International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

FAKE NEWS DETECTION USING MACHINE LEARNING

PRERNA SANTOSH OZA[1],Dr.Mrs.PRATIBHA ADKAR[2]

MCA Department,PES Modern College Of Engineering Pune,India

Abstract :- Fake news on different platforms is spreading widely and is a matter of serious concern, as it causes social wars and permanent breakage of the bonds established among people. A lot of research is already going on focused on the classification of fake news. We can solve this issue with the help of machine learning in Python. This research contains Introduction, Importing libraries and datasets, Data Preprocessing, Preprocessing and analysis of News column, Converting text into Vectors, Model training, Evaluation, and Prediction

keywords: Data Preprocessing and analysis, Converting text into vectors, Model training, Evaluation and Prediction, Logistic Regression, Decision Tree Classifier and TfidfVectorizer.

I. INTRODUCTION

[1] [2]In today‘s society, most of the news consumption by people is through different social media platforms, since it is the most easy and convenient way of sharing news to each other. But with this comes the risk of widespread dissemination of fake news. These fake news not just adversely affect an individual but it also affects the society as a whole. Today our world is fighting against covid19. This pandemic not just destroyed the livelihood of many people but also destroyed many families. Amidst these problems, fake news just acts as a fuel to the fire. These misinformation conceal healthy behavior and encourage erroneous activities which aid in the spread of virus and lead to poor mental and physical health outcomes in people. Thus, it is very important to stop the chain of fake news from the root itself. This can be done only if have the proof whether the given news is real or fake and also the source of real news. This is where the seminar will be beneficial site’s success. CSS is the key presentational technology that is used to design websites.

Fake news refers to deliberately false or misleading information that is presented as if it were factual news. It can be spread through various mediums, including social media, news websites, blogs, and even traditional news outlets. Fake news can be created for a variety of reasons, including to deceive people, generate clicks and revenue for websites, or promote a particular agenda or ideology. It often relies on sensationalist or emotionally charged headlines to grab people's attention and generate clicks and shares on social media. Fake news can have serious consequences, including undermining trust in the media, spreading misinformation, and influencing public opinion and policy decisions. It can also lead to increased polarization and divisions in society. It is important to be vigilant and skeptical of news stories, especially those that seem too good (or bad) to be true, come from unfamiliar sources, or are not corroborated by reputable news outlets. Fact-checking websites and media literacy education can also be helpful in identifying and combatting fake news.

II. LITERATURE SURVEY

1. History

[4]Fake news is not a new phenomenon and has been around for centuries. However, the term "fake news" gained prominence in recent years due to the rise of social media and its role in spreading misinformation. In the early days of journalism, there were few regulations, and newspapers often published sensationalist stories without fact-checking. In the 1800s, "yellow journalism" emerged, where newspapers would exaggerate or make up stories to increase circulation and profits. The most famous example of yellow journalism was the coverage of the sinking of the USS Maine, which helped to spark the Spanish-American War. During World War I and II, propaganda was used extensively by governments to influence public opinion and promote their own agenda.

In the 1930s, the Nazi regime in Germany used propaganda to spread anti-Semitic and anti-democratic messages. In the 21st century, the internet and social media have made it easier for fake news to spread quickly and widely[5].

The 2016 US Presidential election saw a surge in fake news stories, with many designed to promote one candidate over another. False information about vaccines, climate change, and other important issues has also been spread online, leading to confusion and mistrust among the public[6].

Overall, the history of fake news shows that it has always been a problem, but technology has made it more pervasive and harder to combat. It is important for individuals and society as a whole to be vigilant and critical of the information they receive, and to support efforts to promote media literacy and fact-checking.

In 2016, there was widespread concern about the role of fake news on social media in the US Presidential election, with many false stories being shared across platforms such as Facebook and Twitter. The issue has continued to be a problem, with false information about COVID-19 and the 2020 US Presidential election being widely shared on social media.

In 2018, it was reported that Facebook had removed hundreds of accounts and pages linked to an Iranian disinformation campaign that had been spreading false information about the US and UK. The same year, Facebook was also criticized for its role in the spread of misinformation in Myanmar, where false information had fueled violence against the Rohingya Muslim minority.Overall, the issue of fake news on social media has continued to be a significant concern, with many efforts being made by social media companies, governments, and individuals to combat it.

