PROJECT REPORT
ON

FAKE NEWS DETECTION USING MACHINE LEARNING

By Asim Maner

ABSTRACT

Indian politics suffered from a great set back due to fake news. Fake news is intentionally written to mislead the audience to believe the false propaganda, which makes it difficult to detect based on news content. The fake news has hindered the mindset of the common people. Due to this widespread of the fake news online it is the need of the hour to check the authenticity of the news. The spread of fake news has the potential for extremely negative impact on society. The proposed approach is to use machine learning to detect fake news. Using vectorisation of the news title and then analysing the tokens of words with our dataset. The dataset we are using is a predefined curated list of news with their property of being a fake news or not. Our goal is to develop a model that classifies a given article as either true or fake.

Key Words:

Fake News, Self Learning, Pattern Matching, Response Generation, Artificial Intelligence, Natural Language Processing, Context Free Grammar, Term Frequency Inverse Document Frequency, Stochastic Gradient Decent, Word2Vec.

INTRODUCTION

"Fake news" is a term that has come to mean different things to different people. At its core, we are defining "fake news" as those news stories that are false: the story itself is fabricated, with no verifiable facts, sources or quotes. Sometimes these stories may be propaganda that is intentionally designed to mislead the reader, or may be designed as "clickbait" written for economic incentives (the writer profits on the number of people who click on the story). In recent years, fake news stories have proliferated via social media, in part because they are so easily and quickly shared online.

In this paper I experiment the possibility to detect fake news based only on textual information by applying traditional machine learning techniques.

In order to work on fake news detection, it is important to understand what is fake news and how they are characterized. The first is characterization or what is fake news and the second is detection. In order to build detection models, it is need to start by characterization, indeed, it is need to understand what is fake news before trying to detect them.

1.1 Fundamental Theories

Fundamental human cognition and behaviour theories developed across various disciplines, such as social sciences and economics, provide invaluable insights for fake news analysis. These theories can introduce new opportunities for qualitative and quantitative studies of big fake news data. These theories can also facilitate building well-justified and explainable models for fake news detection and

intervention, which, to date, have been rarely available. We have conducted a comprehensive literature survey across various disciplines and have identified well-known theories that can be potentially used to study fake news. These theories are provided in Table 2 along with short descriptions, which are related to either (I) the news itself or (II) its spreaders.

I. News-related theories. News-related theories reveal the possible characteristics of fake news content compared to true news content. For instance, theories have implied that fake news potentially differs from the truth in terms of, e.g., writing style and quality, quantity such as word counts, and sentiments expressed. It should be noted that these theories, developed by forensic psychology, target deceptive statements or testimonies but not fake news, though these are similar concepts. Thus, one research opportunity is to verify whether these attributes are statistically distinguishable among disinformation, fake news, and the truth, in particular, using big fake news data.

II. User-related theories. User-related theories investigate the characteristics of users involved in fake news activities, e.g., posting, forwarding, liking, and commenting. Fake news, unlike information such as fake reviews, can “attract” both malicious and normal users. Malicious users spread fake news often intentionally and are driven by benefits. Some normal users (which we denote as vulnerable normal users) can frequently and unintentionally spread fake news without recognizing the falsehood. Such vulnerability psychologically stems from (i) social impacts and (ii) self-impact.

1.2Relevance of the project:

A fake news classification system using different feature extraction methods and different classification algorithms like MultinomialNB, Logistic , Decision Tree, Random Forest and Support Vector Machine , we are going to use it in predicting the news as fake or real. In order to create a real time application, the algorithm should be fed with the most recent data. Data is of different sizes so that should be properly cleaned to get better results. So, we are using different algorithms and feature extraction methods like Word embedding model to get the best result.

1.3 Problem Statement:

Our main aim of the project is to make a machine learning model, with the help of which news can be classified as fake or real with help of different machine learning classification algorithms, deep learning methods and text feature extraction methods for classifying news.

1.4 Objective:

To achieve our goal of developing machine learning model to classify news as fake or real, we need perform following tasks in the same order as stated.

-Data Collection and Analysis

-Preprocessing the data

-Text feature extraction

-Using different classification algorithms

-Finding the best classification algorithm and feature extraction method

-Classifying the news as fake or real.

-Deploying the model.

1.5 Scope of the project:

In future works, we intend to use highly sophisticated classifying approach, like deep learning with sentiment analysis also and consider many text features like publisher, URLs etc., which may increase the accuracy of the classification of news as fake or real. Automatic fake news detection may be done using the latest news and training the model regularly to get the best results. So this can be used as a filter to upload the news.

MARKET DEMAND

The market demand for **fake news detection using machine learning** is growing significantly, driven by the increasing volume of misinformation and disinformation circulating online. Several factors are contributing to this demand:

1. Rising Misinformation and Disinformation:

- **Misinformation** refers to false information spread without malicious intent, while **disinformation** is deliberately false content aimed at misleading or manipulating public opinion. The sheer scale of both types of content, especially on social media and digital platforms, has heightened the need for automated detection systems.
- **High-profile incidents**, such as election interference, health misinformation (e.g., COVID-19), and political propaganda, have amplified concerns over the credibility of news sources and the need for effective detection mechanisms.

2. Government Regulations and Corporate Initiatives:

- Governments across the globe are imposing stricter regulations and laws to counter the spread of false information. Companies, media outlets, and social media platforms are being pressured to take stronger measures, which is accelerating the adoption of AI-driven fake news detection tools.
- **Tech giants like Facebook, Twitter, and Google** have already invested heavily in machine learning algorithms and partnerships with fact-checking organizations to address the problem of misinformation.

3. Machine Learning and AI Capabilities:

- Machine learning (ML) and **Natural Language Processing (NLP)** have evolved considerably, allowing the creation of more sophisticated fake news detection algorithms. These algorithms can analyze vast amounts of data quickly, identify patterns, and assess the credibility of news articles, social media posts, and other content.
- **Sentiment analysis, topic modeling, and deep learning** are also being applied to analyze linguistic nuances, context, and sources to detect whether a piece of news is likely to be fake or misleading.

4. Business Applications:

- Many industries are impacted by fake news, including **advertising**, **media**, **financial services**, and **e-commerce**. In the financial sector, for instance, fake news can lead to stock market volatility and reputational damage. Machine learning solutions for fake news detection are increasingly in demand to mitigate these risks.
- **Advertising companies** are investing in fake news detection to ensure their brands are not associated with misleading content, as it can damage their reputation.

5. Consumer Demand for Trustworthy Information:

- With declining trust in traditional and digital news sources, there is a growing demand from consumers for platforms that can guarantee the authenticity of the information they provide. This has created a market for companies that can offer reliable fake news detection technology, either as a standalone service or integrated into broader content management systems.

6. Investment and Startups:

- The market is attracting investments in **startups** focused on fake news detection and **fact-checking technologies**. For example, companies such as **Logically.ai**, **Factmata**, and **Full Fact** use AI to detect false or misleading content in real-time.
- **Media monitoring** and **cybersecurity** companies are also investing in integrating fake news detection solutions to provide their clients with more comprehensive risk analysis and reputation management tools.

