the news turned out to be fake, the user will need to check the facts and stop the spread of fake news. For this, our next module real news generation comes into play.

CHAPTER 6: IMPLEMENTATION
Here is the code implementation



```
import numpy as np
import pandas as pd
```

```
fake = pd.read_csv("Fake.csv")
real = pd.read_csv("True.csv")
```

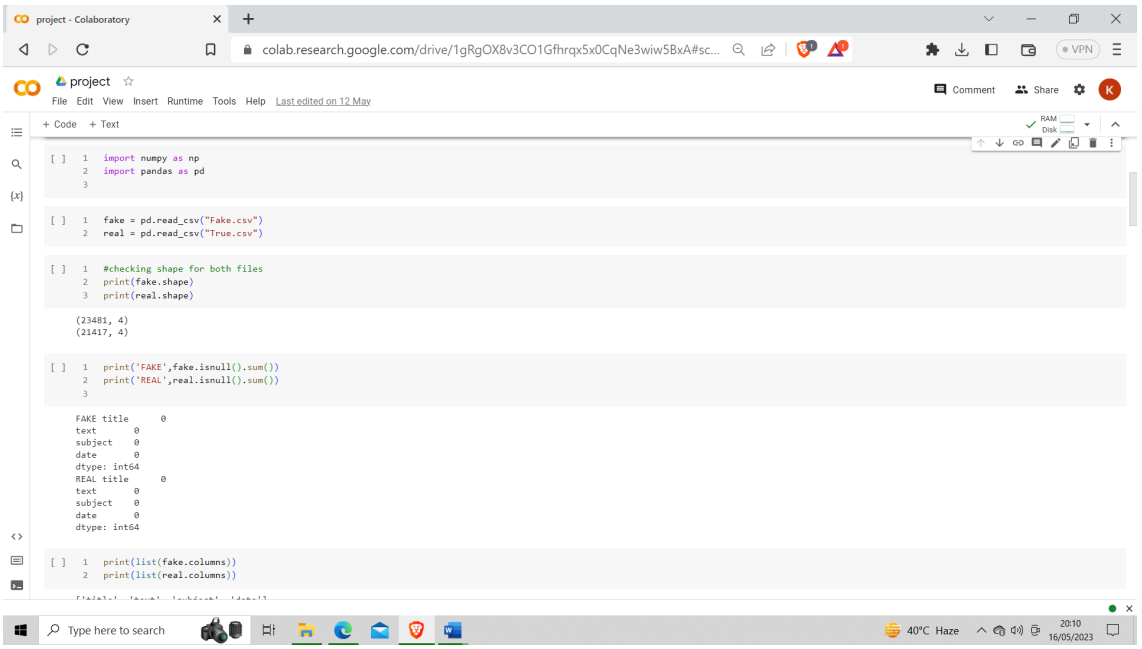
```
#checking shape for both files
print(fake.shape)
print(real.shape)
```

```
(23481, 4)
(21417, 4)
```

```
print('FAKE',fake.isnull().sum())
print('REAL',real.isnull().sum())
```

```
FAKE title      0
text       0
subject    0
date       0
dtype: int64
REAL title      0
text       0
subject    0
date       0
dtype: int64
```

```
print(list(fake.columns))
print(list(real.columns))
```




```
import numpy as np
import pandas as pd
```

```
fake = pd.read_csv("Fake.csv")
real = pd.read_csv("True.csv")
```

