FAKE NEWS PROJECT

Submitted by:

KUNIGIRI NAGARAJU

ACKNOWLEDGMENT

In the present world of competition there is a race of existence in which those are having will to come forward succeed. Project is like a bridge between theoretical and practical working. With this willing I joined this particular project.

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

I would like to express my special thanks to my SME Gulshana Chaudhary who gave me the golden opportunity to do this wonderful project on the topic **Fake News Project**, which also helped me in doing a research and I came to know about so many new things.

I am really thankful to him.

I would also thankful to the online platforms who help me a lot in finishing this project within the limited time.

THANKS AGAIN TO ALL WHO HELPED ME.

INTRODUCTION

- ## Business Problem Framing

- The proliferation of fake news has become a major issue in recent years, affecting not only individual lives but also the stability and credibility of various organizations and governments. The spread of false information can have significant consequences, such as manipulating public opinion, spreading rumors and paranoia, and even inciting violence.
- As a result, there is a growing need for technology to help identify and mitigate the spread of fake news. The problem of fake news can be framed as a business problem in the following ways:
- Reputation Management: Organizations, especially those in the media and politics, need to protect their reputation and credibility by ensuring that they are not spreading false information.
- Adverse Effects on Society: Fake news can cause harm to individuals, communities, and even entire societies. It can create misunderstandings, spread rumors and paranoia, and even incite violence.
- Legal Liabilities: The spread of false information can result in legal consequences for individuals, organizations, and governments. This could include lawsuits, fines, and damage to reputation.
- Financial Consequences: The spread of fake news can result in financial losses for organizations, governments, and even individuals. This could include lost revenue, increased costs associated with correcting false information, and damage to reputation and credibility.
- In conclusion, the problem of fake news is a complex and multi-faceted business problem that requires a comprehensive solution to address its various implications and consequences.

- ## Conceptual Background of the Domain Problem

The issue of fake news has become increasingly prevalent in recent years, particularly with the rise of social media. Fake news refers to misleading, false, or fabricated information that is spread deliberately to deceive people. The spread of fake news can have serious consequences, including damaging reputations, spreading misinformation, and even affecting political outcomes.

So we build a machine learning model that helps to understand that which is real and fake. We do Exploratory Data Analysis to visualize the data graphically which is easy to understand. We also used some statistical technique to see the insights of the data.

- **Review of Literature**

- There is a growing body of literature on the problem of fake news and its impact on society. Researchers from various disciplines, including computer science, journalism, psychology, and sociology, have studied the issue from different perspectives and have proposed various solutions to address it.

- In the field of computer science, researchers have focused on developing algorithms and machine learning models to automatically detect fake news. These solutions typically involve analyzing various features of news articles, such as text, images, and videos, to identify patterns and anomalies that are indicative of false information. For example, studies have shown that fake news articles tend to contain certain types of words or phrases, and that they are often shared more widely on social media than legitimate news articles.

- In the field of journalism, researchers have studied the impact of fake news on news consumption and trust in media. They have found that the spread of false information can undermine the credibility of traditional news organizations and reduce public trust in journalism. Additionally, fake news can also create confusion and misunderstandings among the public, leading to a lack of consensus on important issues.

- In the field of psychology, researchers have studied the reasons why people believe and spread false information, as well as the ways in which fake news affects individuals and societies. For example, studies have shown that people are more likely to believe and share false information if it aligns with their existing beliefs or biases. Additionally, fake news can also have a polarizing effect, causing people to become more entrenched in their views and less likely to engage in productive dialogue.

- The Literature on fake news provides a comprehensive understanding of the problem and its implications, as well as various solutions to address it. The research highlights the need for interdisciplinary solutions that combine the expertise of computer scientists, journalists, psychologists, and others to effectively mitigate the spread of false information.

- **Motivation for the Problem Undertaken**

- Every investigation begins with ideas that are further developed and inspired to address a variety of situations and circumstances.
- The client wants some predictions that could help them in further investment and improvement in selection of comments. So to help them we make this project.

- My motivation behind this project is to do the proper research because research as a process for finding a solution to a problem after making a deep analysis and conducting studies of relevant factors. In general, research is a method designed to ensure that the information obtained is reasonable and supported by the quantitative and qualitative data, and that involves a systematic process. It includes the process of designing research methods, collecting and describing.