Market Forecast:

- The **global fake news detection market** is expected to grow significantly in the next few years. According to some industry reports, the fake news detection market size was valued at approximately **USD 324.2 million in 2022** and is expected to reach **USD 1,504.7 million by 2027**, growing at a compound annual growth rate (CAGR) of 28.5% from 2022 to 2027.
- Growth is driven by **increased awareness**, **rising adoption of AI-based tools**, and **corporate social responsibility** initiatives in technology and media sectors.

Challenges:

- **False positives/negatives**: Machine learning models may struggle to perfectly classify news, leading to either incorrectly labeling true information as false or failing to detect certain fake news.
- **Deepfake Technology**: The emergence of deepfake videos and images has made detection more challenging. ML models need to adapt to multimodal content (text, images, videos).
- **Ethical Concerns**: There is ongoing debate about the balance between controlling fake news and preserving freedom of speech, which could influence regulations and the development of these technologies.

LITERATURE SURVEY

Multinomial Naive Bayes Classification Model for Sentiment Analysis Muhammad Abbas proposed a Text categorization is the task of determining a document it belongs to a series of pre-specified class documents. The automatic classification scheme can greatly promote the classification process. Along with the rapid growth of information on the Internet, the classification of texts is a general and important research field trend in the search for information.

A Survey on Decision Tree Algorithm For Classification Brijain R Patel proposed a model that explained Now-a-days the data stored in a database and which is used for application is huge. This explosive growth in data and database has generated an urgent need for new techniques and tools that can intelligently automatically transform the processed data into useful information and knowledge. Hence data mining has become a research are with increasing importance. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classis or to predict future data trends. Such analysis can help provide us with better understanding of the data at large.

Random Forest Leo Breiman proposed a model that Significant improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class. In order to grow these ensembles, often random vectors are generated that govern the growth of each tree in the ensemble. An early example is bagging (Breiman [1996]), where to grow each tree a random selection (without replacement) is made from the examples in the training set.

Research on Text Classification for Identifying Fake News

We use a Multinomial NB Decision tree, Random forest and XGBoost classifier in this paper. The real and fake social news is classified by three models: TF-IDF, Word2vec, and Word2vec weighted by TF-IDF. The idea of Word2vec model is to use the average of all word vectors to represent the document. Among them, the calculation of word2vec is implemented by Gensim[11]. In this paper, we measure the effectiveness of the algorithm by the precision rate, recall rate and F1 score. The precision rate is the ratio of the document number that the classifier correctly judge them to belong to the category and the document number belonging to the category, which denoted by the letter P. The recall rate is the proportion of the document number that the classifier correctly judge them to be the category and the document number belonging to the category, denoted

by the letter R. The F1 score is the geometric mean of the recall rate (R) and the precision rate (P). It is a kind of indicator used to comprehensively evaluate the classification effect.

The three algorithms of TF-IDF, Word2vec and Word2vec weighted by TF-IDF have good performance in the field of real and fake news classification. The average F1 score of the three classification algorithms has reached more than 87%. Among them, the precision rate of fake news recognition based on word2vec is 4.31% higher than TF-IDF algorithm, and the average F1 score is 4.25% higher. This paper proposes a method that combines Word2vec and TF-IDF to the field of fake news identification, which achieves good performance. Comparing the precision rate of identifying real news, the Word2vec weighted by TF- IDF is 3.11% higher than Word2vec, the average value of P, R and F1 based on Word2vec weighted by TF-IDF are higher than TF-IDF and Word2vec. which verifies the effectiveness of Word2vec weighted by TF-IDF algorithm in the field of fake news identification.

They used three models to identify fake news. Both of the TF-IDF and Word2vec perform well in the field of fake news identification, but the Word2vec weighted by TF- IDF algorithm is most suitable for distinguishing between real and fake news.

SOFTWARE REQUIREMENT SPECIFICATION

Operating Environment

Operating System: Windows 10/11 Linux

Processor: Intel i5 3.5 GHz or higher

RAM: 4GB or higher

Graphics: Nvidia

Software Interfaces

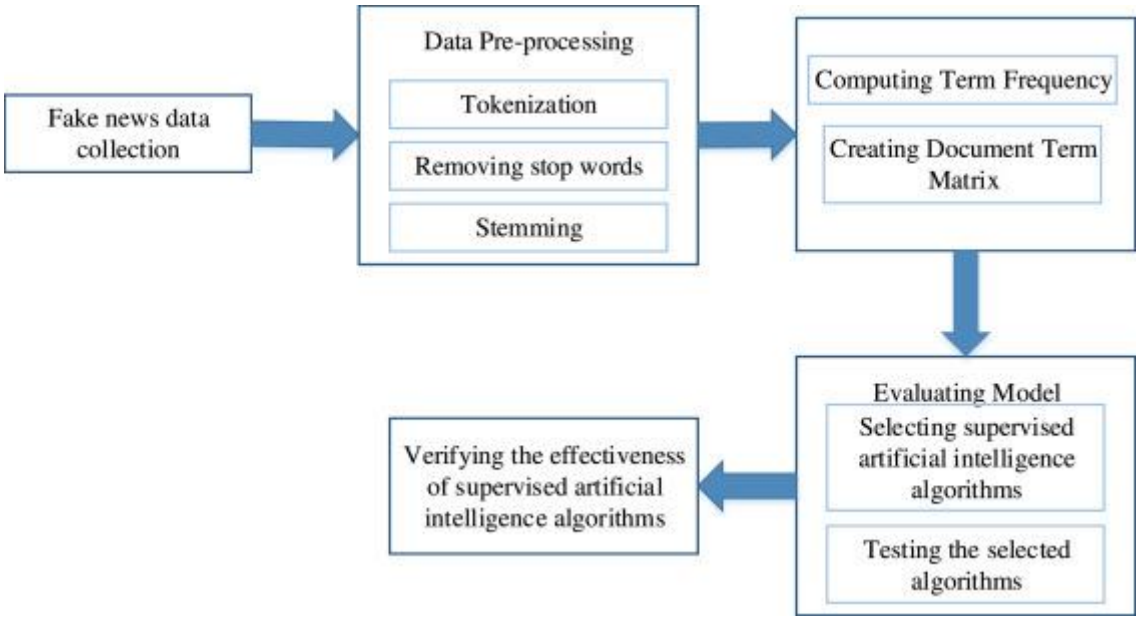
Jupyter Notebook , Google Colab

ARCHITECTURE

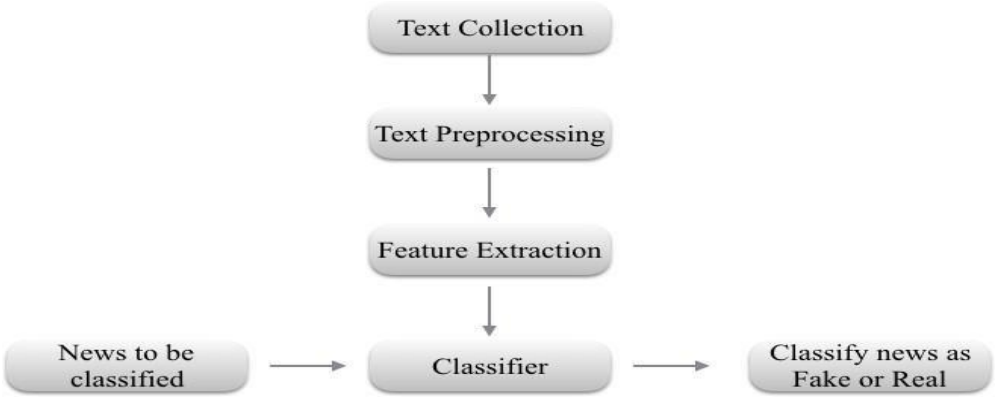
METHODOLOGY

4.1 Proposed Framework

In my proposed framework, I am expanding on the current literature by introducing ensemble techniques with various linguistic feature sets to classify news articles from multiple domains as true or fake. The ensemble techniques along with Linguistic Inquiry and Word Count (LIWC) feature set used in this research are the novelty of our proposed approach.