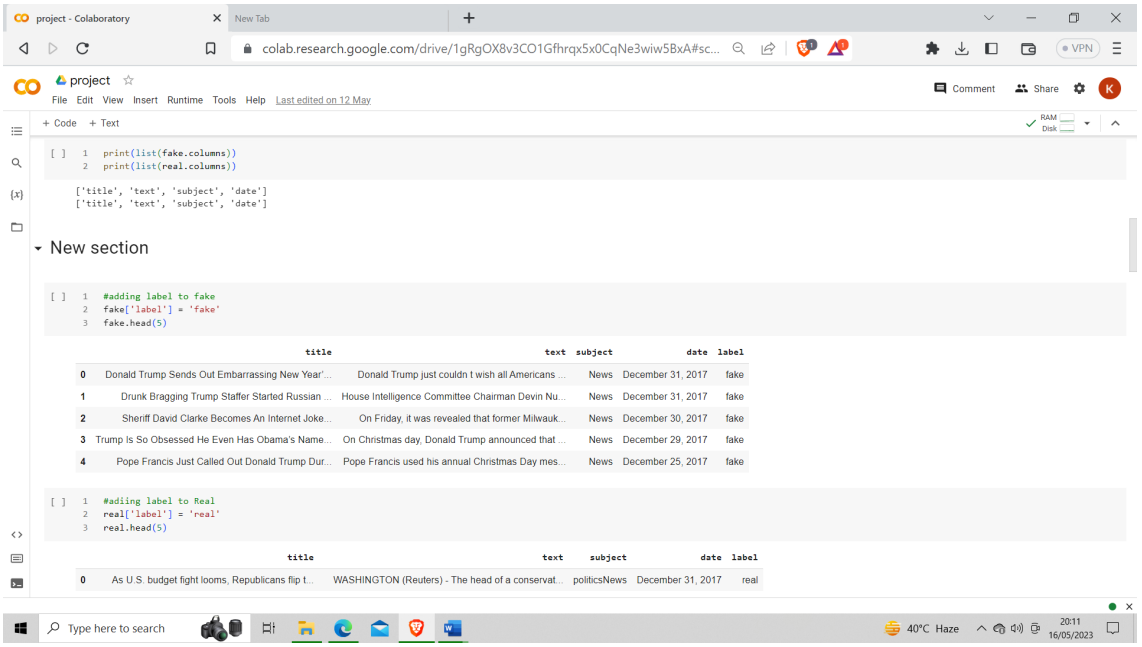
```
#checking shape for both files
print(fake.shape)
print(real.shape)
```