2. Problem faced by social media

Social media have enhanced the experience of news consumption due to its cost effective, easily accessible and widely distributable characteristic. [2]However, it has made an average internet user easily vulnerable to consuming news that is intentionally or unintentionally distorted which can have drastic consequences and puts an individual and society at risk.

[3] [4]Therefore, detecting fake news especially on social media poses a relatively new and unique problem because of which it provides a wide range of research opportunities to tackle such challenges. One such challenge is the different ways in which a news is falsified. Fake news can vary greatly from satirical, inflated news articles that are misinterpreted as genuine to articles that make use of sensationalist, clickbait headlines to grasp the attention of users. News articles can even be fabricated and manipulated with intention to deceive, harm or influence public opinion that may result in confirmation bias or political polarization. Since fake news also usually emerge out of developing critical real time events, it is difficult to properly check and verify the quality of data itself.

Since fake news is riddled with factual inaccuracies, it can mitigate the influence of real news by competing with it. [5]In this project, we propose a system that makes use of machine learning algorithms and various feature extraction methods to detect fake news by cross verifying from various other trusted news sites while also generating and displaying real news from trusted sources in the form of a website. Through this project, we aim to obtain maximum accuracy in fake news detection and real news generation to obtain a perfect result.

- 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.
- The new model can be constructed from algorithms like Logistic Regression and Decision Tree Classifier algorithm.

3. Features of fake news

[7]Fake news can have several features, including:

- Misleading or false information: Fake news often contains information that is not true, exaggerated or manipulated to support a particular agenda or perspective.
- Sensational or clickbait headlines: Fake news stories often have sensational or exaggerated headlines that are designed to grab the reader's attention and generate clicks or shares on social media.
- Lack of credible sources: Fake news often lacks credible sources or references to back up its claims. Instead, it may rely on hearsay or anonymous sources.
- Polarizing or divisive content: Fake news stories often contain content that is polarizing or divisive, which can be used to manipulate public opinion or stir up emotions.
- Repetition: Fake news stories may be repeated or shared multiple times across various social media platforms, blogs, or websites to create the impression of widespread consensus or credibility.
- Inconsistent or contradictory information: Fake news stories may contain inconsistent or contradictory information, which can create confusion or undermine the credibility of the story.

III. DETECTION OF FAKE NEWS AND RESULT ANALYSIS

1. Steps of detection

- Importing Libraries and Datasets
- Data Preprocessing
- Preprocessing and analysis of News column
- Converting text into Vectors
- Model training, Evaluation, and Prediction

Fake News Detection

Dataset

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

	title	text	subject	date	class
0	Donald Tr	Donald Tri	News	Decembe	0
1	Drunk Bra	House Int	News	Decembe	0
2	Sheriff Da	On Friday,	News	Decembe	0
3	Trump Is S	On Christr	News	Decembe	0
4	Pope Fran	Pope Fran	News	Decembe	0
5	Racist Ala	The numb	News	Decembe	0
6	Fresh Off	Donald Tri	News	Decembe	0
7	Trump Sa	In the wak	News	Decembe	0
8	Former Cl	Many peo	News	Decembe	0
9	WATCH: E	Just when	News	Decembe	0
10	Papa John	A centerp	News	Decembe	0
11	WATCH: F	Republica	News	Decembe	0
12	Bad News	Republica	News	Decembe	0
13	WATCH: L	The media	News	Decembe	0
14	Heiress T	Abigail Di	News	Decembe	0
15	Tone Dea	Donald Tri	News	Decembe	0
16	The Inter	A new ani	News	Decembe	0
17	Mueller S	Trump sup	News	Decembe	0
18	SNL Hilari	Right now	News	Decembe	0
19	Republica	Senate Ma	News	Decembe	0
20	In A Hear	It almost s	News	Decembe	0
21	KY GOP St	In this #M	News	Decembe	0
22	Meghan M	As a Dem	News	Decembe	0
23	CNN CALL	Alabama i	News	Decembe	0

2. Importing libraries and dataset

The libraries used are:

Pandas: For importing the dataset.