There are numerous reputed websites that post legitimate news contents which are used for fact checking. In addition, there are open repositories which are maintained by researchers to keep an up-to-date list of currently available datasets and hyperlinks to potential fact checking sites that may help in countering false news spread. However, we selected three datasets for our experiments which contain news from multiple domains (such as politics, entertainment, technology, and sports) and contain a mix of both truthful and fake articles, and merged the three datasets into large dataset. The datasets are available online and are extracted from Kaggle.



Text Collection

The dataset was taken from Kaggle. There was two csv files True.csv and Fake.csv. Fake dataset file was 23481 and True dataset file was 21417 text news. This two dataset then concatenated into one data set which contains 'text' and 'class' columns.

Text Preprocessing

After a text is obtained, we start with text preprocessing. Text preprocessing includes:
Converting all letters to lower case
Removing numbers

Removing punctuations, accent marks Removing white spaces
Removing stop words
Lemmatizing the word

Feature Extraction

Text needs to be converted into numbers before it is used with a machine learning algorithm. For classification of documents, documents are taken as input and a class label is generated as output by the predictive algorithm. The documents need to be converted into fixed-length vectors of numbers for the algorithm to take them as input. The input for the machine learning algorithm are the words encoded as integers or floating point values.

Bag of Words (BOW)

We make the list of unique words in the text corpus called vocabulary. Then we can represent each sentence or document as a vector with each word represented as 1 for present and 0 for absent from the vocabulary.

Count Vectorizer

Count Vectorizer generates an encoded vector that contains the length of the entire vocabulary coupled with the frequency of each word by which it appears in the document.

Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency (TF) = (Number of times term t appears in a document)/(Number of terms in the document)

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Inverse Document Frequency (IDF) = log(N/n), where, N is the number of documents and n is the number of documents a term t has appeared in. The IDF of a rare word is high, whereas the IDF of a frequent word is likely to be low. Thus having the effect of highlighting words that are distinct.

$$idf(w) = log(\frac{N}{df_t})$$

We calculate TF-IDF value of a term as = TF * IDF.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Word Embedding

It is a representation of text where words that have the same meaning have a similar representation. In other words it represents words in a coordinate system where related words, based on a corpus of relationships, are placed closer together. Word embeddings are in fact a class of techniques where individual words are represented as real-valued vectors in a predefined vector space. Each word is mapped to one vector. Each word is represented by a real-valued vector.

Classifier

The feature vectors are send to the classifier to classify the news as fake or not.

4.2. Algorithms

We used the following learning algorithms in conjunction with our proposed methodology to evaluate the performance of fake news detection classifiers.

4.2.1. Naïve Bayes

Naive Bayes is a probabilistic classifier inspired by the Bayes theorem under a simple assumption which is the attributes are conditionally independent.

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i)$$

The classification is conducted by deriving the maximum posterior which is the maximal P(Ci|**X**) with the above assumption applying to Bayes theorem. This assumption greatly reduces the computational cost by only counting the class distribution. Even though the assumption is not valid in most cases since the attributes are dependent, surprisingly Naive Bayes has able to perform impressively.

4.2.2. Logistic Regression

Logistic regression is named for the function used at the core of the method, the logistic function.

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the

environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modelled is a binary value (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^ (b0 + b1*x) / (1 + e^ (b0 + b1*x))

$$\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

4.2.3. Support Vector Machine (SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

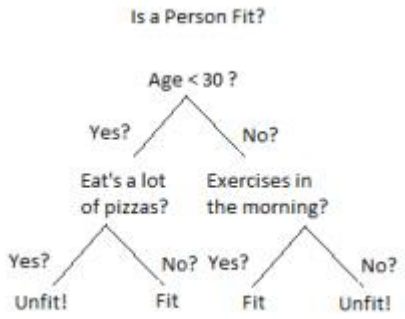
SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

$$h(\boldsymbol{x}_i) = sign(\ \sum_{j=1}^{s} \alpha_j\ y_j\ K(\boldsymbol{x}_j, \boldsymbol{x}_i) + b\ )$$

$$K(\boldsymbol{v}, \boldsymbol{v}') = \exp(\frac{||\boldsymbol{v} - \boldsymbol{v}'||^2}{2\gamma^2})$$

4.2.4. Decision Tree Learning

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.



4.2.5. Random Forest

Random forest (RF) is an advanced form of decision trees (DT) which is also a supervised learning model. RF consists of large number of decision trees working individually to predict an outcome of a class where the final prediction is based on a class that received majority votes. The error rate is low in random forest as compared to other models, due to low correlation among trees. Our random forest model was trained using different parameters; i.e., different numbers of estimators were used in a grid search to produce the best model that can predict the outcome with high accuracy. There are multiple algorithms to decide a split in a decision tree based on the problem of regression or classification. For the classification problem, we have used the Gini index as a cost function to estimate a split in the dataset.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Why Analysis?

Machine Learning is one of the most popular sub-fields of Artificial Intelligence. Machine learning concepts are used almost everywhere, such as Healthcare, Finance, Infrastructure, Marketing, Selfdriving cars, recommendation systems, chatbots, social sites, gaming, cyber security, and many more. Currently, Machine Learning is under the development phase, and many new technologies are continuously being added to Machine Learning. It helps us in many ways, such as analyzing large chunks of data, data extractions, interpretations, etc. Hence, there are unlimited numbers of uses of Machine Learning. In this topic, we will discuss various importance of Machine Learning with examples. So, let's start with a quick introduction to Machine Learning. Machine Learning is a branch of Artificial Intelligence that allows machines to learn and improve from experience automatically. It is defined as the field of study that gives computers the capability to learn without being explicitly programmed. It is quite different than traditional programming.

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies. Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semisupervised learning and reinforcement learning. The type of algorithm data scientists choose to use depends on what type of data they want to predict.

1. **Supervised learning:** In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

2. **Unsupervised learning:** This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.

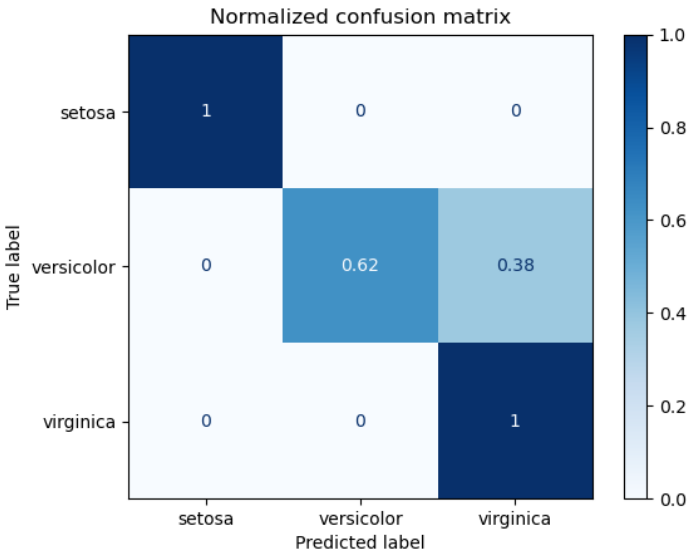
3. **Semi-supervised learning:** This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set. **4 .Reinforcement learning:** Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

4.3. Datasets

The datasets used in this study are open source and freely available online. The data includes both fake and truthful news articles from multiple domains. The truthful news articles published contain true description of real-world events, while the fake news websites contain claims that are not aligned with facts.

