FAKE NEWS DETECTION

PENDEM SATHWIK
CSE
IIIT Sri City
Andhra Pradesh, India
Email: sathwik.p21@iiits.in

CH.LAKSHMI NISHITHA
CSE
IIIT Sri City
Andhra Pradesh, India
Email: lakshminishitha.c21@iiits.in

***Abstract*—Fake news is the false news which influences people in a wrong direction. These days fake news is spreading so fast and becoming viral with the increased use of social media. It can be any kind of news like general, political, sports etc. Also in recent years the geopolitical news is being manipulated by different countries to destroy stock market, to spread their propaganda, to destroy value of currency. The manipulated political news can make people take wrong decisions in selecting their representatives which would affect the entire country. Sometimes fake news talks about emergency and dangerous things can make people panic or do wrong things. So it is very important for people to know whether a news is real or fake with high accuracy. By keeping all these in mind we made a project on FAKE NEWS DETECTION to make sure the NEWS that is being spread is accurate. In this project we used a dataset from Kaggle containing both genuine and fake news regarding politics. In this project we used different natural language processing techniques for data cleaning, data preprocessing and model building using N-gram which gave an accuracy of 91**

1. Introduction

In earlier days there are only two sources for getting news. One is the printed news paper and the other is radio. In those days spreading of fake news is very difficult due to less number of sources. But in today's world there are various sources. We have a great number of websites on Internet sharing news. Also many social media platforms have been established in which a small news can become more viral even if it is real or fake. This has become a major problem for the people to know about the real news. They are unable to distinguish between the fake news and the real news.

Suppose there is a news that Corona is starting again in India. By seeing this news people won't think whether it is real or fake but start panic. As we all know the previous effect caused people start panic and take wrong decisions. Also if there is news about a political party about the good activities they have done but which were not done in real. By seeing this news people get impressed about the party and vote for that party in upcoming elections. But the reality is the news is fake.

1.1. Motivation

Finding and detecting fake news is like being a guardian for the truth in the news. It's important because fake news spreads wrong information quickly in people. Detecting fake news helps keep our information accurate, especially in important things like elections. It protects how our society works, making sure people know the truth and reducing the chances of problems. Beyond that, it keeps our money safe, our online world secure, and helps us stay healthy during tough times. So we all know detecting fake news can make better decisions together.

2. State of the art/Background

Fake news detection is very important for the people in the world of social media. Here are some proposed approches from different research papers which would help for knowing the accurate model:

1) **New explainability method for BERT-based model in fake news detection**-The objective of this paper is to approach BERT-based fake news detectors which often utilise Natural Language Processing and Deep Learning for fake news detection. https://www.nature.com/articles/s41598-021-03100-6
2) **Fake News Detection Using Natural Language Processing and Logistic Regression**-In this paper the problem of fake is detected using machine learning techniques and give verifiable news. The paper identifies counterfeit news using Logistic Regression. This model successfully labels a said article with up to 80 percent accuracy. https://ieeexplore.ieee.org/document/9563292
3) **Comparison of Random Forest and Gradient Boosting algorithms for detecting fake news articles on media**-In this paper applies the chosen algorithms Random Forest and Gradient. Among

the two selected supervised machine learning algorithms the research found that Gradient Boosting algorithm has better Accuracy and better F1 scores to detect fake news articles on given data sets. https://www.diva-portal.org/smash/get/diva2:1679658/FULLTEXT02

