INTRODUCTION TO DATA SCIENCE AND ANALYTICS

Project Delivery #4:
Predictive Analysis

Project Title:
Social Media Sentiment Analysis

1. Features

The dataset contains 1,600,000 tweets extracted using the twitter api. The tweets have been classified from 0 (negative) to 4 (positive). The dataset contains 6 fields which are target as integer, ids as integer, date as date, flag as string, user as string and text as string.These 6 fields are shown below.

- target: The polarity of the tweet (0 - negative, 2 - neutral, 4 - positive)
- ids: The id of the tweet.
- date: The date of the tweet.
- flag: The query. If there is no query, then this value is NO_QUERY.
- user: The user that tweeted.
- text: The text of the tweet

Target	Ids	Date	Flag	User	Text
0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOn	@switchfoot http://twitpic.com/2y1zl - Awww, t
0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by text
0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Mana
0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm
0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit
0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollyw	@Tatiana_K nope they didn't have it
0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?

Figure 1. A sample from the dataset

At the beginning, our dataset had 6 features which were target, id, date, query, user and text. We chose two of them for our purpose which are target and text. We can see that the entropy decreases significantly after this transformation.

Information gain

First entropy of dataset = 41.08269441306875

Entropy after preprocess = 14.73368002815221

2. Classification/Regression

For classification/regression experiments, the test set percentage is set to be 20%. 6 different models that are applied are CNN Model-1, CNN Model-2, LSTM Model-1, LSTM Model-2, Naive Bayes Model-1 and Naive Bayes Model-2. Below, precision, recall, f1 score and accuracy of the models are shown.

CNN Model - 1 :
Conv1D = 64
Dense = 512
Dense = 512
1024 batch size

CNN Model - 2 :
Conv1D = 31
Dense = 256
Dense = 256
512 batch size

LSTM Model - 1 : 1024 Batch size
LSTM Model - 2 : 512 Batch size

Naive Bayes Model - 1 : Multinomial, count vectorizer
Naive Bayes Model - 2 : Multinomial, Use TF-IDF

Naive Bayes Model - 1 (CountVectorizer) :

	precision	recall	f1-score	support
Negative	0.76	0.77	0.77	159493
Positive	0.77	0.76	0.76	158973
accuracy			0.76	318466
macro avg	0.77	0.76	0.76	318466
weighted avg	0.77	0.76	0.76	318466

Naive Bayes Model - 2 (tf-idf):

	precision	recall	f1-score	support
Negative	0.76	0.77	0.76	159493
Positive	0.76	0.75	0.76	158973
accuracy			0.76	318466
macro avg	0.76	0.76	0.76	318466
weighted avg	0.76	0.76	0.76	318466

LSTM Model - 1 :

	precision	recall	f1-score	support
Negative	0.78	0.79	0.79	159493
Positive	0.79	0.78	0.78	158973
accuracy			0.78	318466
macro avg	0.78	0.78	0.78	318466
weighted avg	0.78	0.78	0.78	318466

LSTM Model - 2 :

	precision	recall	f1-score	support
Negative	0.77	0.81	0.79	159493
Positive	0.80	0.76	0.78	158973
accuracy			0.78	318466
macro avg	0.78	0.78	0.78	318466
weighted avg	0.78	0.78	0.78	318466

CNN Model - 1 :

	precision	recall	f1-score	support
Negative	0.78	0.78	0.78	159493
Positive	0.78	0.78	0.78	158973
accuracy			0.78	318466
macro avg	0.78	0.78	0.78	318466
weighted avg	0.78	0.78	0.78	318466