4.4. Performance Metrics

To evaluate the performance of the algorithms, I used confusion matrix. Confusion matrix is a tabular representation of a classification model performance on the test set, which consists of four parameters: true positive, false positive, true negative, and false negative.



MODELS IMPLEMENTATIONS

Multinomial Naive Bayes

Naive Bayes

```
dct = dict()

from sklearn.naive_bayes import MultinomialNB

NB_classifier = MultinomialNB()
pipe = Pipeline([('vect', CountVectorizer()),
                 ('tfidf', TfidfTransformer()),
                 ('model', NB_classifier)])

model = pipe.fit(X_train, y_train)
prediction = model.predict(X_test)
print("accuracy: {}%".format(round(accuracy_score(y_test, prediction)*100,2)))

dct['Naive Bayes'] = round(accuracy_score(y_test, prediction)*100,2)
```

```
accuracy: 80.43%
```

Decision Tree

Random Forest

```
from sklearn.ensemble import RandomForestClassifier

pipe = Pipeline([('vect', CountVectorizer()),
                 ('tfidf', TfidfTransformer()),
                 ('model', RandomForestClassifier(n_estimators=50, criterion="entropy"))])

model = pipe.fit(X_train, y_train)
prediction = model.predict(X_test)
print("accuracy: {}%".format(round(accuracy_score(y_test, prediction)*100,2)))
dct['Random Forest'] = round(accuracy_score(y_test, prediction)*100,2)
```

Decision Tree

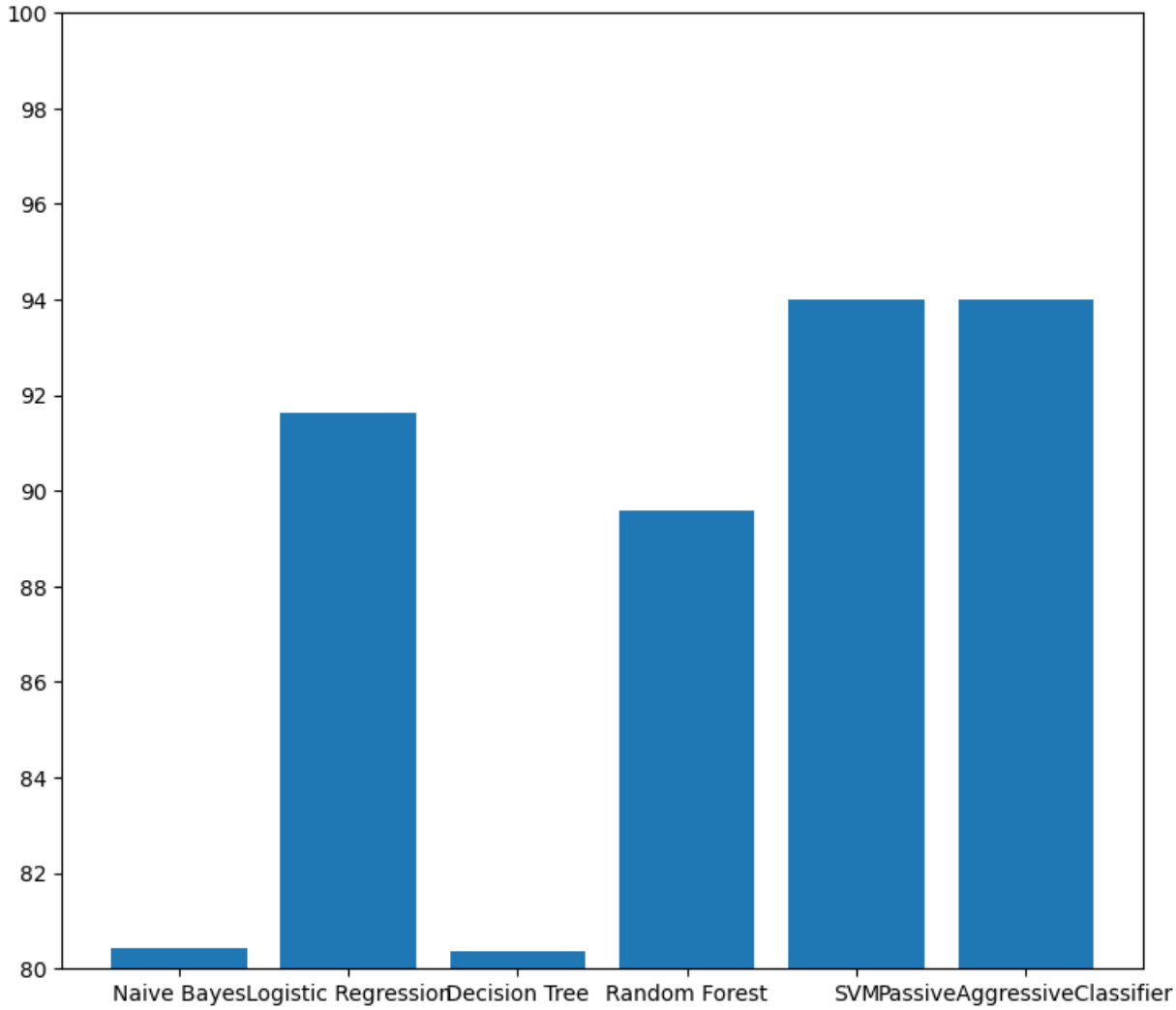
```
from sklearn.tree import DecisionTreeClassifier

# Vectorizing and applying TF-IDF
pipe = Pipeline([('vect', CountVectorizer()),
                 ('tfidf', TfidfTransformer()),
                 ('model', DecisionTreeClassifier(criterion= 'entropy',
                                                  max_depth = 20,
                                                  splitter='best',
                                                  random_state=42))])
# Fitting the model
model = pipe.fit(X_train, y_train)

# Accuracy
prediction = model.predict(X_test)
print("accuracy: {}%".format(round(accuracy_score(y_test, prediction)*100,2)))
dct['Decision Tree'] = round(accuracy_score(y_test, prediction)*100,2)
```

```
accuracy: 80.35%
```

Model Comparision



PassiveAgressiveClassifer performed better with accuracy 94.16%

BUSINESS OPPORTUNITIES

The business opportunities in fake news detection using machine learning (ML) are growing as misinformation and disinformation continue to pose challenges for various sectors. Here are some key business opportunities in this area:

1. Media & Journalism

- **Fake News Detection Services**: ML-powered platforms can offer real-time fact-checking and fake news detection services to media outlets. These tools help journalists verify the authenticity of news stories and protect the credibility of their reporting.
- **Content Validation Software**: Media companies can integrate AI solutions that scan articles, videos, and social media posts for misinformation, helping reduce the spread of false information.
- **Third-Party Verification**: Agencies or startups could specialize in fact-checking services and offer B2B solutions to media companies looking to enhance their content verification processes.