4) **Fake news detection using logistic regression algorithm with machine learning**-With the help of Machine learning and natural language processing, it is tried to aggregate the news and later determine whether the news is real or fake using Logistic regression.The proposed model is working well and defining correctness of the results 97.21 percent of accuracy . https://assets.researchsquare.com/files/rs-3156168/v1/f87824c6-4d9e-4f19-a27f-818a851b1de9.pdf?c=1689691469
5) **Fake News Accuracy using Naive Bayes Classifier**–This paper helps us to detect the accuracy of the fake news using Naive Bayes classification. Here the data is divided into test dataset and train dataset and the train dataset is divided into groups of similar information. Test data is later matched with these groups and accuracy is found using Naive Bayes classifier. https://www.ijrte.org/wpcontent/uploads/papers/v8i1C2/A11660581C219.pdf
6) **Fake news detection using Decision tree and adaboost**-The proposed method using decision trees achieves high accuracy in fake news detection, which demonstrates the effectiveness of decision trees in feature extraction. Furthermore, Ada boost is shown which further improves the overall classification performance. https://www.eurchembull.com/uploads/paper/65c4c56c14a269d4156a4f34a28538ef.pdf
7) **Fake News Detection using Bi-directional LSTM**-Recurrent Neural Network-The paper presents a fake news detection model based on Bi-directional LSTM-recurrent neural network.The result shows the superiority in terms of accuracy of Bi-directional LSTM model over other methods namely CNN, vanilla RNN and unidirectional LSTM for fake news detection. https://www.sciencedirect.com/science/article/pii/S1877050920300806
8) **Fake News Detection Using Naive Bayes and Support Vector Machine Algorithm**-In this paper we make use of two classification models, naïve Bayes and support vector machine. https://www.researchgate.net/publication/334192809_A_Study_on_Fake_News_Detection_Using_Naive_Bayes_SVM_Neural_Networks_and_LSTM
9) **Fake News Detection Using Machine Learning**-The aim of this paper is to try to tackle the growing problems with fake news.During this paper two classification models are used: Naive Bayes and TF-IDF Vectorizer. https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040
10) **Fake news detection: A hybrid CNN-RNN based deep learning approach**-This work proposes a novel hybrid deep learning model that combines convolutional and recurrent neural networks for fake news classification. The model was successfully validated on two fake news datasets (ISO and FA-KES), achieving detection results that are significantly better than other non-hybrid baseline methods. https://www.sciencedirect.com/science/article/pii/S2667096820300070

3. Proposed System

The approach we used for fake news detection is by using traditional NLP techniques such as text cleaning, word level tokenization, lemmatization, stop word removal, word sense disambiguation, N-gram model.

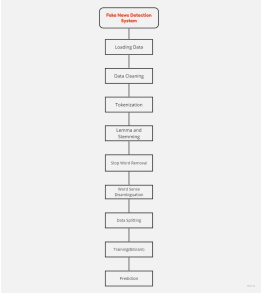

Figure 1: The steps involved in this system.

3.1. Data Collection

Collected to datasets from Kaggle.

- Namely true.csv containing all the genuine news
- False.csv containing all the fake news

3.2. Data Cleaning

Cleaned the data by removing unwanted special characters in the tuples of dataset which includes â€šÃº¹¬ Now our true.csv containing some redundant data in the tuple which is source name of the text. So, we cleaned the data by removing all the source names of the data. Now our duty is to merge the data sets. So how? We need to identify the data which is true or which is fake after merging the data. For this analysation We observed 2 methods personally, one is observing all the subjects Fake news have subjects:

- News
- Left-News
- USNews
- Politics

- Government News
- Middle-east News

True news has subjects:

- Political News
- World News

Based on the subject I can identify whether it is fake or genuine but I think maintaining the subject column is redundant and it has 6 different values which may hard for compare. Our alternative approach is to add a new column to both fake and true csv files and for fake I'll set them to 0 and whereas for genuine or true news I'll set them to 1.This is effective than comparing with subjects we need shorter memory and we have only 2 values to compare either 0 means fake 1 means true.

Dropping Unwanted Columns: Now we dropped redundant columns such as title, subject and date.

3.3. Tokenization

It is the process of breaking entire sentence or news articles into smaller components such as words.For each procedure we used a function tokenize where each news article is broken intlo simpler words. Also, the punctuation marks are removed. Now after tokenization we get the array of words.

