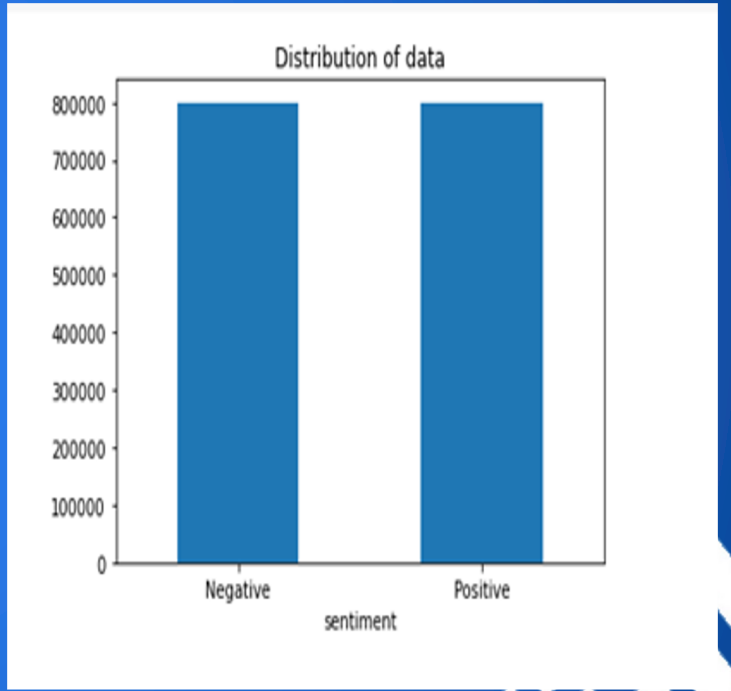
TWITTER SENTIMENT ANALYSIS

BY:
RAHUL GUPTA : MT20065
ANJALI : MT20082
COURSE: MACHINE LEARNING
INSTRUCTOR: TANMOY CHAKRABORTY

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Dataset :

The dataset obtained from Kaggle contains 80000 positive instances and 80000 negative instances of twitter data. Thus it is balanced but requires some preprocessing.



Previous Work and Approach:

- Tweets are preprocessed and cleaned, then feature are extracted using Tf-idf unigram.
- Naive Bayes and Logistic Regression model were trained on the extracted feature.

Baseline Model:

- Logistic Regression achieved f1 score of 0.78 and accuracy of 0.78.

Approach

1.Preprocess Data

2.Feature Extraction

3.ML Classifier and Prediction

Data Preprocessing :
Lowercasing
Removing urls,
Usernames,
non alphabets,
consecutive alphabets
and stop words.

Preprocessed data is fed to feature extractor which creates feature vector.

Features extracted are passed to the classifier, model built thus is used to predict the sentiment of tweets

Approach 1 - Feature extraction by Tf-idf :

Using unigram :

- The basic feature that was considered was of unigrams - takes into account one word at a time.

Using bigram :

- A bigram is a sequence of two words.
- Thus the tfidf is constructed taking into account two words at a time.

Using unigram and bigram :

- In this approach both unigrams and bigram are used to construct the tf-idf vector and then the model is trained on this vector.

Comparison between variations:

- Training models on unigram and bigram features together performed better than training on only unigram or bigram feature.
- Accuracy of SVM and Logistic Regression was 0.79 on test set.

MODEL Tf-idf	ACCURACY	F1 SCORE
Naive Bayes (unigram)	0.77	0.77
Logistic Regression (unigram)	0.78	0.78
Naive Bayes (bigram)	0.73	0.71
Logistic Regression (bigram)	0.73	0.73
Naive Bayes (unigram and bigram)	0.78	0.77
Logistic Regression (unigram and bigram)	**0.79**	**0.79**
SVM (unigram and bigram)	**0.79**	**0.79**
XGBClassifier (unigram and bigram)	0.72	0.71
MLPClassifier (unigram and bigram)	0.76	0.76

Approach 2- Feature extraction by Word2Vec:

- Word2Vec creates distributed numerical representations of word features, such as the context of individual words.