2. Social Media Platforms

- **Automated Monitoring Tools**: Social media companies can leverage ML to identify and flag fake news automatically. These tools can reduce the manual effort required to moderate content and ensure that platforms maintain a high level of trust among users.
- **Content Filtering for Advertisers**: ML-driven fake news detection can help advertisers avoid placing ads alongside false or controversial content, protecting their brand reputation.

3. Government & Regulatory Agencies

- **Public Policy & Regulatory Compliance**: Governments are increasingly focused on combating disinformation, especially during elections or public health crises. ML solutions for fake news detection can be marketed to governments for improving oversight and regulation of digital content.
- **Law Enforcement**: Agencies focused on cybercrime and online fraud can utilize fake news detection tools to monitor online threats, misinformation campaigns, and foreign interference.

4. Education & Research

- **Educational Tools**: ML-powered platforms can be used to develop educational tools that teach students and the general public how to identify fake news. Companies or startups can create software to enhance media literacy programs in schools, universities, and NGOs.

- **Research & Data Analytics**: ML tools can gather and analyze large datasets of news articles and social media posts to study patterns in misinformation, offering valuable insights for academia, governments, and private companies.

5. Cybersecurity

- **Digital Risk Management**: Fake news detection can be integrated into cybersecurity offerings, protecting businesses from reputational damage, financial scams, or data breaches that result from disinformation.
- **Enterprise Solutions**: Enterprises can use ML models to monitor social channels and media outlets for fake news that could damage their brand, enabling swift damage control.

6. E-Commerce & Online Marketplaces

- **Product and Brand Protection**: E-commerce platforms can benefit from ML systems that detect fake reviews, counterfeit product listings, and misleading promotional content, safeguarding both consumers and brands.
- **Consumer Trust Tools**: Companies can market AI-driven tools to help users distinguish between authentic and fraudulent product reviews or advertisements, enhancing the integrity of their platforms.

7. Legal & Compliance

- **Defamation & Libel Cases**: Law firms specializing in defamation can use fake news detection software to gather evidence in legal cases involving false information spread online.
- **Corporate Compliance**: ML-based fake news detection can help corporations ensure that their communications and marketing materials comply with regulations, avoiding legal and reputational risks.

8. Advertising & Marketing

- **Content Verification for Advertisers**: Advertisers and marketers can use ML-based solutions to verify the authenticity of news sources and ensure their campaigns run on reputable websites.
- **Brand Safety Solutions**: Companies can offer brand safety services to advertisers, ensuring that their ads are not shown alongside misleading or false content.

9. Public Health

- **Health Information Verification**: With the rise of fake news in healthcare, especially during events like the COVID-19 pandemic, there is a growing demand for solutions that verify the authenticity of health-related information. ML models can detect and flag misinformation about treatments, vaccines, or public health guidelines.
- **Pharmaceutical Companies**: Pharma companies can use ML-powered tools to monitor for false claims or harmful misinformation about their products and respond swiftly to protect their brand reputation.

10. Content Moderation Services

- **Third-Party Moderation Services**: Startups can offer fake news detection as a third-party service to online platforms that lack the resources or infrastructure to build their own moderation systems. This could be useful for forums, niche social platforms, and independent news websites.

By capitalizing on these opportunities, companies can address a pressing global issue while simultaneously tapping into a growing market.

BUSINEES MODELLING

A fake news detection company, **CredCheck**, targets media companies, social media platforms, and financial firms. They offer a **SaaS model** where companies pay a monthly subscription to access their API, which scans millions of articles/posts and flags potential misinformation.

CredCheck also partners with government institutions to monitor and report on election-related misinformation, and they sell custom-built solutions to large corporations worried about brand reputation. Revenue streams include subscriptions, data licensing, and custom consulting projects. Their costs involve development, cloud infrastructure, and ongoing model improvements.

FINANICIAL MODELLING

Monthly Profit=Subscription cost*Paid Customer Count- (Maintenance, Production, Training Cost)

Average Yearly Profit=Average Monthly Profit *12

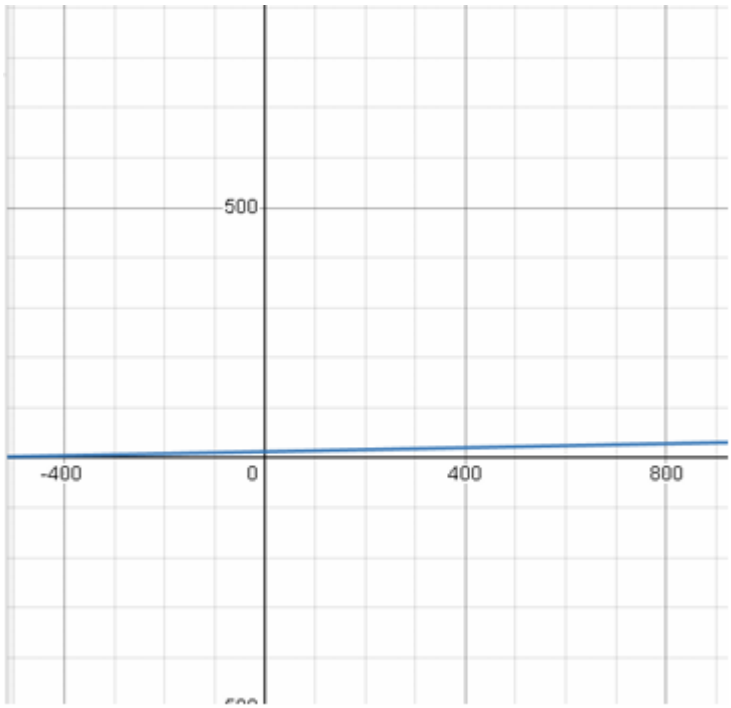
Assuming Profit=x, Subscription Cost=50, Customer Count=y and Cost to Produce, Maintain=599

Equation would be

22

x=50y-599

Graph would look like this



So in above conditions there must be a minimum 12 customers in order to start earning some profit.

The graph shows linear growth rate of profit with respect to customer count.

*Price of Subscription and Cost to Produce are still flexible and prone to change

Cost to Produce includes:

Training Cost,Maintenance Cost,Data Security,Marketing,etc.

RESULT

In this project accuracy, precision, recall, f1-Score is measured for classification of news as fake or real.

Confusion Matrix:

It is a common way of presenting true positive (tp), true negative (tn), false positive (fp) and false negative (fn) predictions. Those values are presented in the form of a matrix where the Y-axis shows the true classes while the X-axis shows the predicted classes.

Accuracy:

It measures how many observations, both positive and negative, were correctly classified.
Accuracy = (TP + TN) / (TP + TN + FP + FN)

CLASSIFER	ACCURACY
Naïve Bayes	80.43%
Logistic Regression	91.63%
Decision Tree	80.35%
Random Forest	89.58%
Support Vector Machine (SVM)	94%
Passive Aggressive	94.16%

CONCLUSION

In this project three different feature extraction methods like Count Vectorizer,TF- IDF Vectorizer,Word Embedding has been used. Decision Tree, Random Forest have been used to classify the news as fake or real.