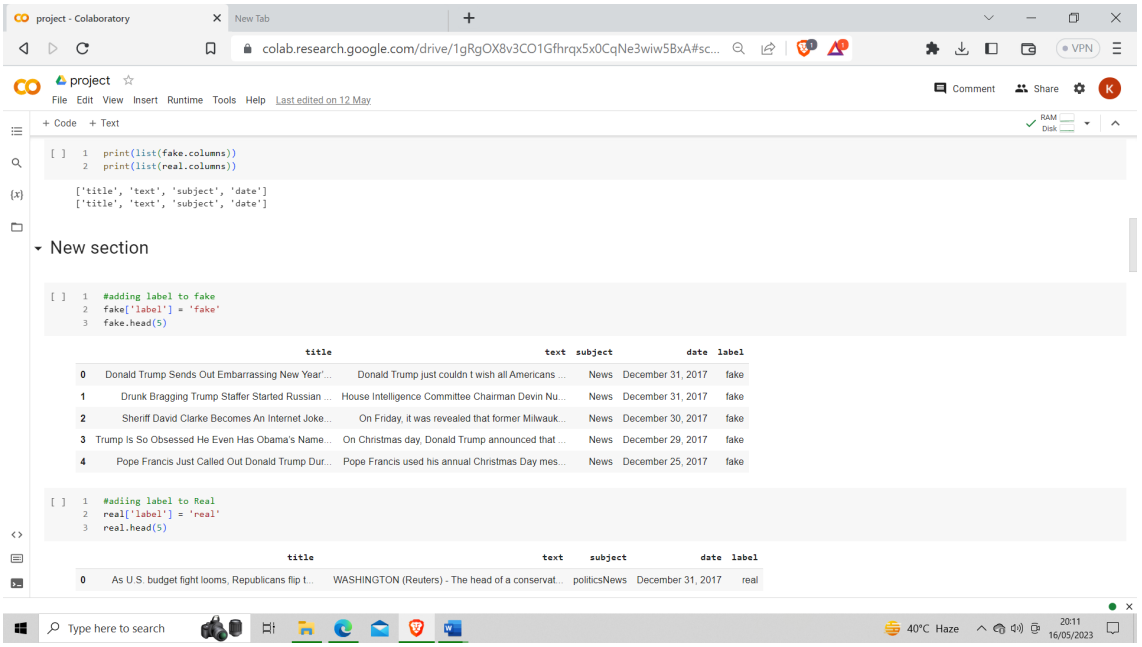
```
(23481, 4)
(21417, 4)
```

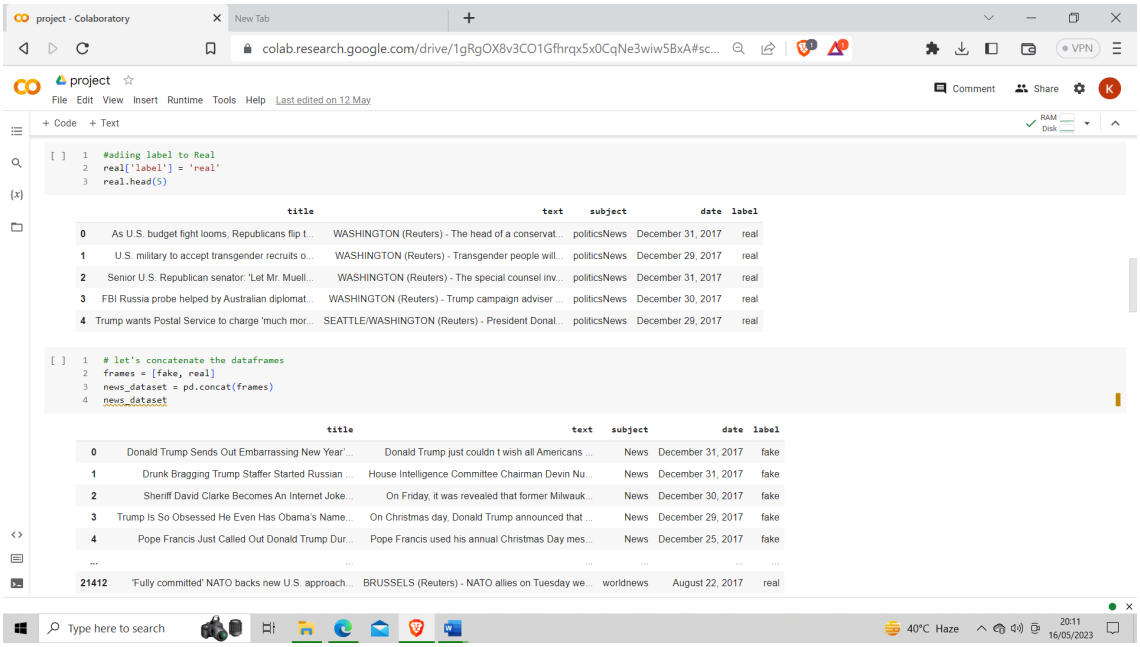
```
print('FAKE',fake.isnull().sum())
print('REAL',real.isnull().sum())
```

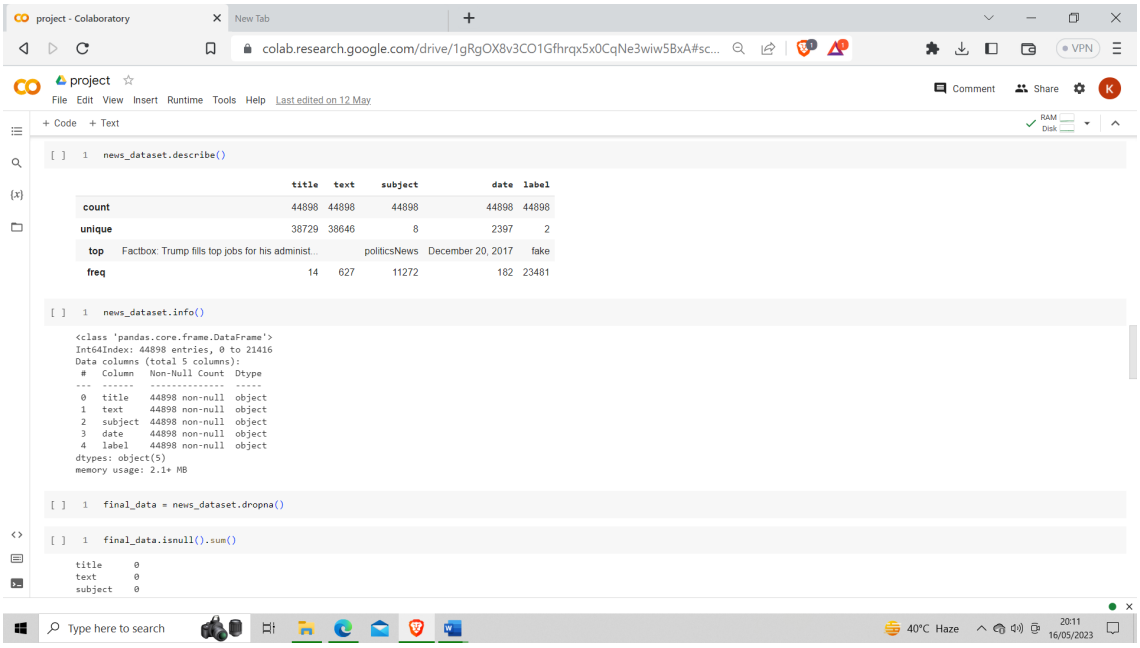
```
FAKE title      0
text       0
subject    0
date       0
dtype: int64
REAL title      0
text       0
subject    0
date       0
dtype: int64
```

