CS 491 NLP Project Final Report

NO.	NAME	Contribution in project
1.	Shashwat Shah	Multinomial NB Count
2.	Aman Khandelwal	Multinomial NB Tf-IDF
3.	Sagnik Mitra	DataSet loading and comparison

Reference Paper Title:Multi-source multi-class fake news detection

Authors: Hamid Karimi, Proteek Chandan Roy, Sari Saba-Sadiya, and Jiliang Tang

Index

INTRODUCTION

Fake news is a term that has been used to describe very different issues, from satirical articles to completely fabricated news and plain government propaganda in some outlets. Fake news, information bubbles, news manipulation and the lack of trust in the media are growing problems with huge ramifications in our society. However, in order to start addressing this problem, we need to have an understanding on what Fake News is. Only then can we look into the different techniques and fields of machine learning (ML), natural language processing (NLP) and artificial intelligence (AI) that could help us fight this situation.

How do we define ‘Fake News’?
It is a made-up story with an intention to deceive. This definition focuses on two dimensions: the intentional and the fact that the story is made up. This implies that honest mistakes (no matter how major they are, as long as they are accidental) are not considered to be fake news. The challenge lies, of course, on how to prove intentionality. Also, “deceive” is a relatively vague concept that I believe was purposely used by the writer to allow for a multitude of situations to be included in the definition: propaganda, completely fabricated news, partial lies, omissions or unsupported accusations. This shows already one of the major challenges with fake news, measuring it or even defining it

properly could very quickly become a subjective matter, rather than an objective metric. Despite all these drawbacks, several people and organisations have tried to categorised fake news in different ways.

Fake news comprises of 7 types of fake content:

- False Connection
- False Context
- Manipulated content
- Satire or Parody
- Misleading Content
- Imposter Content
- Fabricated content

We might believe that fake news only exists for political advantage, but this is not the only reason. In fact, it might not even be the main one. The reasons behind fake news include media manipulation and propaganda, political and social influence, provocation and social unrest and financial profit. Finding ways to determine fake news from real news is a challenge most Natural Language Processing folks want to solve. There is significant difficulty in doing this properly and without penalizing real news sources.
In our research, we studied the problem of fake news detection. In particular, we aim to answer two major research questions –

(1) Analysed data with python libraries (Numpy and Pandas) and train our model with Multinomial Naïve Bayes algorithm and (2) Compared the results of TF-IDF Vectorizer and Count Vectorizer mathematically.

Objectives

1.First objective is to collect necessary datasets

2.Divide the datasets into training and test sets

3.Implemented TF-IDF to evaluate how important is a word in a document

4.Using Multinomial Naive Bayes method to plot confusion matrix for our test data

Work Done

1. We take a quick look at the data and to do so, we use a Pandas DataFrame and check the shape, head and apply any necessary transformations.

2. Separate the labels and set up training and test datasets.

3. I decided to focus on using the longer article text. Because I knew I would be using bag-of-words and Term Frequency–Inverse Document Frequency (TF-IDF) to extract features, this seemed like a good choice. Using longer text will hopefully allow for distinct words and features for my real and fake news data.

4. We Building Vectorizer Classifiers(CountVectorizer and TfidfVectorizer) to get a good idea if the words and tokens in the articles had a significant impact on whether the news was fake or real.

5. We compare TF-IDF versus bag-of-words(aka CountVectorizer).

6. I used the scikit library to build some easily-readable confusion matrices. A confusion matrix shows the proper labels on the main diagonal (top left to bottom right). The other cells show the incorrect labels, often referred to as false positives or false negatives.

7. Comparing different classification models.
    a. Multinomial Naive Bayes with Count Vectors
    b. Multinomial Naive Bayes with Tf-Idf Vectors

Experiments and Results

We first Investigate Fake News Detection with Scikit-Learn Detecting so-called "fake news" is no easy task. First, there is defining what fake news is -- given it has now become a political statement. If you can find or agree upon a definition, then you must collect and properly label real and fake news (hopefully on similar topics to best show clear distinctions). Once collected, you must then find useful features to determine fake from real news.

```
df = pd.read_csv('fake_or_real_news.csv')
```

```
df.shape
```

```
(6335, 4)
```

```
df.head()
```

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

```
df = df.set_index('Unnamed: 0')
```

```
df.head()
```