Analytical Problem Framing

- **Mathematical/ Analytical Modelling of the Problem**

1. The problem of fake news can be modeled mathematically and analytically in several ways to help address it. Here are some of the common modeling approaches:

Text-Based Analysis: Fake news articles often contain certain patterns and anomalies in their text that can be used to identify false information. For example, fake news articles may contain certain words, phrases, or grammatical structures that are indicative of false information. Researchers have used techniques such as Natural Language Processing (NLP) and text classification algorithms to automatically detect fake news based on these patterns.

Network Analysis: The spread of fake news can also be modeled as a network, where nodes represent news articles or social media posts and edges represent the relationships between them (e.g., sharing, commenting, etc.). Researchers have used graph theory and network analysis techniques to identify key nodes and pathways in the network that are responsible for the spread of false information. This can help to target and intervene to prevent the spread of fake news.

Machine Learning: Machine learning algorithms can be used to train models on large datasets of news articles to automatically detect fake news. This involves using features such as the text, images, and videos of news articles, as well as information about their authors, sources, and distribution networks, to train the models. Researchers have used various machine learning algorithms, such as decision trees, support vector machines, and neural networks, to develop fake news detection models.

Bayesian Analysis: Bayesian analysis is a statistical approach that can be used to estimate the likelihood that a news article is false, given certain features or evidence. This involves using Bayes' theorem to update the probabilities of different hypotheses based on new evidence, such as the text and images of a news article or the credibility of its source. Researchers have used Bayesian analysis to develop models that can automatically detect fake news in real-time.

Mathematical and analytical modeling approaches can be used to address the problem of fake news in various ways. These approaches can help to automatically detect false information and prevent its spread, as well as to understand the patterns and mechanisms behind the spread of fake news.

We also create new columns to check the length of data before and after cleaning the comment feature to check the distribution of our data.

We use seaborn library to plot the target data and using wordcloud to for getting the sense of loud words in Malignant and Benign comments.

Similarly we can also make wordcloud for seprate columns where you can check the loud words for particular features because there are six target features.

Before making the model we convert our text into vectors so for that we use technique known as **TF-IDF Vectorizer**

- **Data Sources and their formats**
  - In this project the sample data is provided to us from our client database. The dataset is in csv (comma seprated values) format.
  - The data set contains the fake and real set, which has both real and fake 44898 samples in both datasets. All the data samples contain 4 fields which includes ‘title’, ‘text’, ‘subject’, ‘date’.
  - The “date” and “title” column which is of not much important so can be droppe

- **Data Preprocessing Done**
  - First we check the information of the given dataset because it tells that how many rows and columns are present in our dataset and data type of the columns whether they are object, integer or float.
  - Dropping duplicates rows if present in dataset.
  - Then we check for the null values present in our dataset. If null values are present then drop it because we cannot able to fill the text data.
  - To visualize the amount of missing values in different-2 columns we use Missingno library if needed .

- After that we check the summary statistics of our dataset. This part tells about the statistics of our dataset i.e. mean, median, max value ,min values and also it tell whether outliers are present in our dataset or not.
- We also check the correlation of our target features with each other. If columns are highly correlated with each other let's say 90% or above then remove those columns to avoid multicolinearity problem
- We also create new columns to check the length of data before cleaning the Input feature column.
- In data cleaning we use mailny five steps using function:

- Removing HTML tags
- Removing special characters
- Converting everything to lowercase
- Removing stopwords
- Using WordNetLemmatization for lemmatization

- We create new column (clean_text) after removing puncuations, stopwords from input feature to check how much data is cleaned.

- **Hardware and Software Requirements and Tools Used**

❖ **Hardware:**
- Processor—Intel (R) Core(TM) i5-2430M CPU @ 2.40GHz
- Installed Memory(RAM)—8.00 GB
- System type—64-bit Operating System

❖ **Software**: Windos 10 Pro

❖ We have used Python Package because it is powerful and general purpose programming language.
- **NumPy**—It is a math library to work with ndimensional arrays. It enables us to do computation effectively and regurarly. For working with arrays, dictionary, functions data type we need to know NumPy.
- **Pandas**—It is high level Python library and easy to use for data importing , manipulation and data analysis.
- **Matplotlib**—It is a plotting that provide 2D and 3D plotting.