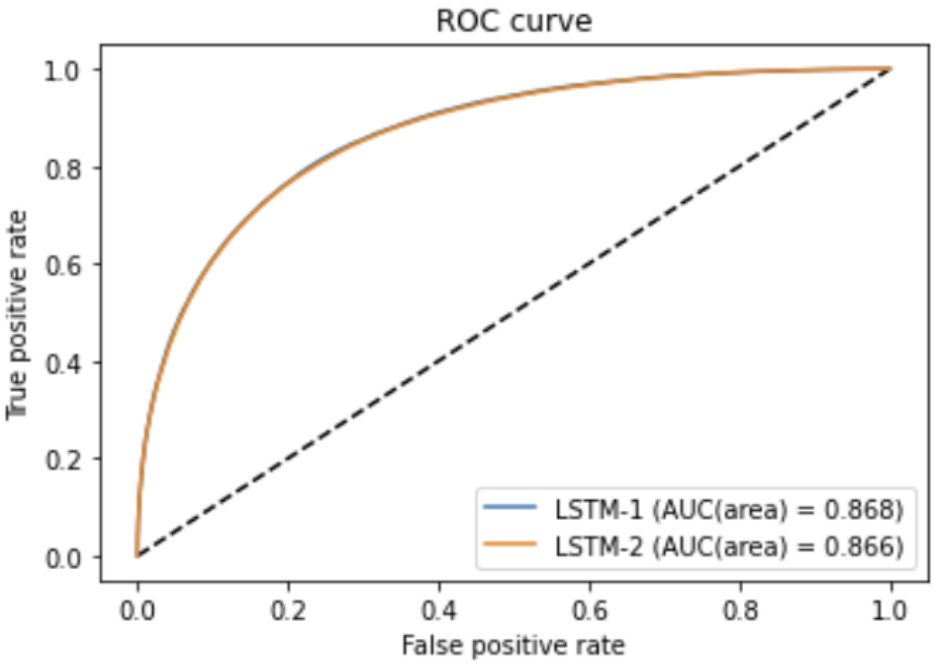
CNN Model - 2 :

	precision	recall	f1-score	support
Negative	0.80	0.73	0.76	159493
Positive	0.75	0.82	0.78	158973
accuracy			0.77	318466
macro avg	0.78	0.77	0.77	318466
weighted avg	0.78	0.77	0.77	318466

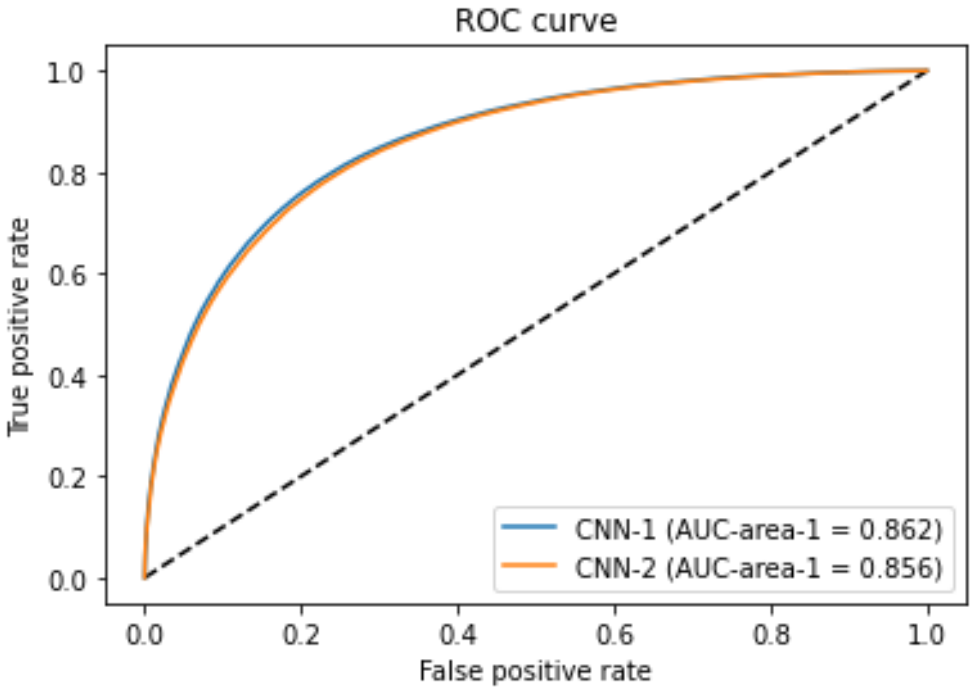
3. ROC Curve

After determining the evaluation metrics, ROC curves of the models are formed. Also AUC values are calculated and shown at the bottom of each graph.