Seaborn/Matplotlib: For data visualization.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Importing the downloaded dataset

```
data = pd.read_csv('/content/News (1).csv')

data.head()
```

OUTPUT:

	Unnamed: 0	title	text	subject	date	class
0	0.0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0.0
1	1.0	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0.0
2	2.0	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0.0
3	3.0	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0.0
4	4.0	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0.0

3. Data Preprocessing

The shape of the dataset can be found by the below code.

```
data.shape
```

OUTPUT:

(44919, 5)

As the title, subject and date column will not going to be helpful in identification of the news. So, drop these column.

```
data = data.drop(["title", "subject","date"], axis = 1)
```

Now, we have to check if there is any null value (we will drop those rows)

```
data.isnull().sum()
```

OUTPUT:

text 0

class 0

So there is no null value.

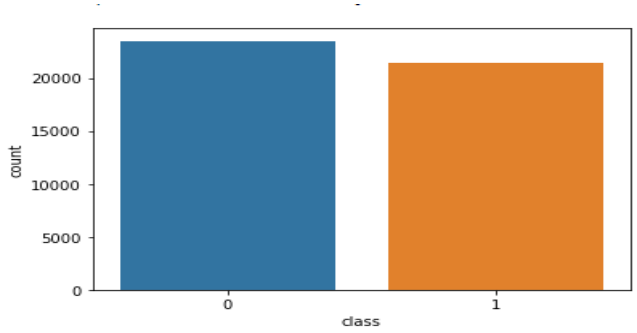
Now we have to shuffle the dataset to prevent the model to get bias. After that we will reset the index and then drop it. Because index column is not useful to us

```
# Shuffling
data = data.sample(frac=1)
data.reset_index(inplace=True)
data.drop(["index"], axis=1, inplace=True)
```

Now Let's explore the unique values in the each category using below code.

```
sns.countplot(data=data,
              x='class',
              order=data['class'].value_counts().index)
```

OUTPUT:



4. Preprocessing and analysis of News column

Firstly we will remove all the stopwords, punctuations and any irrelevant spaces from the text. For that NLTK Library is required and some of it's module need to be downloaded. So, for that run the below code.

```
from tqdm import tqdm
import re
import nltk
nltk.download('punkt')
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem.porter import PorterStemmer
from wordcloud import WordCloud
```

Once we have all the required modules, we can create a function name preprocess text. This function will preprocess all the data given as input.

```
def preprocess_text(text_data):
    preprocessed_text = []

    for sentence in tqdm(text_data):
        sentence = re.sub(r'[^\w\s]', '', senten
        preprocessed_text.append(' '.join(token.
                                          for token in s
                                          if token not i

    return preprocessed_text
```

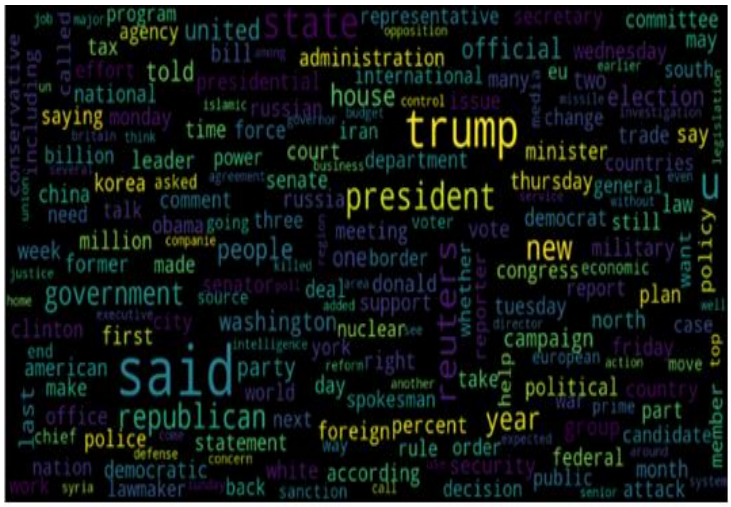
To implement the function in all the news in the text column, run the below command

```
preprocessed_review = preprocess_text(data['text'].values)
data['text'] = preprocessed_review
```