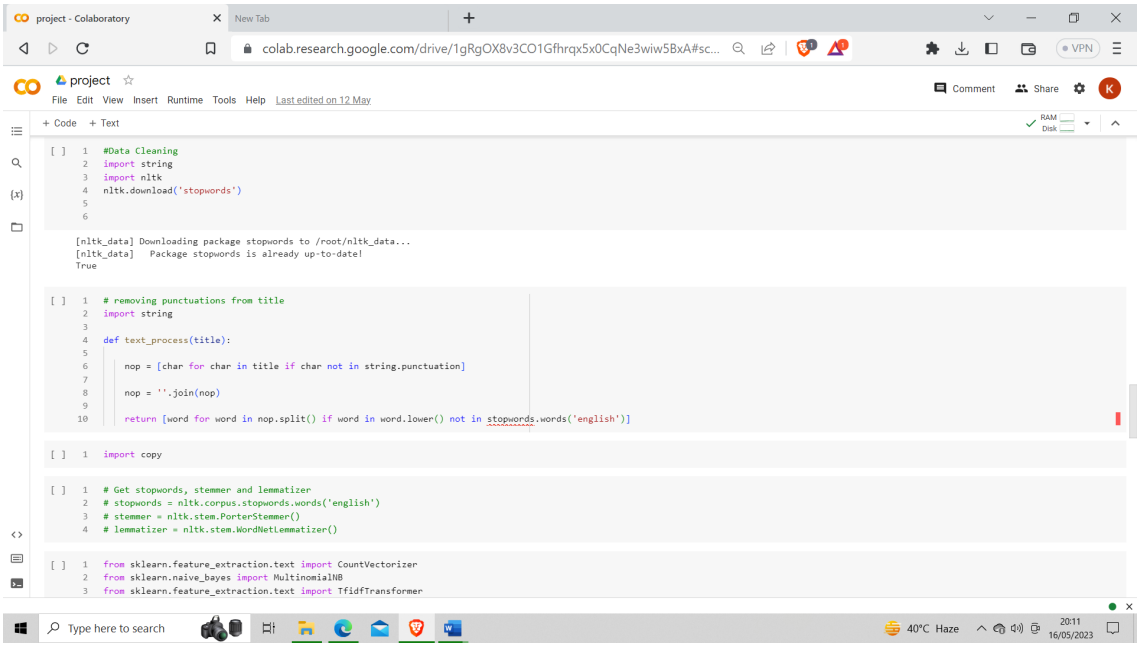
```
print(list(fake.columns))
print(list(real.columns))
```

