Fake vs Real News Identification

Binny Manojkumar Naik

Background

Fake news refers to the spread of misinformation or false information presented as if it were real news. It can have serious consequences for individuals and society, leading to a distorted view of reality, division and mistrust, and even violence. Detecting fake news is important because it helps ensure that people have access to accurate and reliable information, which is essential for making informed decisions and participating in a healthy democracy.

Google News

BBC NEWS

Apple News

Objective

The objective is to use Natural Language Processing and Deep Learning techniques to detect the fake news article.

Dataset: Fake vs Real News Data

Dataset contains 72,134 labeled news article (1 = real, 0 = fake).

title	text	label
Bobby Jindal, raised Hindu, uses story of Chri...	A dozen politically active pastors came here f...	0
May Brexit offer would hurt, cost EU citizens ...	BRUSSELS (Reuters) - British Prime Minister Th...	0
Schumer calls on Trump to appoint official to ...	WASHINGTON (Reuters) - Charles Schumer, the to...	0
No Change Expected for ESPN Political Agenda D...	As more and more sports fans turn off ESPN to ...	0
Billionaire Odebrecht in Brazil scandal releas...	RIO DE JANEIRO/SAO PAULO (Reuters) - Billionai...	0
...	...	...
Racist Prick Spits On Black Pair, Yells N****...	Are we a post-racial nation? No, we re not, no...	1
Florida Judge Blames Rape Victim For Attendin...	The Ultra Music Festival in Miami is one of th...	1
WIKILEAKS EMAIL SHOWS CLINTON FOUNDATION FUNDS...	An email released by WikiLeaks on Sunday appea...	1
WATCH: Giuliani Demands That Democrats Apolog...	You know, because in fantasyland Republicans n...	1
GOP Rep. Just Achieved The IMPOSSIBLE By Outd...	Trump continued to claim credit for things wit...	0

[1] Fake News Detection. https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification

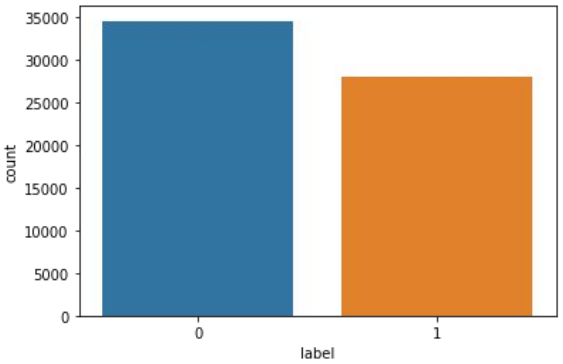
Data Preparation

- Remove Stopwords
- Removing Null Values
- Lowercase all characters
- Remove punctuations
- Remove non-English words
- Random sampling

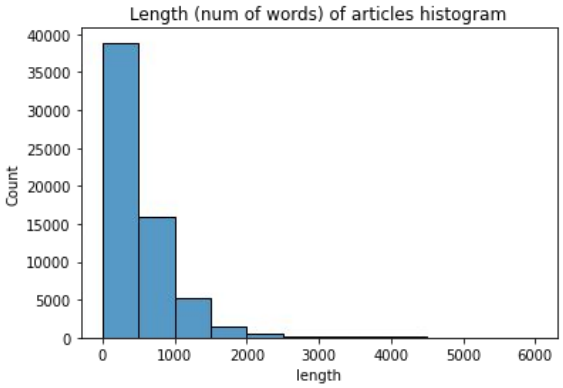
Number of Train and Test Set Samples after Cleaning

	Class 0	Class 1
Training Set	24,202 (55.2%)	19,633 (44.8%)
Test Set	10,372 (55.2%)	8,415 (44.8%)