Let's visualize the WordCloud for fake and real news separately

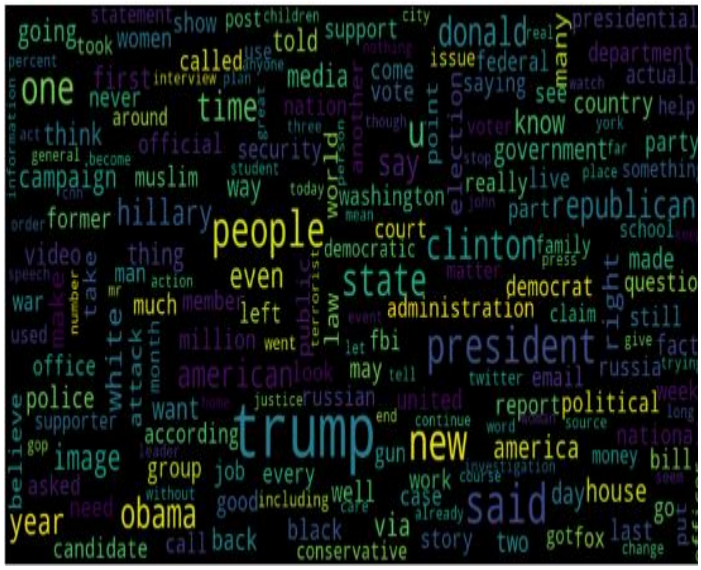
```
# Real
consolidated = ' '.join(
    word for word in data['text'][data['class'] == 1].astype(str))
wordCloud = WordCloud(width=1600,
                      height=800,
                      random_state=21,
                      max_font_size=110,
                      collocations=False)
plt.figure(figsize=(15, 10))
plt.imshow(wordCloud.generate(consolidated), interpolation='bilinear')
plt.axis('off')
plt.show()
```

OUTPUT :



```
# Fake
consolidated = ' '.join(
    word for word in data['text'][data['class'] == 0].astype(str))
wordCloud = WordCloud(width=1600,
                      height=800,
                      random_state=21,
                      max_font_size=110,
                      collocations=False)
plt.figure(figsize=(15, 10))
plt.imshow(wordCloud.generate(consolidated), interpolation='bilinear')
plt.axis('off')
plt.show()
```

OUTPUT :



Now, Let's plot the bargraph of the top 20 most frequent words.

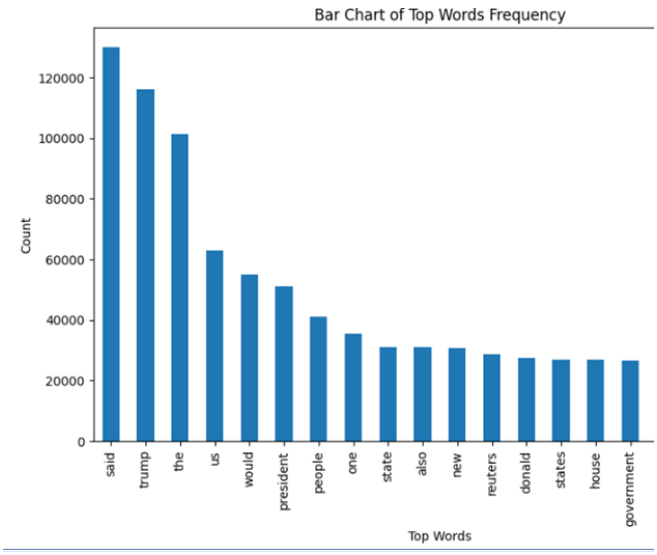
```
from sklearn.feature_extraction.text import CountVectorizer


def get_top_n_words(corpus, n=None):
    vec = CountVectorizer().fit(corpus)
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx])
                  for word, idx in vec.vocabulary_.items()]
    words_freq = sorted(words_freq, key=lambda x: x[1],
                        reverse=True)
    return words_freq[:n]


common_words = get_top_n_words(data['text'], 20)
df1 = pd.DataFrame(common_words, columns=['Review', 'count'])

df1.groupby('Review').sum()['count'].sort_values(ascending=False).plot(
    kind='bar',
    figsize=(10, 6),
    xlabel="Top Words",
    ylabel="Count",
    title="Bar Chart of Top Words Frequency"
)
```

OUTPUT :



5. Converting text into Vectors

Before converting the data into vectors, split it into train and test.

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression

x_train, x_test, y_train, y_test = train_test_split(data['text'],
                                                    data['class'],
                                                    test_size=0.25)
```

Now we can convert the training data into vectors using TfidfVectorizer.

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorization = TfidfVectorizer()
x_train = vectorization.fit_transform(x_train)
x_test = vectorization.transform(x_test)
```

6. Model training, Evaluation, and Prediction

Now, the dataset is ready to train the model.