Decision tree and Random forest with Count Vectorizer and take more time to train. By using the classification algorithms we got highest accuracy with algorithm and with TF-IDF feature extraction with 0.94 accuracy.Even though we got the same accuracy.

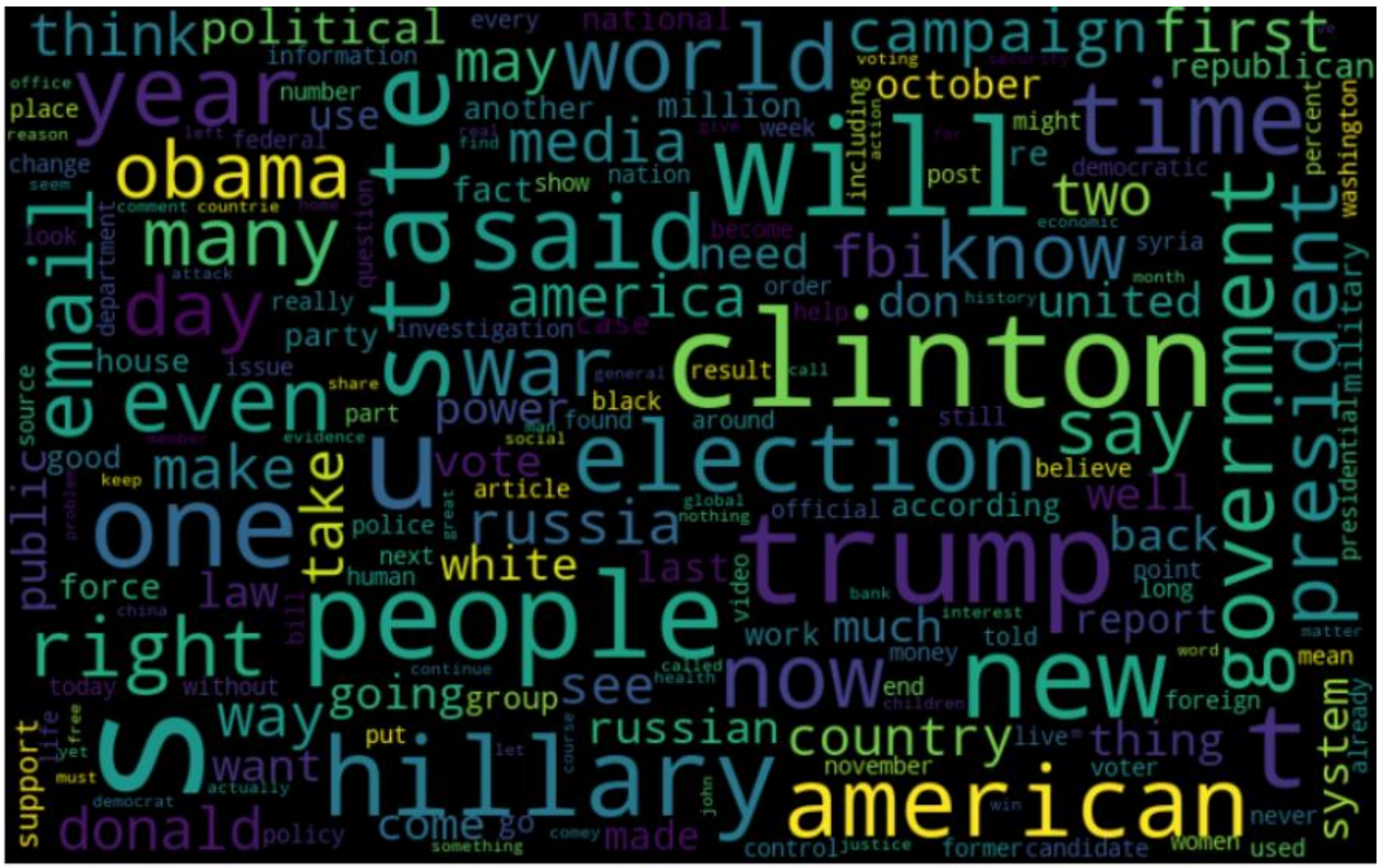
FUTURE SCOPE

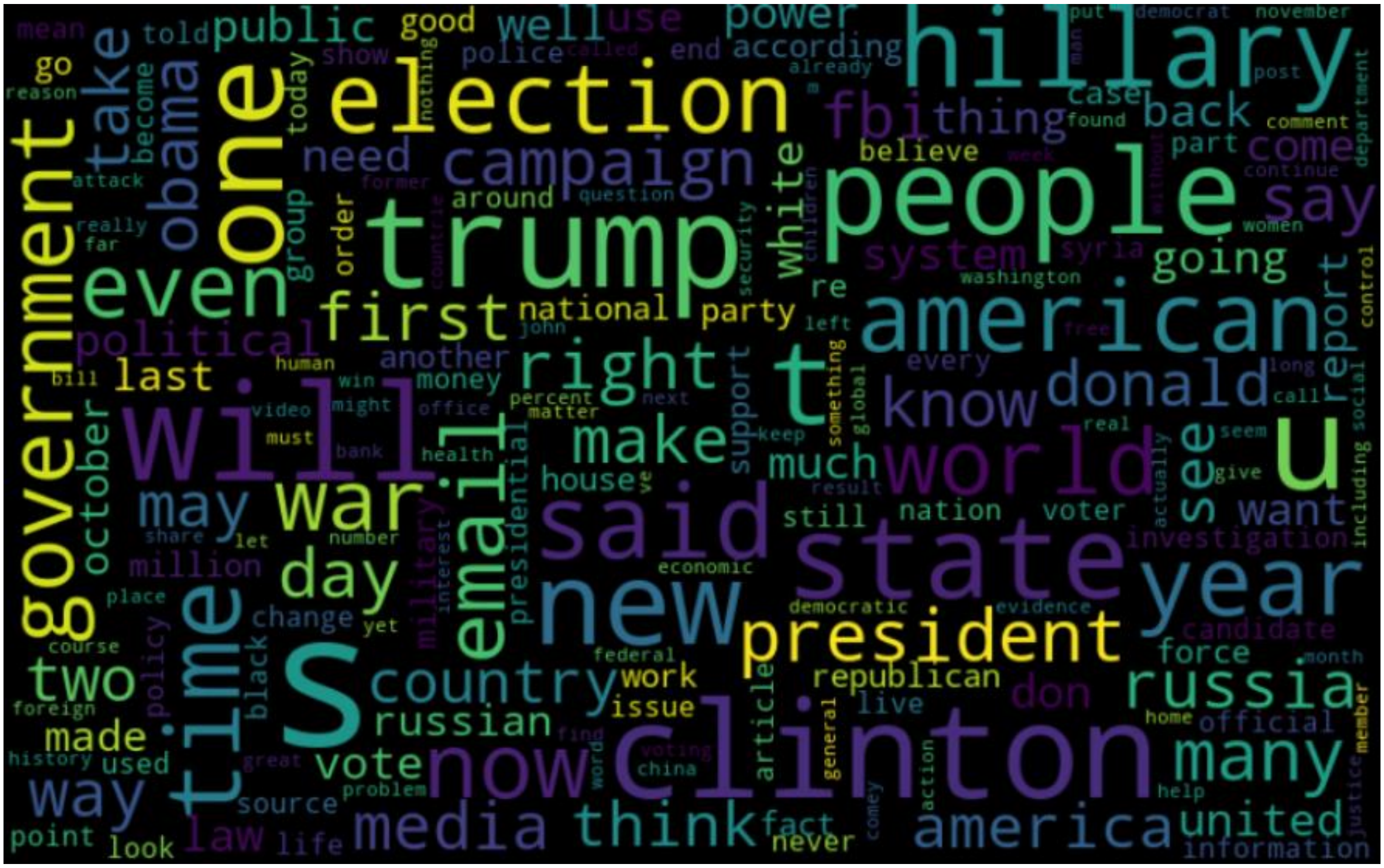
In future we can also use deep learning methods and sentiment analysis to classify the news as fake or real which may get high accuracy and we can extract further useful text like publication of the news, url domain etc.,