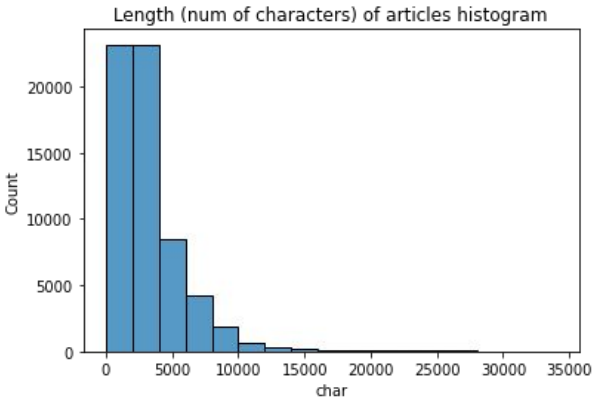
Data Visualizations



- Countplot of fake & news article

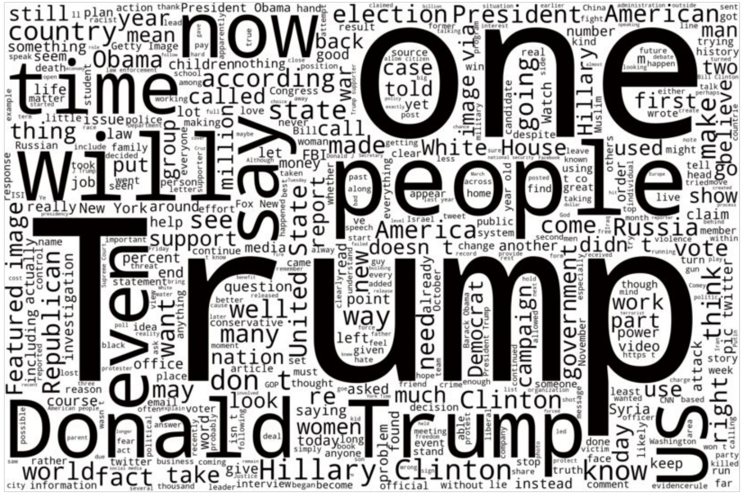


- Histogram of number of words of article

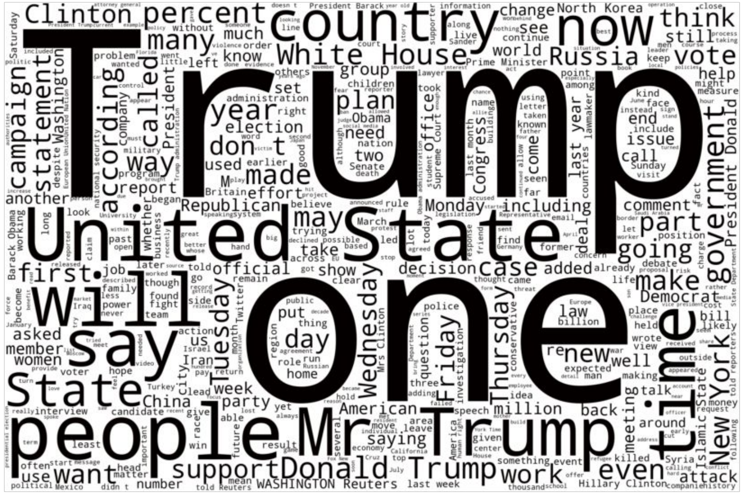


- Histogram of number of characters of article

Data Visualizations



- WordCloud of Real News Articles



- WordCloud of Fake News Articles

Text Representations

- Count Vectorization
- TF-IDF Vectorization
- GloVe Word Embeddings

Models

Neural Network:

- Multi Layer Perceptron with Glove Word Embeddings - Baseline
- Multi Layer Perceptron with TF IDF Vectorizer
- Multi Layer Perceptron with CountVectorizer
- Bidirectional LSTM with Glove Word Embeddings
- Bidirectional Encoder Representations from Transformers (BERT) Glove Word Embeddings