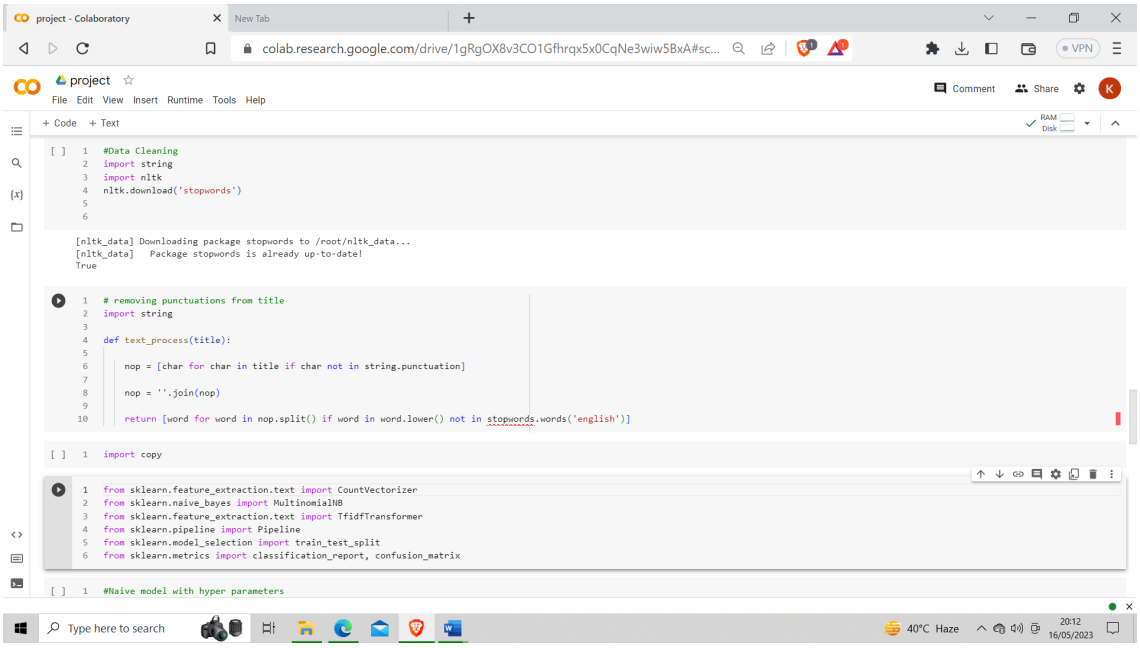

project - Colaboratory
New Tab
colab.research.google.com/drive/1gRgOX8v3CO1Gfhrqx5x0CqNe3wiw5BxA#sc...
project
File Edit View Insert Runtime Tools Help Last edited on 12 May
Comment
Share
+ Code + Text
[ ] 1 print(list(fake.columns))
2 print(list(real.columns))
['title', 'text', 'subject', 'date']
['title', 'text', 'subject', 'date']
New section
[ ] 1 #adding label to fake
2 fake['label'] = 'fake'
3 fake.head(5)
title text subject date label
0 Donald Trump Sends Out Embarrassing New Year'... Donald Trump just couldn t wish all Americans ... News December 31, 2017 fake
1 Drunk Bragging Trump Staffer Started Russian ... House Intelligence Committee Chairman Devin Nu... News December 31, 2017 fake
2 Sheriff David Clarke Becomes An Internet Joke... On Friday, it was revealed that former Milwauk... News December 30, 2017 fake
3 Trump Is So Obsessed He Even Has Obama's Name... On Christmas day, Donald Trump announced that ... News December 29, 2017 fake
4 Pope Francis Just Called Out Donald Trump Dur... Pope Francis used his annual Christmas Day mes... News December 25, 2017 fake
[ ] 1 #adiing label to Real
2 real['label'] = 'real'
3 real.head(5)
title text subject date label
0 As U.S. budget fight looms, Republicans flip t... WASHINGTON (Reuters) - The head of a conservat... politicsNews December 31, 2017 real
Type here to search
40°C Haze
20:11
16/05/2023


project - Colaboratory
New Tab
colab.research.google.com/drive/1gRgOX8v3CO1Gfhrqx5x0CqNe3wiw5BxA#sc...
project
File Edit View Insert Runtime Tools Help Last edited on 12 May
Comment
Share
+ Code + Text
[ ] 1 print(list(fake.columns))
2 print(list(real.columns))
['title', 'text', 'subject', 'date']
['title', 'text', 'subject', 'date']
New section
[ ] 1 #adding label to fake
2 fake['label'] = 'fake'
3 fake.head(5)
title text subject date label
0 Donald Trump Sends Out Embarrassing New Year'... Donald Trump just couldn t wish all Americans ... News December 31, 2017 fake
1 Drunk Bragging Trump Staffer Started Russian ... House Intelligence Committee Chairman Devin Nu... News December 31, 2017 fake
2 Sheriff David Clarke Becomes An Internet Joke... On Friday, it was revealed that former Milwauk... News December 30, 2017 fake
3 Trump Is So Obsessed He Even Has Obama's Name... On Christmas day, Donald Trump announced that ... News December 29, 2017 fake
4 Pope Francis Just Called Out Donald Trump Dur... Pope Francis used his annual Christmas Day mes... News December 25, 2017 fake
[ ] 1 #adiing label to Real
2 real['label'] = 'real'
3 real.head(5)
title text subject date label
0 As U.S. budget fight looms, Republicans flip t... WASHINGTON (Reuters) - The head of a conservat... politicsNews December 31, 2017 real
Type here to search
40°C Haze
20:11
16/05/2023