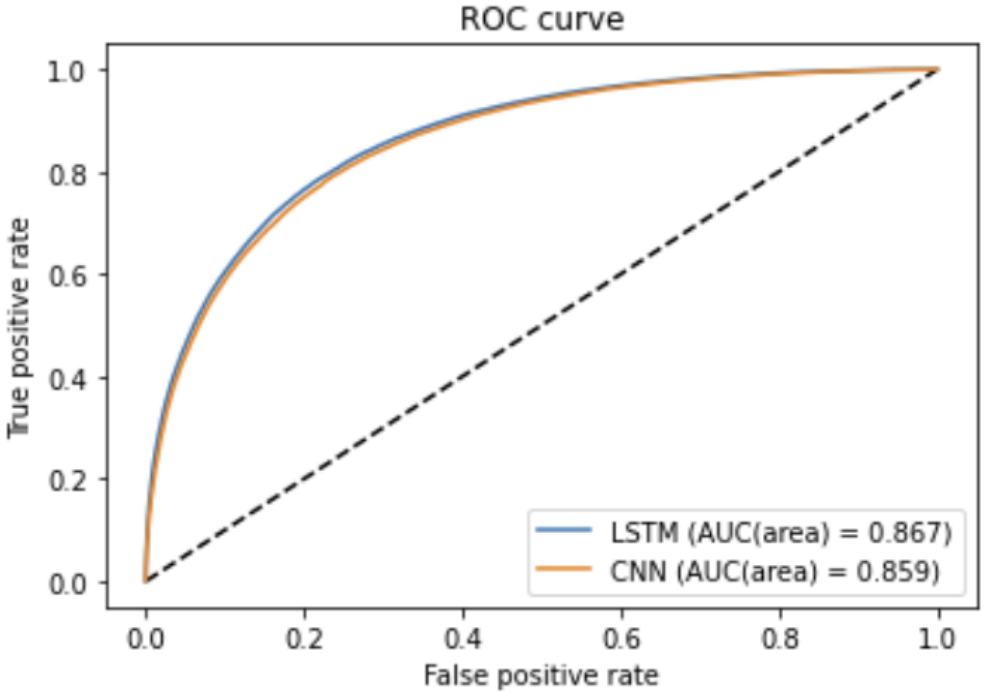
ROC Curve of CNN Model-1 and CNN Model-2 :



ROC Curve of LSTM Model-1 and LSTM Model-2 :



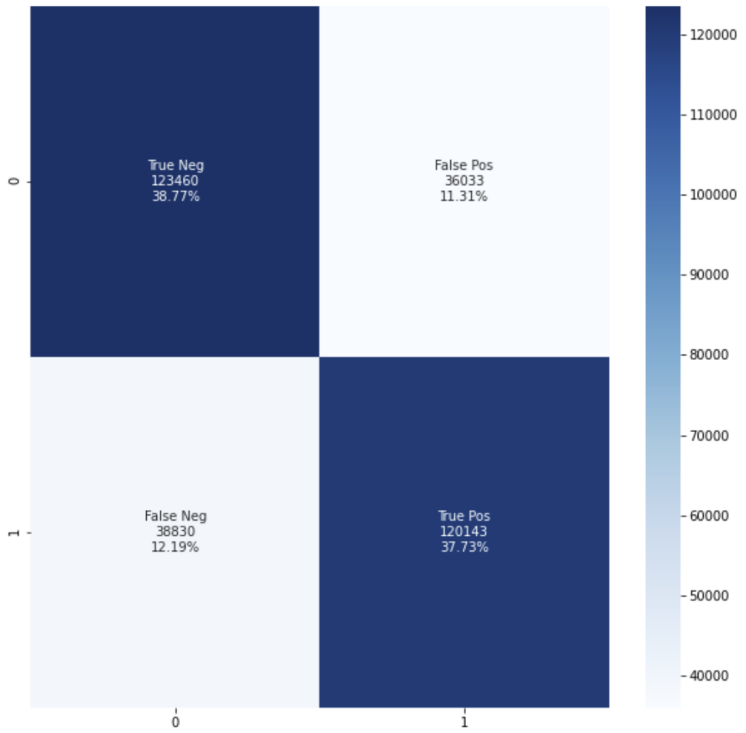
ROC Curve of best LSTM model and best CNN model :



