FAKE NEWS DETECTION

PROBLEM OVERVIEW:

Develop a machine learning program to identify when a news source may be producing fake news. We aim to use a corpus of labeled real and fake new articles to build a classifier that can make decisions about information based on the content from the corpus.

DATASET DESCRIPTION:

- Train.csv : A full training dataset with the following attributes.
    - id: unique id for a news article
    - title: the title of a news article
    - author: author of the news article
    - text: the text of the article; could be incomplete
    - label: a label that marks the article as potentially unreliable
        - 1: unreliable
        - 0: reliable
- test.csv: A testing training dataset with all the same attributes at train.csv without the label.

REQUIREMENTS:- numpy , tensorflow , pandas , nltk , gensim , keras , matplotlib

Note:- I approached this problem using three models and then compared their accuracy.

MODEL 1: LSTM (Long Short Term Memory)

We clean the raw text data and count the frequency of each word and give each word an unique ID. After truncating and padding the list, we transfer the string to a fixed length integer vector while preserving the word order information. Finally we use word embedding to transfer each word ID to a 32-dimension vector. Then we feed the processed training data into the LSTM unit to train the model.

Accuracy = 93.72%

MODEL 2: NAIVE - BAYES

This is one of the simplest approaches to classification in which a probabilistic approach is used. We convert the dataset into a frequency table, then create a likelihood table by finding probabilities. Then,we use a Naive Bayesian equation to calculate probability for each class.

Accuracy = 72.31%

MODEL 3: SVM (Support Vector Machine)

After cleaning and embedding the raw text data, the word-embedding is transferred to a feature vector, which is then fed into a Support Vector Machine (SVM) with Radial Basis Function Kernel.