project - Colaboratory
New Tab
colab.research.google.com/drive/1gRgOX8v3CO1Gfhrqx5x0CqNe3wiw5BxA#sc...
project
File Edit View Insert Runtime Tools Help Last edited on 12 May
Comment
Share
+ Code + Text
RAM
Disk
[ ] 1 #adiing label to Real
2 real['label'] = 'real'
3 real.head(5)
title text subject date label
0 As U.S. budget fight looms, Republicans flip t... WASHINGTON (Reuters) - The head of a conservat... politicsNews December 31, 2017 real
1 U.S. military to accept transgender recruits o... WASHINGTON (Reuters) - Transgender people will... politicsNews December 29, 2017 real
2 Senior U.S. Republican senator: 'Let Mr. Muell... WASHINGTON (Reuters) - The special counsel inv... politicsNews December 31, 2017 real
3 FBI Russia probe helped by Australian diplomat... WASHINGTON (Reuters) - Trump campaign adviser ... politicsNews December 30, 2017 real
4 Trump wants Postal Service to charge 'much mor... SEATLE/WASHINGTON (Reuters) - President Donal... politicsNews December 29, 2017 real
[ ] 1 # let's concatenate the dataframes
2 frames = [fake, real]
3 news_dataset = pd.concat(frames)
4 news_dataset
title text subject date label
0 Donald Trump Sends Out Embarrassing New Year'... Donald Trump just couldn t wish all Americans ... News December 31, 2017 fake
1 Drunk Bragging Trump Staffer Started Russian ... House Intelligence Committee Chairman Devin Nu... News December 31, 2017 fake
2 Sheriff David Clarke Becomes An Internet Joke... On Friday, it was revealed that former Milwauk... News December 30, 2017 fake
3 Trump Is So Obsessed He Even Has Obama's Name... On Christmas day, Donald Trump announced that ... News December 29, 2017 fake
4 Pope Francis Just Called Out Donald Trump Dur... Pope Francis used his annual Christmas Day mes... News December 25, 2017 fake
... ...
21412 'Fully committed' NATO backs new U.S. approach... BRUSSELS (Reuters) - NATO allies on Tuesday we... worldnews August 22, 2017 real
Type here to search
40°C Haze
20:11
16/05/2023


project - Colaboratory
New Tab
colab.research.google.com/drive/1gRgOX8v3CO1Gfhrqx5x0CqNe3wiw5BxA#sc...
project
File Edit View Insert Runtime Tools Help Last edited on 12 May
Comment
Share
+ Code + Text
RAM
Disk
[ ] 1 news_dataset.describe()
title text subject date label
count 44898 44898 44898 44898 44898
unique 38729 38646 8 2397 2
top Factbox: Trump fills top jobs for his administ... politicsNews December 20, 2017 fake
freq 14 627 11272 182 23481
[ ] 1 news_dataset.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 44898 entries, 0 to 21416
Data columns (total 5 columns):
Column Non-Null Count Dtype
--- ------ -------------- -----
0 title 44898 non-null object
1 text 44898 non-null object
2 subject 44898 non-null object
3 date 44898 non-null object
4 label 44898 non-null object
dtypes: object(5)
memory usage: 2.1+ MB
[ ] 1 final_data = news_dataset.dropna()
[ ] 1 final_data.isnull().sum()
title 0
text 0
subject 0
Type here to search
40°C Haze
20:11
16/05/2023


project - Colaboratory
New Tab
colab.research.google.com/drive/1gRgOX8v3CO1Gfhrqx5x0CqNe3wiw5BxA#sc...
project
File Edit View Insert Runtime Tools Help Last edited on 12 May
Comment
Share
+ Code + Text
#Data Cleaning
import string
import nltk
nltk.download('stopwords')
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
True
# removing punctuations from title
import string
def text_process(title):
nop = [char for char in title if char not in string.punctuation]
nop = ''.join(nop)
return [word for word in nop.split() if word in word.lower() not in stopwords.words('english')]
import copy
# Get stopwords, stemmer and lemmatizer
# stopwords = nltk.corpus.stopwords.words('english')
# stemmer = nltk.stem.PorterStemmer()
# lemmatizer = nltk.stem.WordNetLemmatizer()
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import TfidfTransformer
Type here to search
40°C Haze
20:11
16/05/2023


project - Colaboratory
New Tab
colab.research.google.com/drive/1gRgOX8v3CO1Gfhrqx5x0CqNe3wiw5BxA#sc...
project
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
#Data Cleaning
import string
import nltk
nltk.download('stopwords')
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
True
# removing punctuations from title
import string
def text_process(title):
nop = [char for char in title if char not in string.punctuation]
nop = ''.join(nop)
return [word for word in nop.split() if word in word.lower() not in stopwords.words('english')]
import copy
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix
#Naive model with hyper parameters
Type here to search
40°C Haze
20:12
16/05/2023

project - Colaboratory

colab.research.google.com/drive/1gRgOX8v3CO1Gfhrqx5x0CqNe3wiw5BxA#sc...