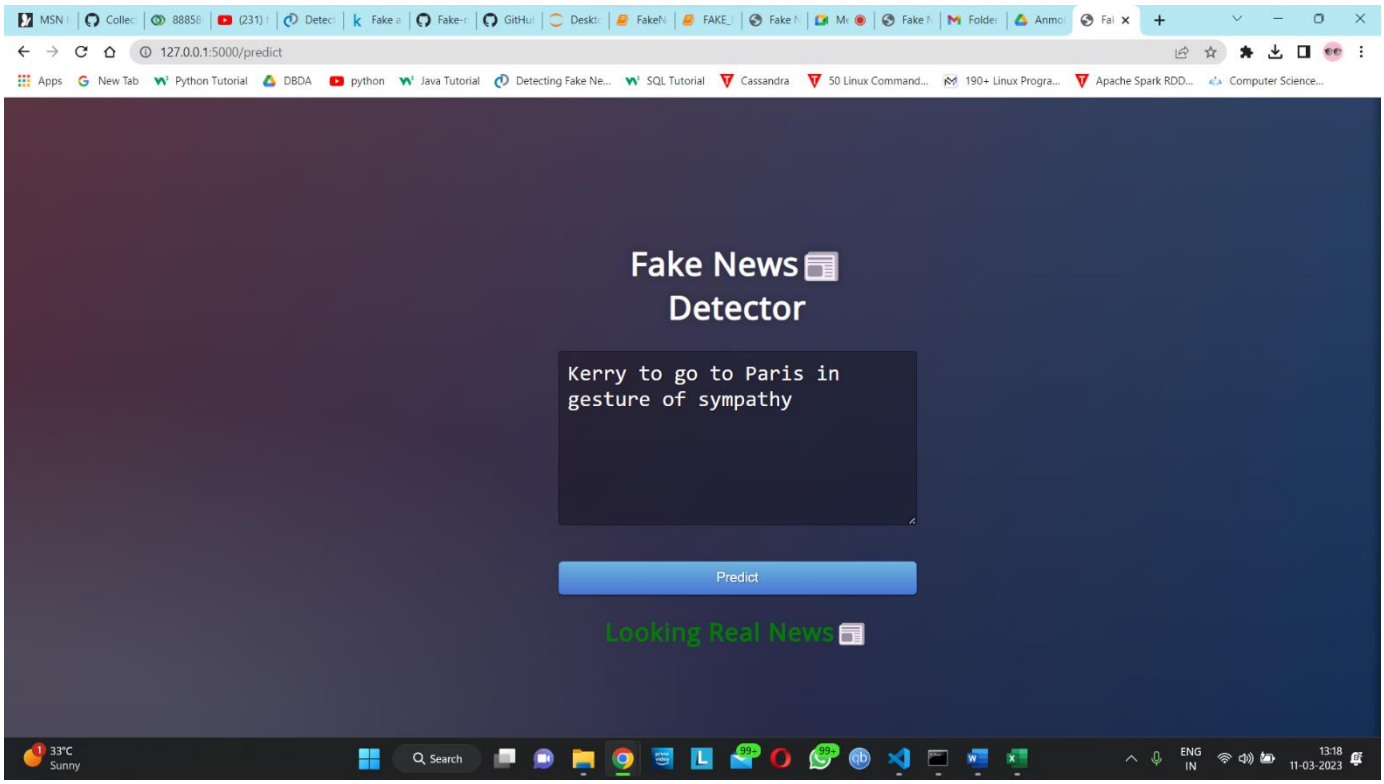
We can use more data for training purposes - In machine learning problems usually availability of more data significantly improves the performance of a learning algorithm. The dataset, which we used in this project contains only around 28000 articles. This number is quite small, and a dataset with larger number of news articles from different sources would be of a great help for the learning process as news from different sources will involve larger vocabulary and greater content.