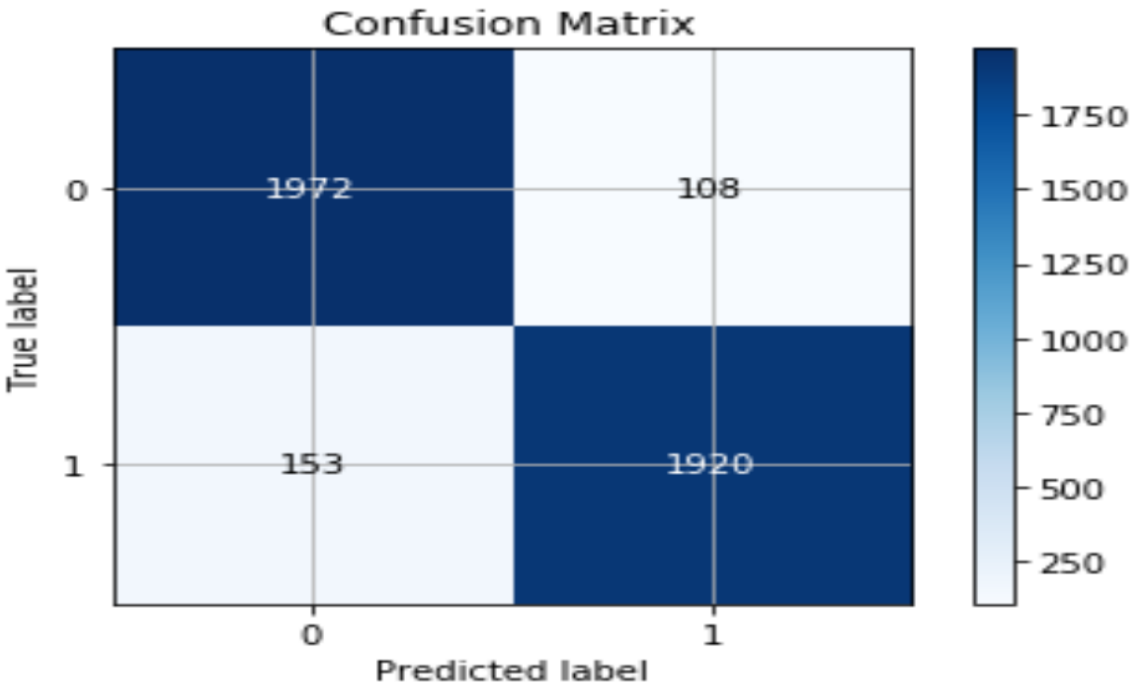
Accuracy = 91.76%

COMPARISON OF RESULTS:-

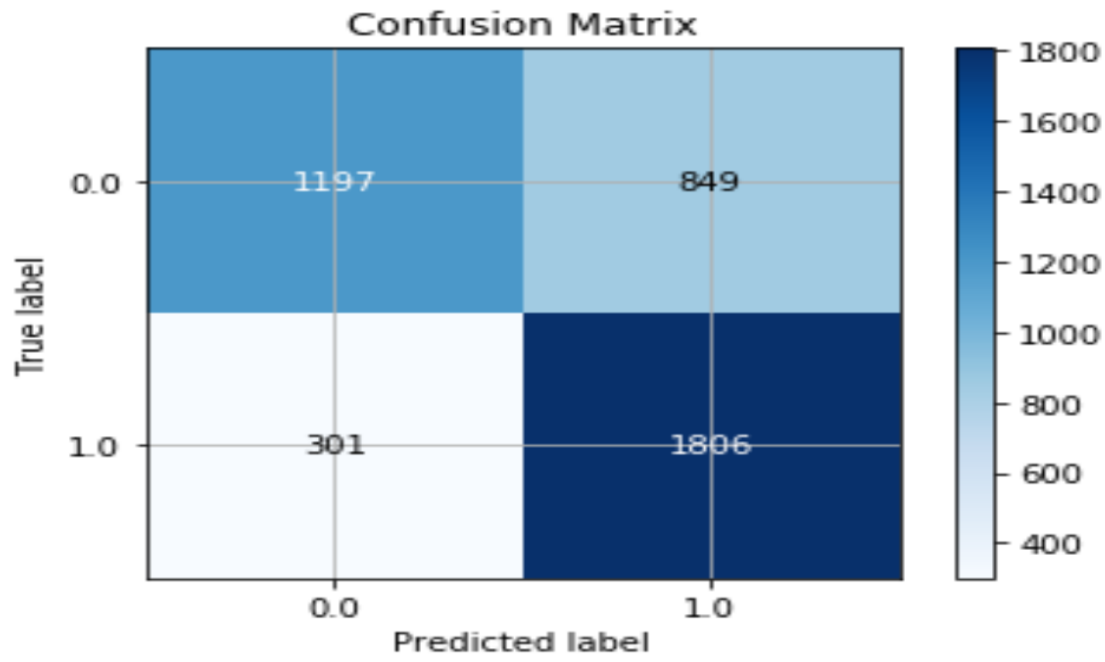
S.NO.	MODEL	ACCURACY
1.	LSTM	93.72 %
2.	NAIVE BAYES	72.31 %
3.	SVM	91.76 %

CONFUSION MATRICES:-

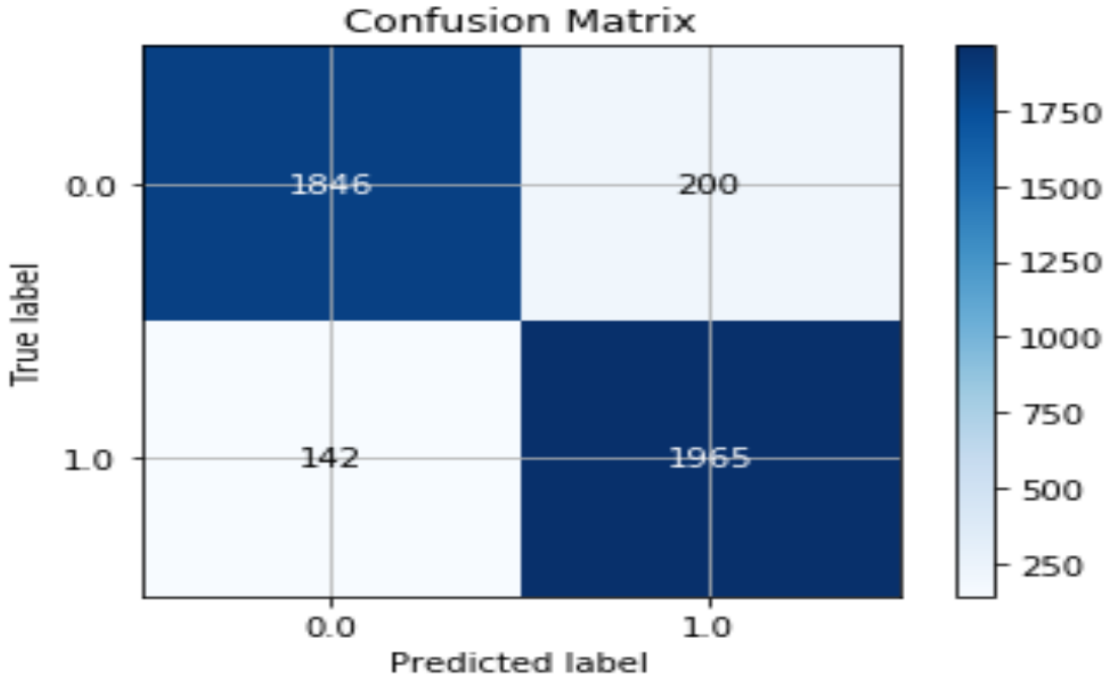
1. LSTM



2. NAIVE BAYES



3. SVM



<u>Conclusion</u> :- Out of all the three models, the LSTM model was found to be most accurate.

<u>Report by</u>:- Vivek Bhushan (mems190005043)