project

File Edit View Insert Runtime Tools Help

Comment Share

+ Code + Text

```
#Naive model with hyper parameters
piplineTitle = Pipeline([
    ('bow', CountVectorizer(analyzer=text_process)),
    ('tfidf', TfidfTransformer()),
    ('classifier', MultinomialNB()),
])
```

```
X_train, X_test, y_train, y_test = train_test_split(final_data['title'], final_data['label'], test_size=0.2, random_state=123)
```

```
print(piplineTitle.fit(X_train, y_train))
```

```
Pipeline(steps=[('bow',
                 CountVectorizer(analyzer=<function text_process at 0x7ff110de2830>)),
                ('tfidf', TfidfTransformer()),
                ('classifier', MultinomialNB())])
```

```
y_pred = piplineTitle.fit(X_train, y_train).predict(X_test)
y_pred
```

```
array(['real', 'fake', 'real', ..., 'real', 'fake', 'fake'], dtype='<U4')
```

```
clf_report = classification_report(y_test, y_pred)
print('Classification_Report',clf_report)
```

```
Classification_Report              precision    recall  f1-score   support

        fake       0.97      0.99      0.98      4556
        real       0.99      0.97      0.98      4424

    accuracy                           0.98      8980
   macro avg       0.98      0.98      0.98      8980
weighted avg       0.98      0.98      0.98      8980
```

Type here to search

40°C Haze 20:12 16/05/2023

project - Colaboratory

colab.research.google.com/drive/1gRgOX8v3CO1Gfhrqx5x0CqNe3wiw5BxA#sc...

project

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

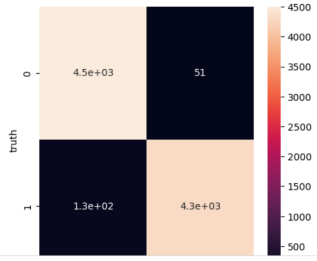
+ Code + Text

```
cnf_matrix = confusion_matrix(y_test, y_pred)
print('Cnfusion Matrix',cnf_matrix)
```

```
Cnfusion Matrix [[4505   51]
 [ 127 4297]]
```

```
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize = (15,5))
sns.heatmap(cnf_matrix, annot=True)
plt.xlabel('predicted')
plt.ylabel('truth')
```

```
Text(33.22222222222222, 0.5, 'truth')
```