3.4. Lemmatization and Stemming

Then we performed lemmatization and stemming to the data

- For lemmatization we defined our own set of rules and change the word into the root word.
- It is difficult for our own set of lemmetization rules to completely lemmitize all words.So, the words which are missed here stemming is performed on them. Here we have written various rules for prefixes and suffixes. After this step most of the words come to the root word by which it will be easier for the next process.

3.5. Stopword Removal

Basically, stopword removal is removing unwanted words.

For stopword removal we're using 2 techniques First approach is to list down the most frequent repeating words and which have less preferences and then we written a function if the word is from the defined set we're going to skip it. We kind of thought that this may not remove all the stopwords because we may miss some crucial stopwords in our defined set of stopwords.

So to cleanup our data more precisely we observed on pattern that stopwords are strings have length less than 3 or 2. Hence we decided to skip all the words having length 2 or below. We have not choosen 3 because me may miss crucial words like cat, dog ,car- - - etc

3.6. Word Sense Disambiguation

Word Sense Disambiguation means identifying the ambiguity in words by using its neighbouring words.

To avoid disambiguity with words we using this step. this step can be included in preprocessing of the data.We written our own rules of ambiguous words.We mentioned the different types of senses for such words and we also mentioned default sense.Since our project is about news detection we mainly focused on the ambiguous words related to politics, news.After comparing all our tokens with set of senses if any ambiguous word is matched with token we proceed for further steps. 2 words before the token and four words.

After the token are chosen to compare with the sense.If any context word matches then the related sense will be displayed instead of that letter else default sense is displayed.

3.7. Data Splitting

The data is splitted into training data and testing data, on whole training data is 80 percent and the testing data is 20 percent.We self-defined a function named train-test-split to split up the data.the function returns test and train data separately.news[text] is sent as features, news[result] is sent as label because it helps to detect it either genuine news or fake news.

3.8. Training using BiGram

By using n-gram we detected whether a news is fake or real A function named get-bi-grams takes tokens as inputs and returns list of bigrams.later these frequencies of bi-grams are stored individually for fake and true dataset.

If the label is 0(indicating fake news), it iterated over the bi-grams and updates the frequencies in the fake-bi-grams dictionary.

If the label is 1 (indicating genuine news), it iterated over the bi-grams and updates the frequencies in the genuine-bi-grams dictionary.

Bi-gram frequencies are utilized for predicting the label of news.

3.9. Prediction

a function predict-label is defined, which takes a document, calculates its bi-grams and predicts its label based on the higher score between fake and true bi-gram frequencies.And then it applied to each of the test data and then predictions are stored.later accuracy is calculated based on the predictions.

4. Results

The following are the outputs in each step:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23468 entries, 0 to 23467
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    23448 non-null  object
 1   text     23448 non-null  object
 2   subject  23448 non-null  object
 3   date     23448 non-null  object
dtypes: object(4)
memory usage: 733.5+ KB

None
```

Figure 2: True News Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21417 entries, 0 to 21416
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    21417 non-null  object
 1   text     21417 non-null  object
 2   subject  21417 non-null  object
 3   date     21417 non-null  object
dtypes: object(4)
memory usage: 669.4+ KB

None
```

Figure 3: Fake News Dataset

```
'sathwik is from planet Earth, he is a student.'
```

	title	text	subject	date	entities	result
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	{'persons': [' Republican', ' Sunday', ' Republ...	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	{'persons': [' Monday', ' Pentagon', ' Friday'...	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	{'persons': [' Russia', ' President Trump', ' ...	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	{'persons': [' George Papadopoulos', ' Austral...	1

Figure 4: After Data Cleaning

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 44885 entries, 0 to 21416
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   text    44865 non-null  object
 1   result  44885 non-null  int64
dtypes: int64(1), object(1)
memory usage: 1.0+ MB