Performance Metrics

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

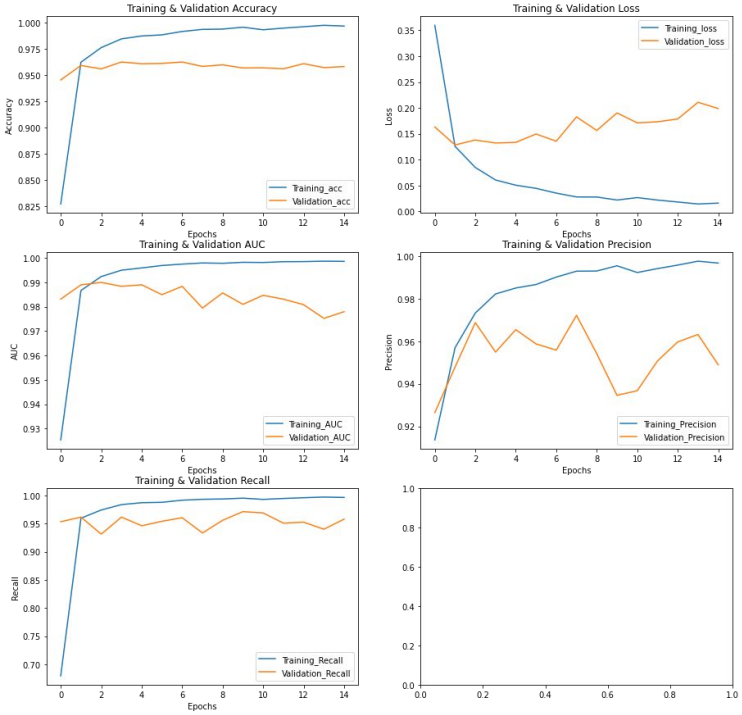
AUC & ROC

Results & Learnings

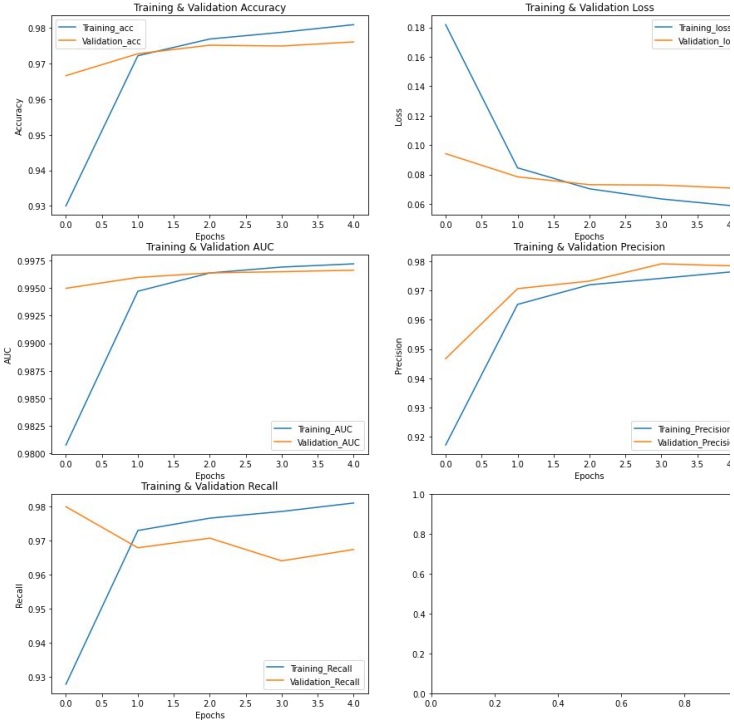
Model	PreProcessing Technique	Test Accuracy
Multi Layer Perceptron	Glove Embedding	55.50%
Multi Layer Perceptron	TF-Idf Vectorizer	93.97%
Multi Layer Perceptron	CountVectorizer	95.57%
BiDirectional LSTM	Glove Embedding	95.84%
BERT Model	Glove Embedding	97.55%