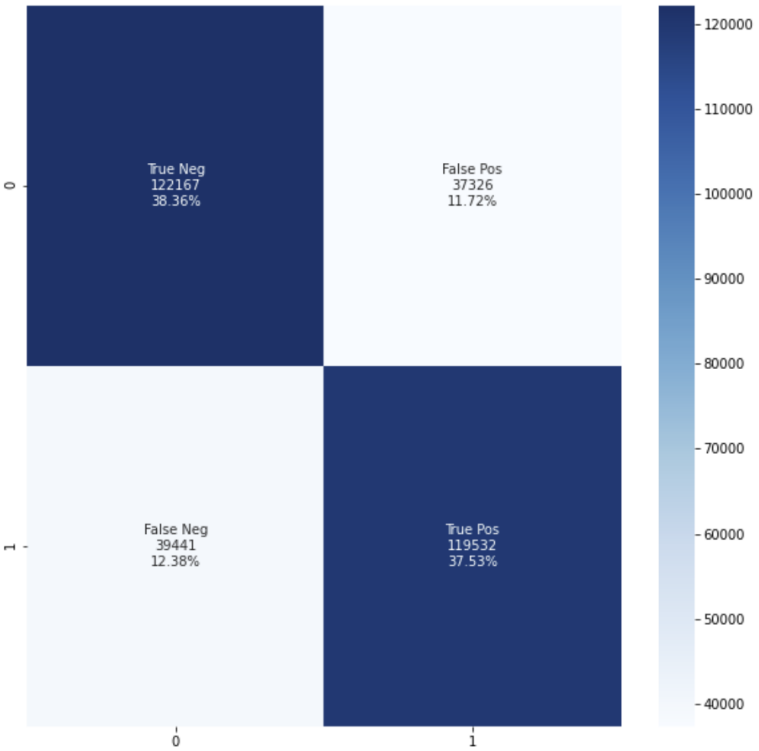
4. Confusion Matrix

Confusion matrices of the 6 model used to train the data, including the best performing model LSTM-1, are as follows:

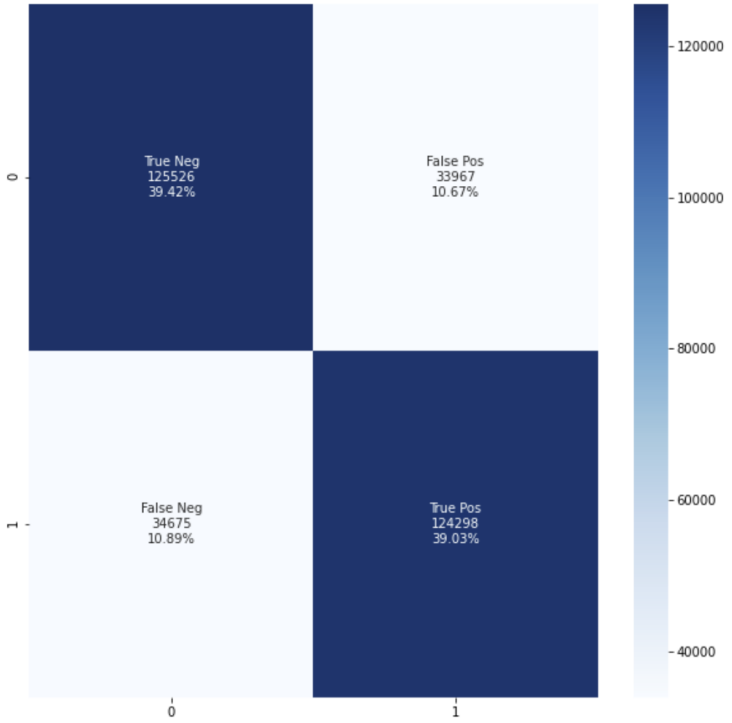
Confusion Matrix of Naive Bayes with Countvectorizer :



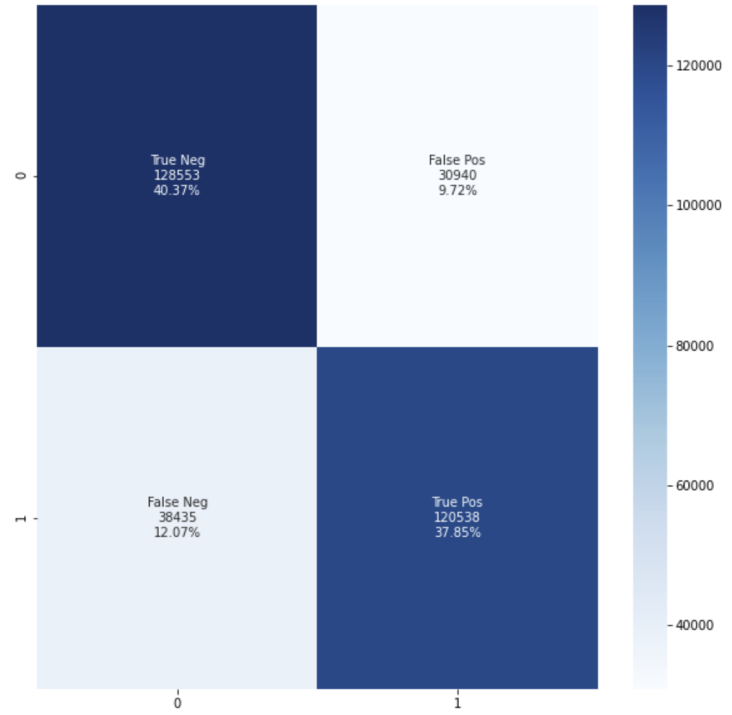
Confusion Matrix of Naive Bayes with TF-IDF :