Analysis of DataSet

We Build Vectorizer Classifiers and to get a good idea if the words and tokens in the articles had a significant impact on whether the news was fake or real, you begin by using CountVectorizer and TfidfVectorizer.

```
count_vectorizer = CountVectorizer(stop_words='english')
count_train = count_vectorizer.fit_transform(X_train)
count_test = count_vectorizer.transform(X_test)
```

```
tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_df=0.7)
tfidf_train = tfidf_vectorizer.fit_transform(X_train)
tfidf_test = tfidf_vectorizer.transform(X_test)
```

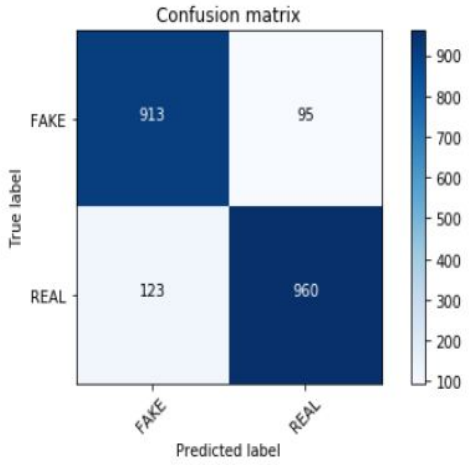
CountVectorizer and TfidfVectorizer

We set a max threshold set at .7 for the TF-IDF vectorizer tfidf_vectorizer using the max_df argument. This removes words which appear in more than 70% of the articles. Also, the built-in stop_words parameter will remove English stop words from the data before making vectors.
There are many more parameters available and you can read all about them in the scikit-learn documentation for TfidfVectorizer and CountVectorizer.

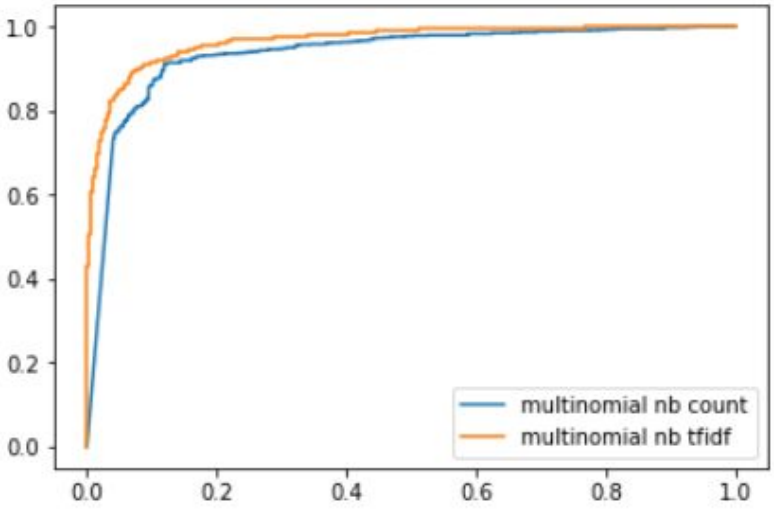
RESULT

I compared Multinomial Naive Bayes on a bag-of-words (CountVectorizer) features as well as on a Term Frequency-Inverse Document Frequency (TfIdfVectorizer) features. I also compared a Passive Aggressive linear classifier using the TF-IDF features. The resulting accuracy ranged from 83% to 93%.

```
accuracy:   0.896
Confusion matrix, without normalization
```



Confusion Matrix



Conclusions and Future Work

Conclusion:-

Fake news is a problem that is heavily affecting society and our perception of not only the media but also facts and opinions themselves.

The fake news detection faces several challenges –

1.How to incorporate multiple sources

2.How to discriminate degrees of fakeness.

We applies Multinomial NB TF-IDF having accuracy of (89.8).And Multinomial NB count having accuracy of (89.3) on our dataset and compare their results. On comparison we get Multinomial NB TF-IDF has more accuracy.

Future Work:-

We believe that this problem is solvable using AI and ML, but it will only be possible if the different communities with expertise about this work together, namely journalists, machine learning

experts and product developers. In addition, We strongly believe that very little can be done without dividing the problem into smaller problems and then combining each one of the potential solutions.

From a more general perspective, We believe that the technology will allow a change in the information consumption habits by showing us different points of view for an interesting event and then empowering the user to decide what to believe. This will not only improve our understanding of the world, but also minimise the polarisation in society.