Machine Learning Models:

- Different machine Learning model is trained on features extracted by the Word2vec.
- Accuracy of the XGBClassifier was better than other machine learning model.

Deep Learning Models:

- Different Neural Network models trained on the feature extracted by the Word2vec.
- Accuracy of CNN+ bidirectional LSTM was found to be 0.76, performed better than other classic machine Learning model.

MODEL Word2vec	ACCURACY	F1 SCORE
Naive Bayes	0.52	0.51
Logistic Regression	0.52	0.31
SVM	0.50	0.51
XGBClassifier	0.59	0.59
MLPClassifier	0.50	0.60
CNN + Bi-LSTM	**0.76**	**0.76**

Approach 3- Feature extraction by pretrained GloVe vectors :

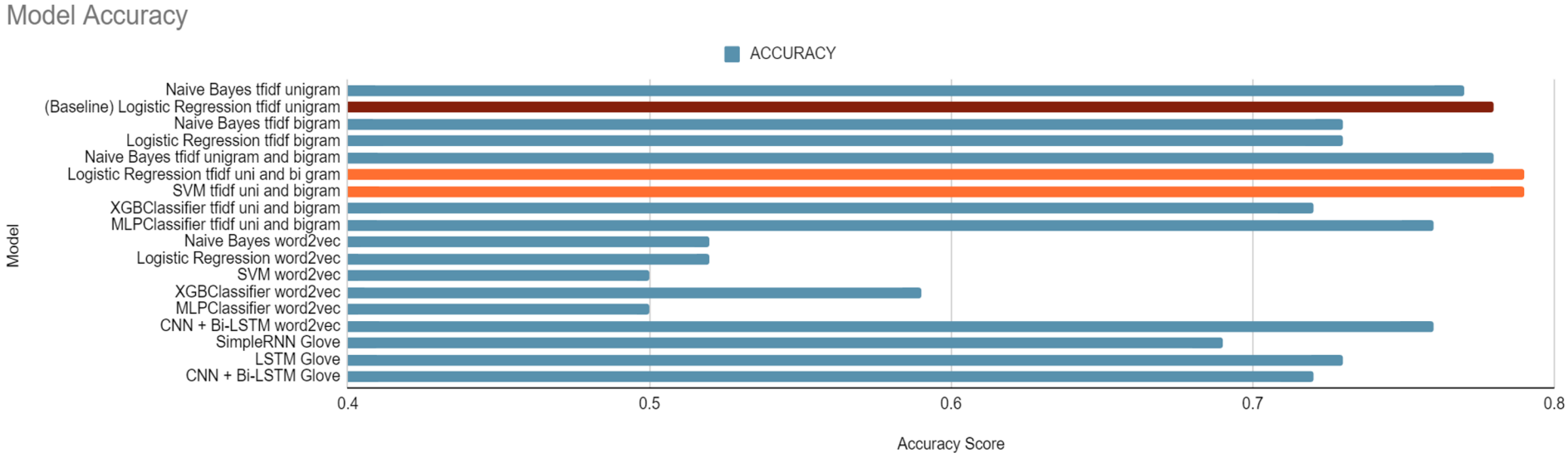
- GloVe stands for global vectors for word representation.
- It is an alternate method to create word embeddings.

Deep Learning model using GLOVE	ACCURACY	F1 SCORE
SimpleRNN	0.69	0.69
LSTM	0.73	0.72
CNN + Bi-LSTM	0.72	0.72

Comparison between Word2Vec and GloVe vector :

- Word2Vec performs better than the pretrained GloVe vector in the neural network model. CNN + Bi-directional LSTM achieved 0.76 accuracy when Word2Vec feature were extracted.

BAR GRAPH COMPARING ACCURACIES OF DIFFERENT MODELS :



Application

- Social Media monitoring
- Customer Service
- Market Research
- Brand Monitoring
- Political Campaigns

Future Scope

- Data Preprocessing using more parameters for better sentiments.
- Updating Dictionary for new Synonym and Antonyms of already existing words.
- Context Sentimental Analysis may be implemented in future for accuracy purposes.