- **Seaborn**-- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- **SciPy**—It is a collection of numerical algorithm and domain specific tool boxes including optimization, statistics and much more.
- **Scikit-learn**—It is a collection of tools and algorithm for machine learning. It works with NumPy and SciPy and it is easy to implement machine learning models.
- **NLTK**-- NLTK is a leading platform for building Python programs to work with human language data.

MODEL/S DEVELOPMENT AND EVALUATION

- **Identification of possible problem-solving approaches (methods)**

Before making the model we convert our text into vectors so for that we use technique known as **TF-IDF Vectorizer** .

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn import feature_extraction, linear_model, model_selection, preprocessing
from sklearn.metrics import accuracy_score
```

- **Run and Evaluate selected models**

Logistic Regression

```
# Vectorizing and applying TF-IDF
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn import feature_extraction, linear_model, model_selection, preprocessing

from sklearn.metrics import accuracy_score

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

# 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['Logistic Regression'] = round(accuracy_score(y_test, prediction)*100,2)
```

Random Forest Classifer:

```
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)
```

Naïve Bayes

```
dct = dict()

from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn import feature_extraction, linear_model, model_selection, preprocessing
from sklearn.metrics import accuracy_score

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)
```

Decision Tree Classifier

```
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)
```

SVM

```
from sklearn import svm

#Create a svm Classifier
clf = svm.SVC(kernel='linear') # Linear Kernel

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

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

From above all results we see that SVM model gives us best accuracy.

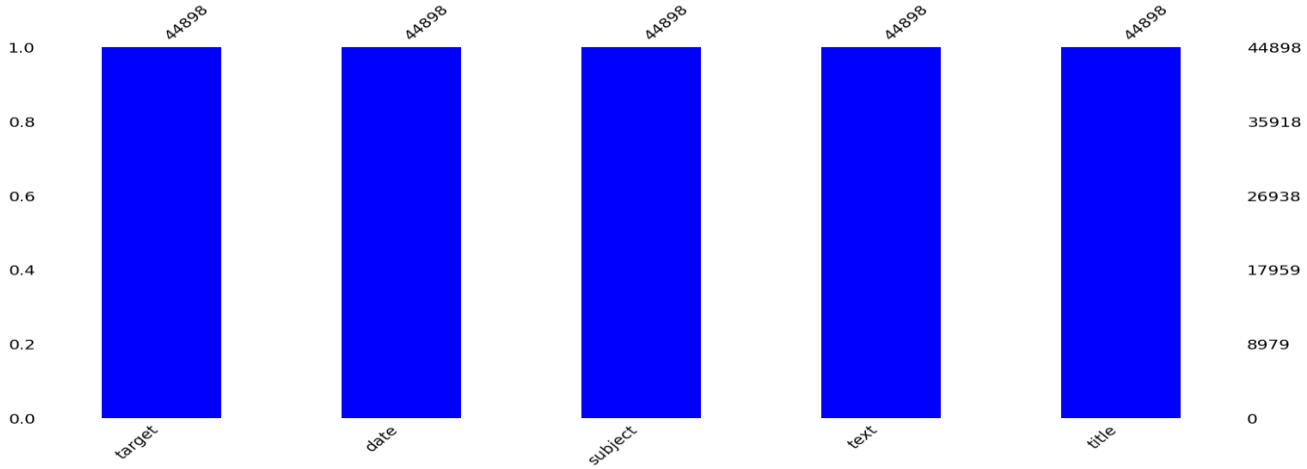
- **Key Metrics for success in solving problem under consideration**

We use accuracy score, classification report and confusion matrix as our evaluation metrics for this project.Precision talks about all the correct predictions out of total positive predictions. Recall means how many individuals were classified correctly out of all the actual positive individuals.