None
```

Figure 5: Dataset after dropping unwanted columns

```
Case for campaign: Match found
Sentence: House intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been und
er the assumption, like many of us, that the Christopher Steele-dossier was what prompted the Russia
investigation so he s been lashing out at the Department of Justice and the FBI in order to protect T
rump. As it happens, the dossier is not what started the investigation, according to documents obtain
ed by the New York Times.Former Trump campaign adviser George Papadopoulos was drunk in a wine bar wh
en he revealed knowledge of Russian opposition research on Hillary Clinton.On top of that, Papadopoul
os wasn t just a covfefe boy for Trump, as his administration has alleged. He had a much larger role,
but none so damning as being a drunken fool in a wine bar. Coffee boys  don t help to arrange a New Y
ork meeting between Trump and President Abdel Fattah el-Sisi of Egypt two months before the election.
It was known before that the former aide set up meetings with world leaders for Trump, but team Trump
ran with him being merely a coffee boy.In May 2016, Papadopoulos revealed to Australian diplomat Alex
ander Downer that Russian officials were shopping around possible dirt on then-Democratic presidentia
l nominee Hillary Clinton. Exactly how much Mr. Papadopoulos said that night at the Kensington Wine R
ooms with the Australian, Alexander Downer, is unclear,  the report states.  But two months later, wh
en leaked Democratic emails began appearing online, Australian officials passed the information about
```

Figure 6: After Word Sense Disambiguiation

```
The sense of 'campaign' in the given context is: political campaign
{'campaign'}
{'political campaign'}
```

Figure 7: After Word Sense Disambiguiation

```
news['text'] = news['text'].apply(tokenize_text)

# function to lemmatize a list of tokens
def lemmatize_tokens(tokens):
    return [lemmatize(token) for token in tokens]

news['text'] = news['text'].apply(lemmatize_tokens)

print(news.head())
```

	text	result
0	[Donald, Trump, just, couldn, t, wish, all, Am...	0
1	[House, Intelligence, Committee, Chairman, Dev...	0
2	[On, Friday, it, wa, reveal, that, form, Milwa...	0
3	[On, Christma, day, Donald, Trump, announc, th...	0
4	[Pope, Franci, us, hi, annual, Christma, Day, ...	0

Figure 8: Tokenization

	text	result
0	[donald, trump, wish, american, happy, new, ye...	0
1	[house, intelligence, committee, chairman, dev...	0
2	[friday, reveal, form, milwaukee, sheriff, dav...	0
3	[christma, day, donald, trump, announc, would,...	0
4	[pope, franci, annual, christma, day, message,...	0
...	...	...
21412	[brussels, reuter, nato, alli, tuesday, welcom...	1
21413	[london, reuter, lexisnexi, provid, legal, reg...	1
21414	[minsk, reuter, shadow, disus, soviet, era, fa...	1
21415	[moscow, reuter, vatican, secretary, state, ca...	1
21416	[jakarta, reuter, indonesia, buy, sukhoi, figh...	1

[44865 rows x 2 columns]

Figure 9: Stemming, Lemmetization, StopWord Removal

```
1, 2017So, to all the people who voted for this a hole thinking he would change on
er, you were wrong! 70-year-old men don t change and now he s a year older.Photo b
tty Images.
The news is predicted to be fake.
```

Figure 10: After Training Model The output

5. Conclusion

In conclusion, fake news detection technology is pivotal in shielding the public from misinformation, preserving the integrity of information sources. While its role in filtering out deceptive content is crucial, careful implementation is essential to avoid unintended censorship or bias. Striking a delicate balance, ethical guidelines must ensure accountability and transparency in deploying these technologies. Concurrently, promoting media literacy remains imperative, empowering individuals to discern between reliable and misleading information. Achieving a more informed society requires not only technological refinement but also an emphasis on critical thinking, culminating in a resilient defense against the pervasive influence of fake news.

6. Future Work

In the above the fake news classification is did by using bigram approach for more efficiency we can go to higher multigram models such as trigram, quadragram where the efficiency will be greater.

We can also Name entity extension to it, such that it checks with existing data if NER not matches any it can simply give cannot decide, because not to decide is better that deciding false assumptiom.

Also to attain more efficiency we can build model in every individual area such as a model only for political news, a model only for film news, a model only for sports news... , if we build like this, they may have greater probabiliy to predict correct.