For training we will use Logistic Regression and evaluate the prediction accuracy using accuracy_score.

```
from sklearn.linear_model import LogisticRegression

model = LogisticRegression()
model.fit(x_train, y_train)

# testing the model
print(accuracy_score(y_train, model.predict(x_train)))
print(accuracy_score(y_test, model.predict(x_test)))
```

OUTPUT :

0.993766511324171

0.9893143365983972

Let's train with Decision Tree Classifier.

```
from sklearn.tree import DecisionTreeClassifier

model = DecisionTreeClassifier()
model.fit(x_train, y_train)

# testing the model
print(accuracy_score(y_train, model.predict(x_train)))
print(accuracy_score(y_test, model.predict(x_test)))
```

OUTPUT :

0.9999703167205913

0.9951914514692787

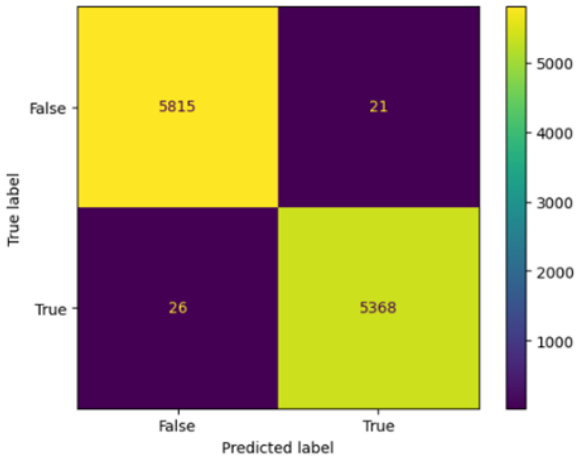
The confusion matrix for Decision Tree Classifier can be implemented with the code below.

```
# Confusion matrix of Results from Decision Tree classification
from sklearn import metrics
cm = metrics.confusion_matrix(y_test, model.predict(x_test))

cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix=cm,
                                            display_labels=[False, True])

cm_display.plot()
plt.show()
```

OUTPUT:



CONCLUSION

With the increased use of social media for news consumption and in prevalence, the widespread distribution of false news has the potential to harm both individuals and society as a whole. Even in the midst of the current covid-19 pandemic, false information on platforms like WhatsApp, Twitter and Facebook can cause panic and have a shocking impact not just on an individual but to a society as a whole. The objective is to detect the fake news through latest technologies and algorithms like- Logistic Regression and Decision Tree Classifier.

FUTURE ENHANCEMENT

Based on the obtained results, the following are the future directions for continuing the research:

- Can test, optimize and cross validate various machine learning models to get good results across different types of news as well as news related to covid-19.

- Can compare various other machine learning algorithms.
- Testing the proposed method in this seminar on a larger dataset to check for accuracy and problems associated.
- After the successful implementation and removing all problems, can try for making this in the form of mobile app.

REFERENCES

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

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

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

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

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

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

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

[8] Gurav, S., Sase, S., Shinde, S., Wabale, P. and Hirve, S., 2019. Survey on Automated System for Fake News Detection using NLP & Machine Learning Approach. International Research Journal of Engineering and Technology (IRJET), 6(01), pp.308-309.

[9] Yang, K.C., Niven, T. and Kao, H.Y., 2019. Fake news detection as natural language inference. arXiv preprint arXiv:1907.07347.

[10] Cui, L. and Lee, D., 2020. Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885.

[11] Qi, P., Cao, J., Yang, T., Guo, J. and Li, J., 2019, November. Exploiting multi-domain visual information for fake news detection. In 2019 IEEE International Conference on Data Mining (ICDM) (pp. 518-527). IEEE.

[12] Srivastava, A., Kannan, R., Chelmis, C. and Prasanna, V.K., 2019, December. RecANt: Network-based Recruitment for Active Fake News Correction. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 940-949). IEEE

[13] Long, Y., 2017. Fake news detection through multi-perspective speaker profiles. Association for Computational Linguistics.

[14] Wang, W. Y. 2017. ‖ liar, liar pants on fire‖: A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648

[15] Jin, Z.; Cao, J.; Zhang, Y.; and Luo, J. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In AAAI, 2972–2978.

[16] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on twitter during the 2016 us presidential election. Science, 363(6425):374–378, 2019.