Visualizations

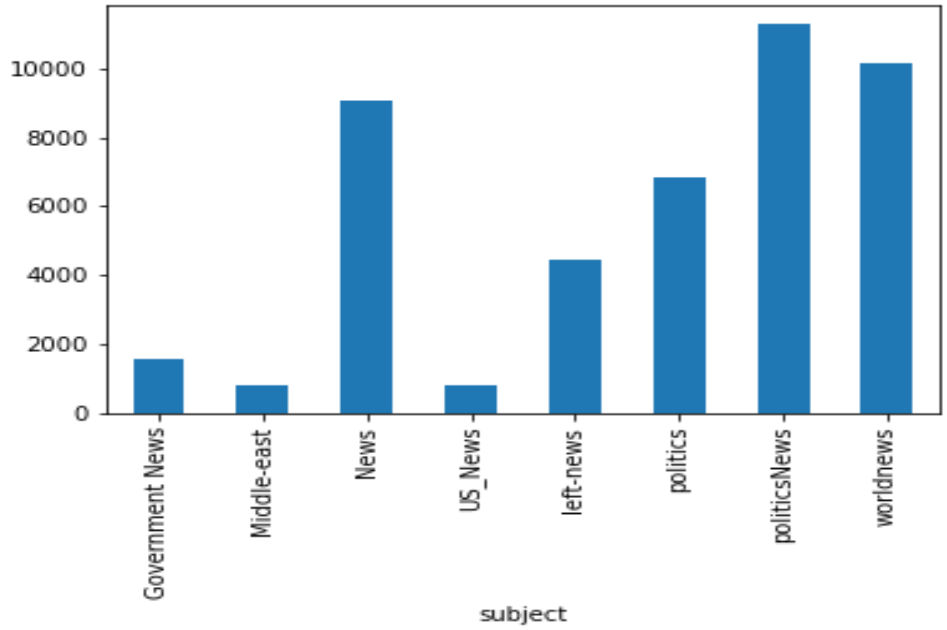
It is the graphical representation of data that is used to check about the presence of outliers, patterns, distribution of the data, etc. There are different data visualization libraries in python that include matplotlib, seaborn, etc. We will make use of the seaborn and matplotlibtolibrary to visualise the dataset.

Plotting the barplot of the null values:



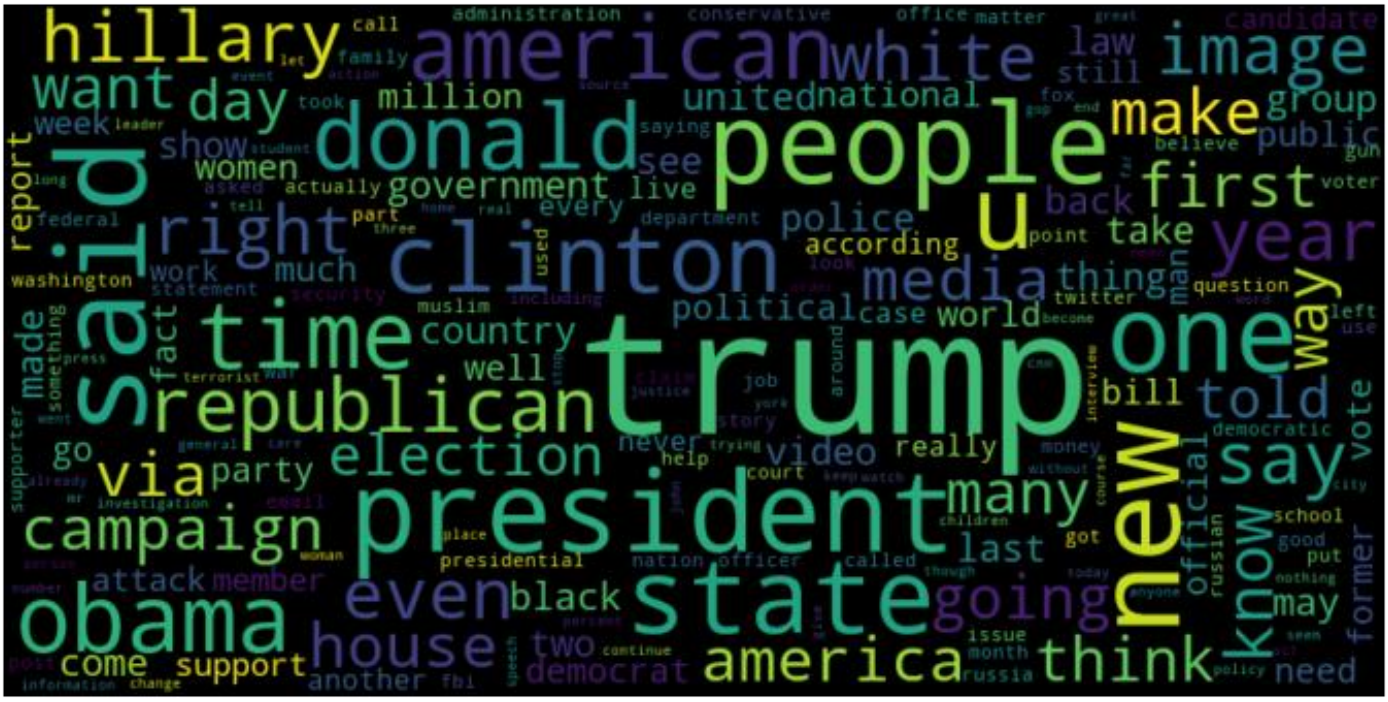
There are no null value present in our dataset.