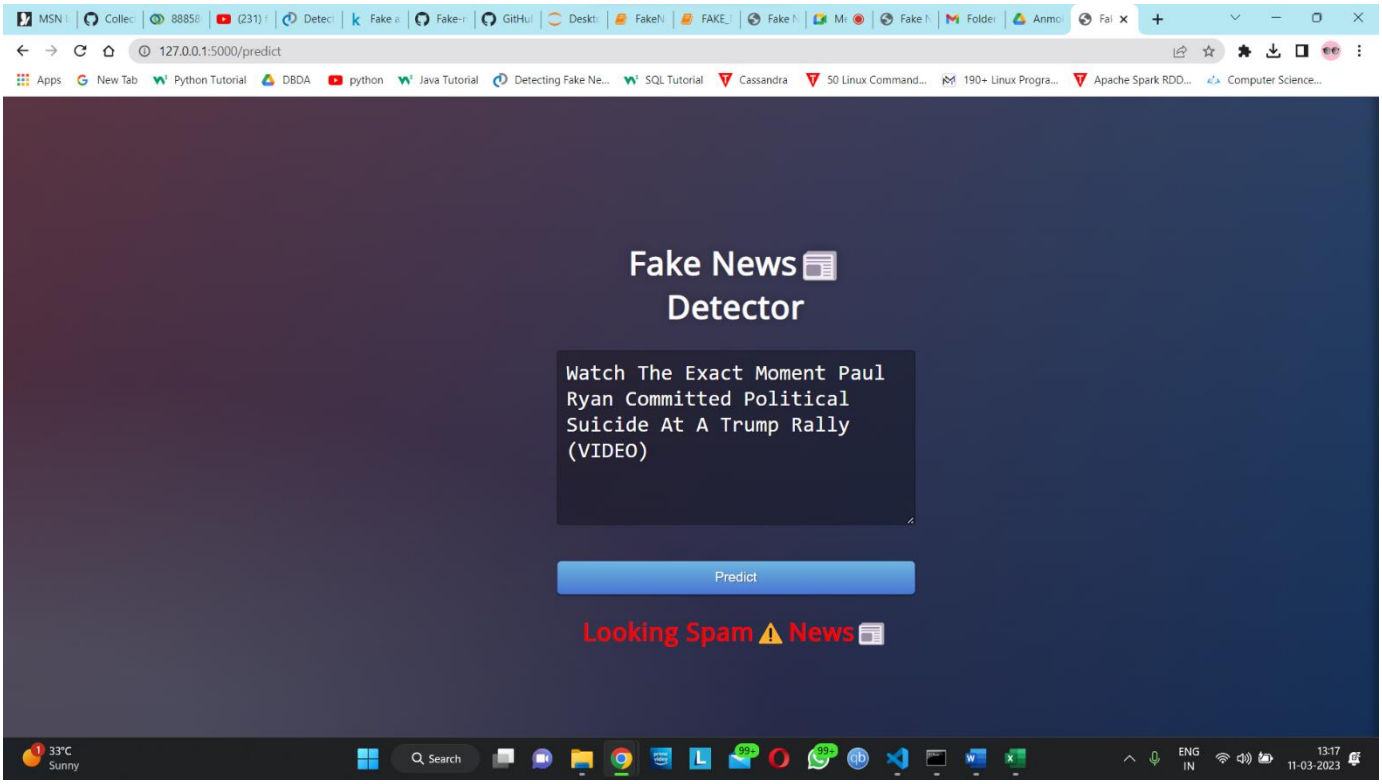
REFERENCES

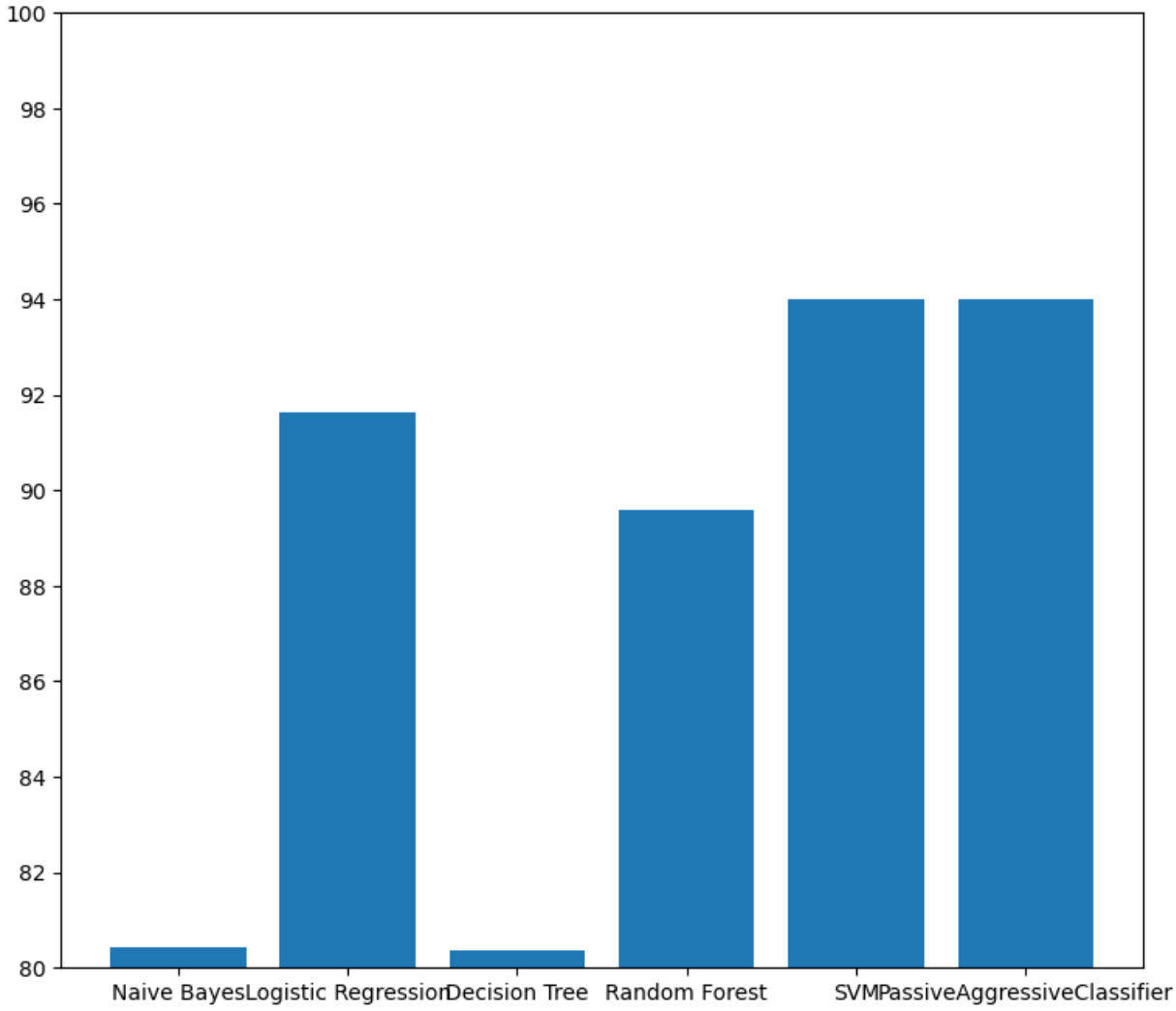
IMAGES FROM THE PROJECT




hillary
people
trump
election
one
government
clinton
said
state
new
will
american
s
time
email
president
year
world
u
russia
campaign
obama
take
even
first
right
t
know
donald
make
war
may
day
two
country
russian
now
media
think
america
united
many
fbi
thing
back
come
say
going
system
report
public
good
well
use
power
told
according
police
end
need
around
white
national
party
political
last
another
money
much
house
october
million
support
see
want
investigation
change
military
presidential
republican
work
issue
vote
made
way
law
source
look
point
life
fact
never
information
official
force
candidate
nation
voter
still
believe
part
case
syria
re
live
article
black
policy
group
order
today
become
go
mean
show
reason
attack
really
far
human
bill
video
might
win
must
office
percent
matter
health
bank
keep
result
next
real
every
long
call
seem
give
share
let
number
interest
place
economic
democratic
evidence
course
foreign
history
yet
federal
month
home
member
help
justice
problem
china
action
comey
general
great
used
find
word
voting
security
children
question
former
week
free
left
john
washington
women
continue
department
comment
post
november
democrat
put
man
already
found
nothing
called
something
global
actually
including
social
control
without


127.0.0.1:5000/predict
Fake News Detector
Kerry to go to Paris in gesture of sympathy
Predict
Looking Real News


127.0.0.1:5000/predict
Fake News Detector
Watch The Exact Moment Paul Ryan Committed Political Suicide At A Trump Rally (VIDEO)
Predict
Looking Spam ⚠ News
33°C Sunny
Search
ENG IN
13:17 11-03-2023



REFERENCES

[1] Fake news detection Akshay Jain, Amey Kasbe 2018 IEEE International Students Conference on Electrical, Electronics and Computer Sciences

[2] A Smart System For Fake News Detection Using Machine Learning 2019 2nd International

Conference on Issues and Challenges in Intelligent Computing Techniques

[3] Research on Text Classification for Identifying Fake News 2018 International Conference on Security,Pattern Analysis, and Cybernetics(SPAC)

https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

https://towardsdatascience.com/nlp-text-preprocessing-a-practical-guide-and-template- d80874676e79

https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

[4] Chaitra K Hiramath and Prof. G.C Deshpande "Fake News Detection Using Deep Learning Techniques" 2019 1st International Conference on Advances in Information Technology

[5] Abhishek Verma, Vanshika Mittal and Suma Dawn "FIND: Fake Information and News Detections using Deep Learning"

[6] Shenhao Zhang,Yihui Wang and Chengxiang Tan "Research on Text Classification for Identifying Fake News" 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)