Results & Learnings

BIDirectional LSTM

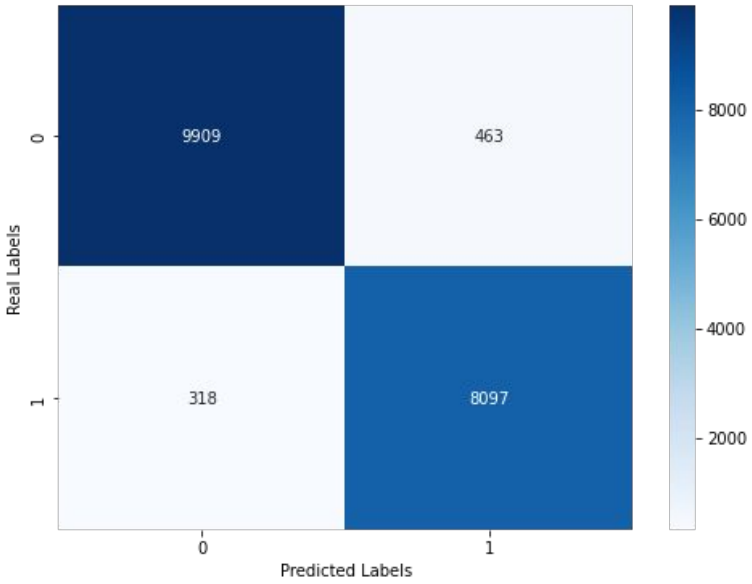


BERT Model

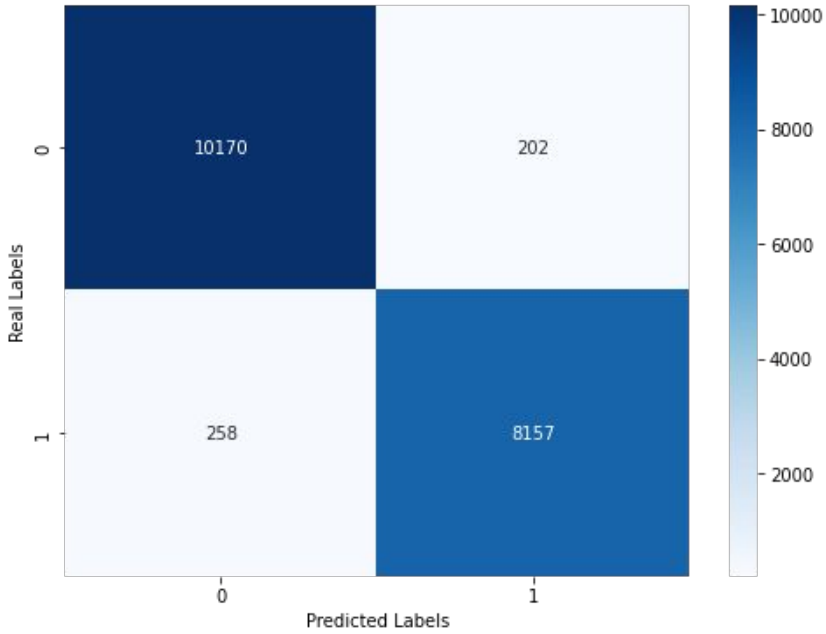


Results & Learnings

BIDirectional LSTM



BERT MODEL

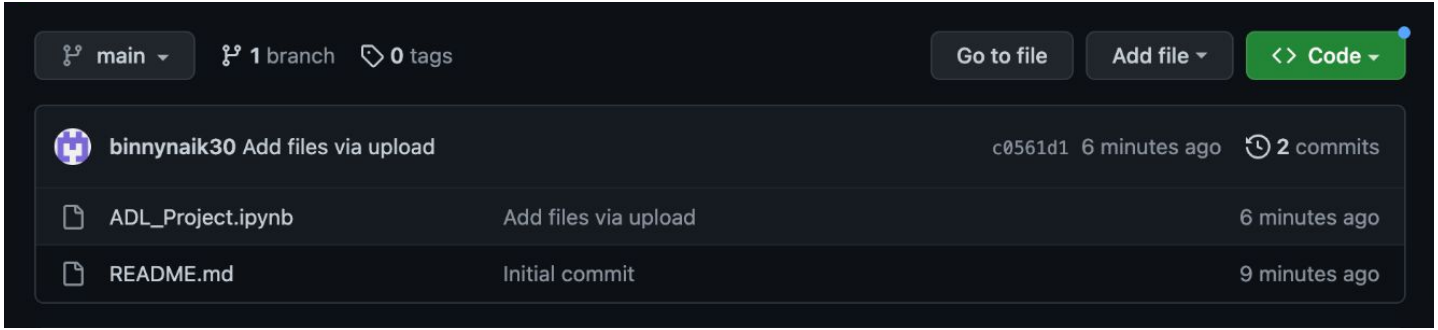


Conclusion

- BERT Model is the best fit for this problem in terms of:
  - Accuracy
  - Precision
  - Recall
  - AUC/ROC
  - Loss

Codebase

https://github.com/binnynaik30/Fake-News-Detection



Future Work

- Implement Word2Vec word embeddings
- More feature engineering (e.g., length of news article, type of question)
- Implement more complex neural networks (e.g., transformer with attention)