Target Label Features:

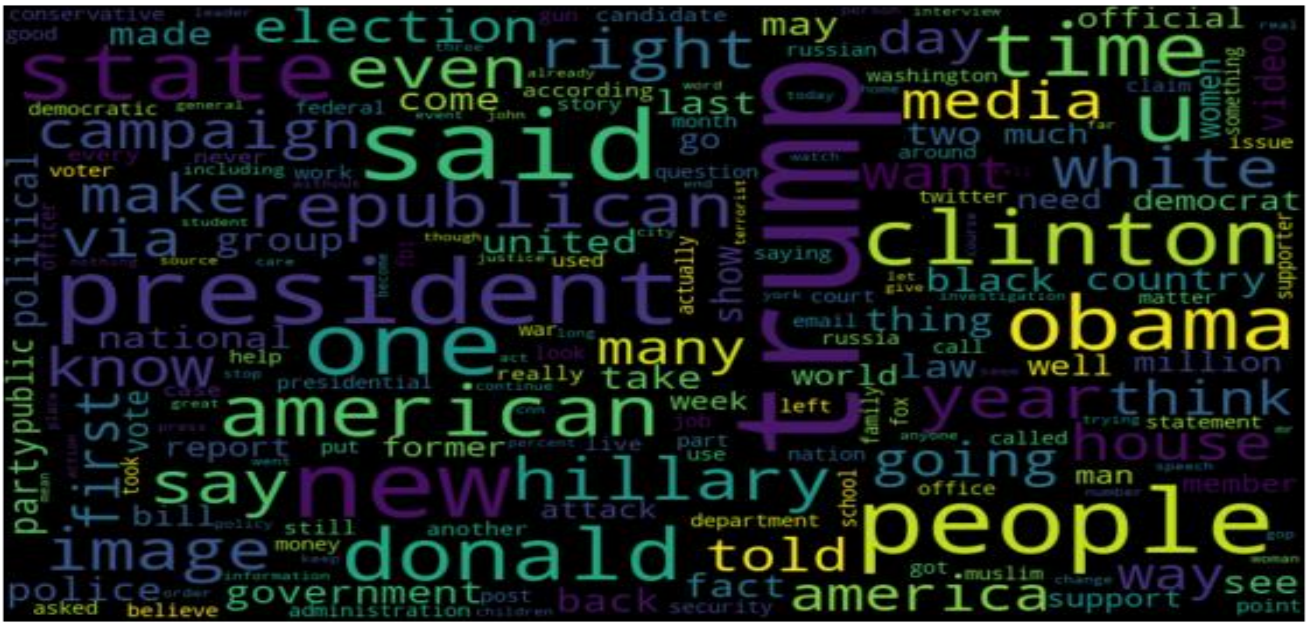


In the above graph we can see the correlation of the target features with respect to each other.

wordcloud for Fake news articles :



Wordcloud for Real news articles:



- **Interpretation of the Results**

From the above interpretation we come to know that this is classification based problem so we have learned to build a complete machine learning model for classification based problem.

The goal of any machine learning problem is to find a single model that will best predict our wanted outcome. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account, and average those models to produce one final model.

On doing this project the biggest problem I have faced is that svm model consumed a lot of time for interpretation of result

We also visualize the data and see the outcomes of our result and also plot the wordcloud to see the frequent words of real articles and as well as fake articles.

CONCLUSION & FUTURE WORK

- **Key Findings and Conclusions of the Study**

We have presented a model for fake news detection through different machine learning techniques. Furthermore, the paper investigated the other methods and compared their accuracies. The model that achieves the highest accuracy is SVM and the highest accuracy score is 99.64%. Fake news detection is an emerging research area which has a scarce number of datasets. There are no data on real-time news or regarding the current affairs. The current model is run against the existing dataset, showing that the model performs well against it. In our future work, news article data can be considered related to recent incidents in the corpus of data. The next step then would be to train the model and analyze how the accuracies vary with the new data to further improve it.