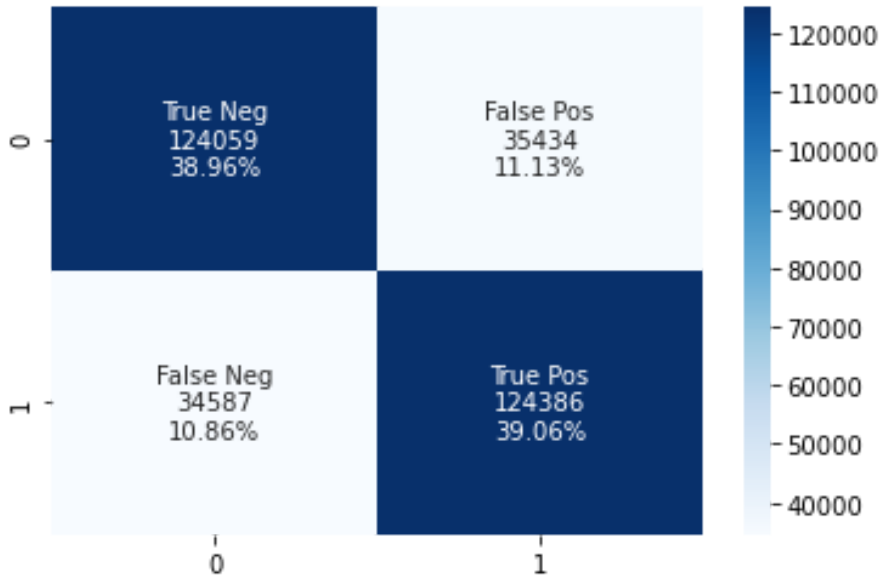
Confusion Matrix of LSTM Model - 1 :



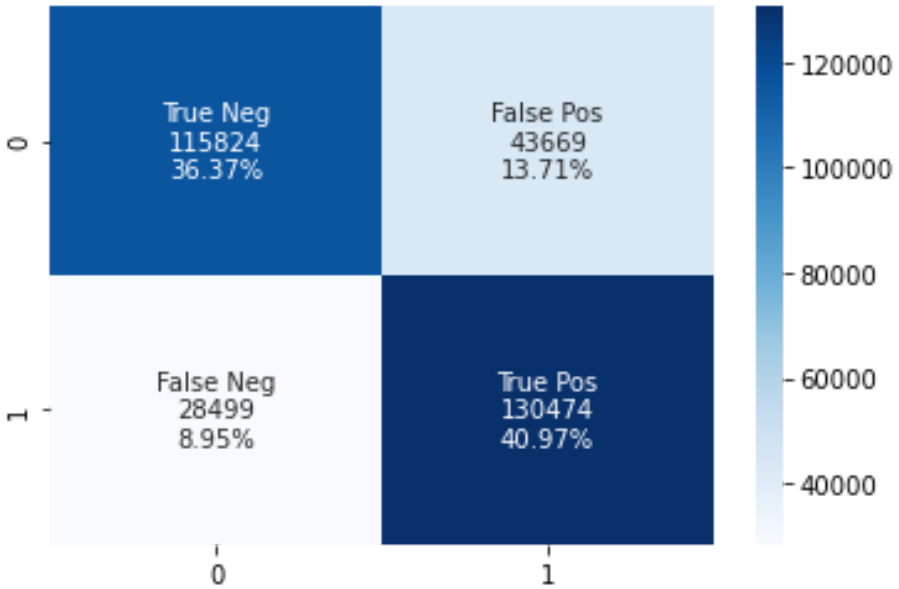
Confusion Matrix of LSTM Model - 2 :



Confusion Matrix of CNN Model - 1 :



Confusion Matrix of CNN Model - 2 :



5. Statistical Significance Analysis

According to Accuracy, P, R, F1, AUC, our best performing model is LSTM model 1 with 1024 batch size and 0.789 accuracy and the closest competitor to LSTM model 1 is CNN model 1 with accuracy 0.781. Multinomial Naive Bayes with tf-idf is the worst performing algorithm among them with accuracy 0.758.