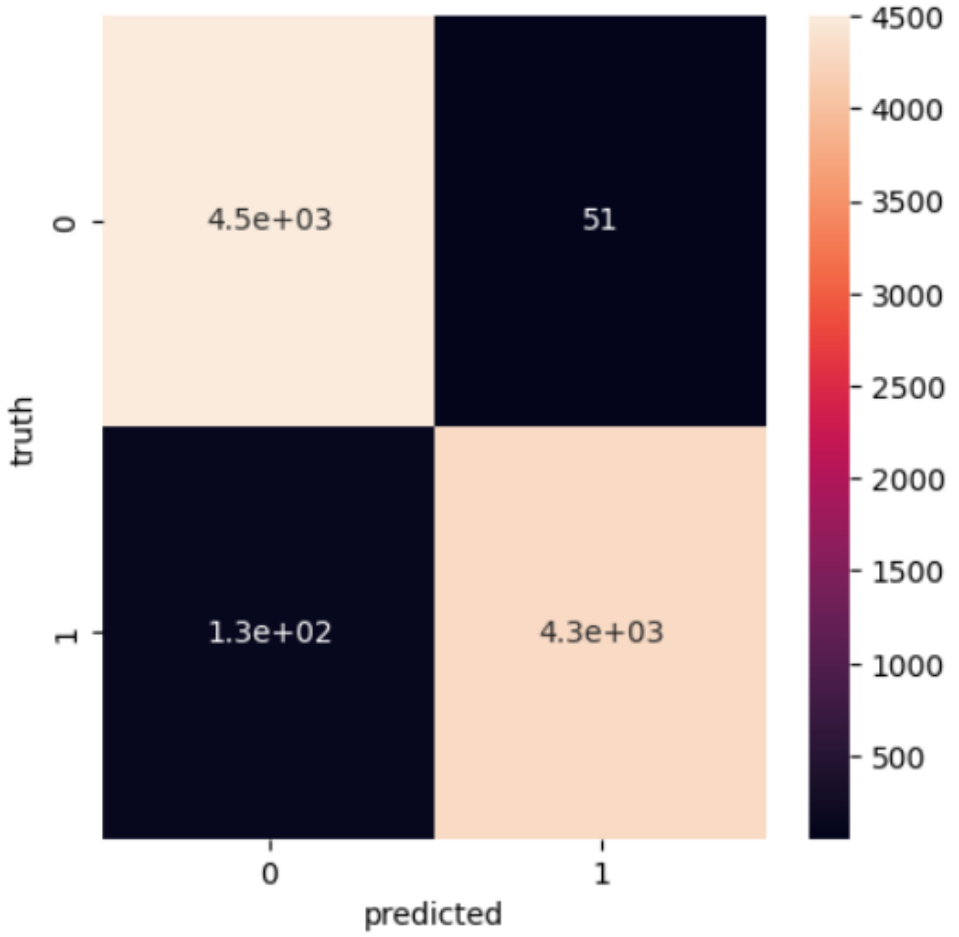
Type here to search

40°C Haze 20:12 16/05/2023

CHAPTER 7: RESULT ANALYSIS

Classification_Report

	Precision	recall	f1-score	support
fake	0.97	0.99	0.98	4556
real	0.99	0.97	0.98	4424
accuracy			0.98	8980
macro avg	0.98	0.98	0.98	8980
weighted avg	0.98	0.98	0.98	8980



CHAPTER 8: CONCLUSION, LIMITATION AND FUTURE SCOPE

CONCLUSION

The Naive Bayes classifier is a popular and widely used algorithm due to its simplicity, speed, and good performance in many text classification tasks. However, for more advanced NLP problems like fake news detection that require capturing complex linguistic relationships, other algorithms like recurrent neural networks (RNNs), convolutional neural networks (CNNs), or transformer-based models may be more suitable.

LIMITATION

limitations of the Naive Bayes classifier specifically in the context of NLP:

Independence assumption: The Naive Bayes classifier assumes that the features used for classification are independent of each other. However, in NLP, this assumption is often violated since words in a text are usually related and can have dependencies. For example, the occurrence of certain words or phrases in a sentence can affect the probability of occurrence of other words. This independence assumption can limit the classifier's ability to capture complex relationships between words or phrases.

Limited handling of out-of-vocabulary words: Naive Bayes classifiers typically rely on the vocabulary present in the training data. Out-of-vocabulary words, which are words not seen during training, pose a challenge for the classifier. These words are often treated as unknown or ignored, leading to potential loss of important information. While techniques like smoothing can help mitigate this issue to some extent, handling out-of-vocabulary words remains a limitation.

Difficulty with handling large feature spaces: In NLP tasks, the feature space can be vast due to the large number of words or n-grams in a corpus. Naive Bayes classifiers, especially the ones that rely on explicit probabilities, can encounter computational challenges when dealing with high-dimensional feature spaces. The probability estimates may become unreliable or require significant computational resources to calculate accurately.

FUTURE SCOPE

Because of its low cost, easy accessibility, and broad distribution, social media has improved the news consumption experience. However, it has rendered the average internet consumer susceptible to consuming news that has been skewed deliberately or inadvertently, which can have serious implications and put a person and society at risk. Thus, we focused on eliminating the problem of fake news from the root itself and also providing people with a consequent genuine news which will help them to gain knowledge and be aware of the facts. Based on our obtained results the following are the future directions for continuing the project:

a)we can use image,video as well as audio data
b)We could compare various other machine learning algorithms.
c) Testing the proposed method in this paper on a larger dataset to check for accuracy and problems associated.
d) Can improve the webpage by using animations and wallpapers and making it more attractive.
e) After the successful implementation and removing all problems, can try for making this in the form of a mobile app.

REFERENCE

[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.