NLP for Fake News Prediction

By: Azami and Wan Zulfadhli

Problem Statement

- Lack of credibility in the news reported

Methodology

- Using NLP
- Logistic Regression Model
- Feature engineering: Bag of Words (BoW), n grams, tfidf

```
news2 = CountVectorizer(min_df=5, ngram_range=(1,2)).fit(X_train)
```
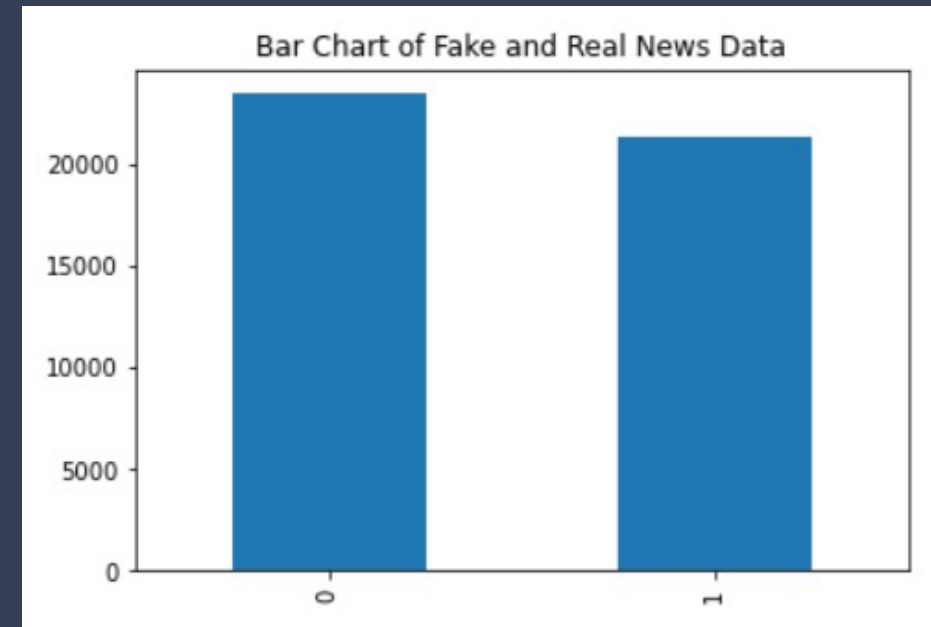
Fake news keywords(lowest coefficient):

- 'via' 'read more' 'us' 'watch' 'just' 'read' 'president trump' 'com' 'via reuters', 'this'

True news keywords(highest coefficient):

- reuters' 'washington reuters' 'reuters the', 'said on', 'on wednesday', 'reuters president', 'said', 'on thursday', 'on friday', 'president donald'

```
df.head()
```

	title	text	subject	date	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1

Bar Chart of Fake and Real News Data

Accuracy: 0.9972160356347439

AUC: 0.9990879732905399

```
[66] model.predict(news2.transform(['donald trump is convicted, as reported recently']))
     array([0])

[67] model.predict(news2.transform(['donald trump is convicted, as reported by reuters']))
     array([1])
```