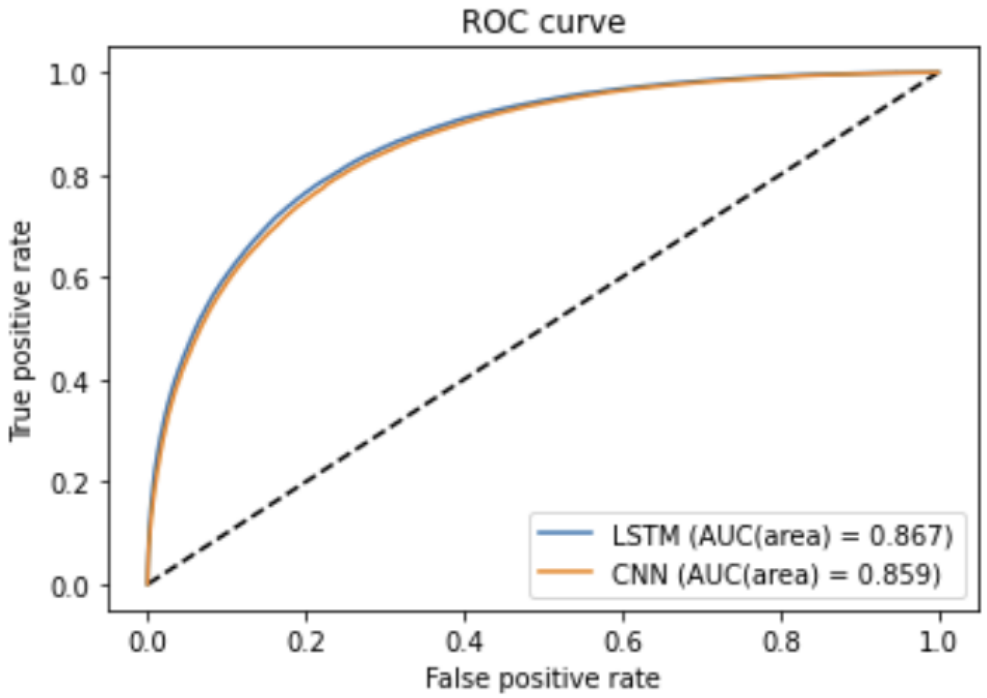
LSTM Model 1

	precision	recall	f1-score	support
Negative	0.78	0.79	0.79	159493
Positive	0.79	0.78	0.78	158973
accuracy			0.78	318466
macro avg	0.78	0.78	0.78	318466
weighted avg	0.78	0.78	0.78	318466

CNN Model 1

	precision	recall	f1-score	support
Negative	0.78	0.78	0.78	159493
Positive	0.78	0.78	0.78	158973
accuracy			0.78	318466
macro avg	0.78	0.78	0.78	318466
weighted avg	0.78	0.78	0.78	318466

ROC - AUC Analysis of Best Performing Models



6. Results

Our raw dataset has unnecessary features for our purpose. Its first entropy value was 41.08. Then we dropped the unnecessary columns, deleted the empty valued rows, and we have obtained an entropy value of 14.73. After this preprocess, we can easily see that there is an important change in entropy values.

After all six experiments, we can see that different LSTM and CNN give us very close accuracy ratios after training. Although there are really low differences, LSTM Model-1 has the best result and Naive Bayes models performed slightly worse.

Naive Bayes models have the best training time durations. It has very good speed compared to LSTM and CNN models. LSTM model-1, LSTM model-2 and CNN model-1 have close training times as each epoch takes 10 to 13 minutes for these models. Although changing the batch size in LSTM did not give an effective result difference, CNN model-2 has a better training time like 7 to 8 minutes for each epoch. Also, its accuracy is really close to the others.

LSTM model-1 has 78.9% accuracy rate with 1024 batch size and LSTM model-2 has 78.6% accuracy rate with 512 batch size. CNN model-1 has 78.2% accuracy rate with 1024 batch size and CNN model-2 has 77.2% accuracy rate with 512 batch size. Both algorithms have better training times with 512 batch size, are better than their 1024 batch sized models and their accuracy rates are really close. As a result of these, we can say that LSTM and CNN models with 1024 batch size are better for accuracy rate. But, models with 512 batch size have close accuracy rates within better training times.

For accuracy rates of Naive Bayes models there is a small difference like 1.5%. As a result of that, we can say that Naive Bayes with the CountVectorizer method gives better results than Naive Bayes with the TF-IDF method.