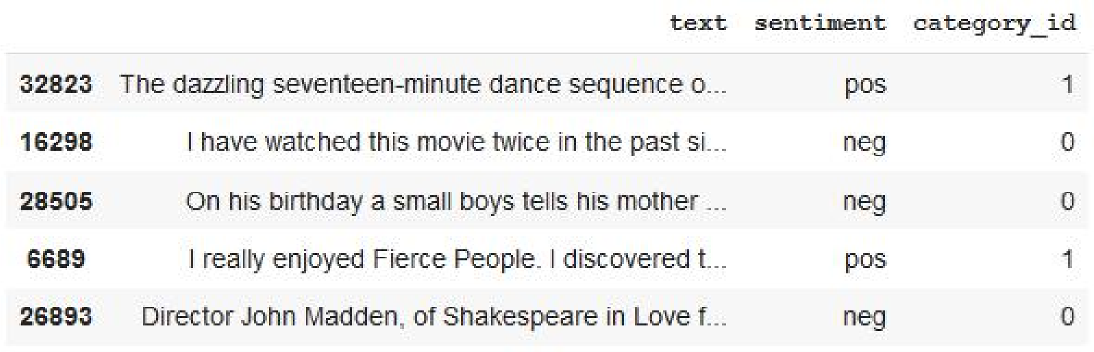
Report

The very first step after loading the dataset was to remove unwanted column such as "Unnamed: 0". Then a separate column with the name 'category_id' was developed which factorized the labels negative and positive as 0 and 1 respectively. Fig. 1 shows how the loaded data looked like.



	text	sentiment	category_id
32823	The dazzling seventeen-minute dance sequence o...	pos	1
16298	I have watched this movie twice in the past si...	neg	0
28505	On his birthday a small boys tells his mother ...	neg	0
6689	I really enjoyed Fierce People. I discovered t...	pos	1
26893	Director John Madden, of Shakespeare in Love f...	neg	0

Fig. 1

Next we plotted a bar graph showing the distribution of the labels in the dataset. We find that both positive and negative labels are equally distributed with 20000 samples in each label. Fig.2 shows the same.

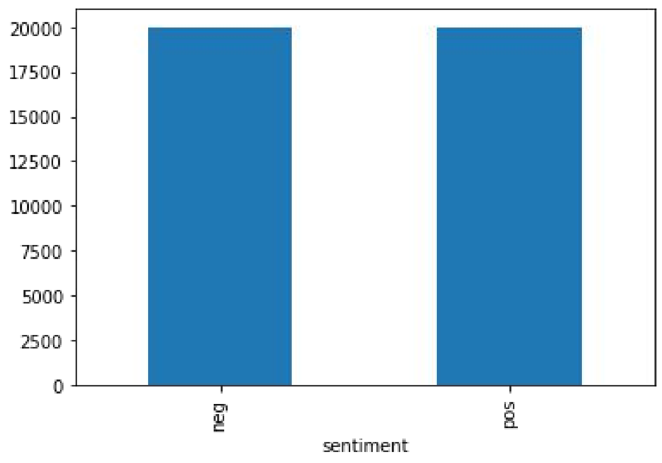


Fig. 2

Now comes the pre-processing part. The TF-IDF vectorizer remap the words in the 40000 different samples in the text column of data frame into features (superset of words) with an importance assigned based on each words frequency in the document and across documents. There were many unwanted words present in the text column of the data which were handled by removing them using the stop-word feature of TF-IDF vectorizer. The maximum number of features was chosen to be 2500 words based on the size of the samples and the length of the vocabulary.

Then we plot this features (randomly selecting 3% of the total samples) on a 2d map (graph) by using a dimensionality reduction technique called TSNE which projects these high dimension features into two dimensions and thus clearly shows the different classes present in the dataset (by forming different clusters). This plotting and dimensionality reduction was made possible with the help of matplotlib and sklearn library in python respectively. Fig.3 shows the same.

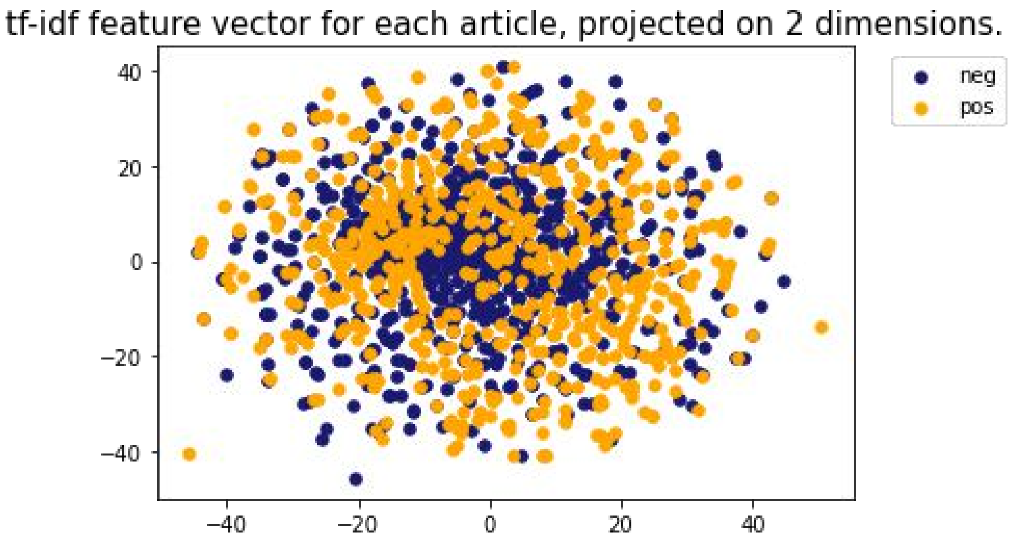


Fig. 3

The original data was divided into features (X) and label (y) which were prepared using TFIDF vectorization, which were then splitted into train (70%) and test (30%) sets. This way we are testing the model on the dataset which was never seen before by the model.

While saving the model into a pickle file, we have selected two most prominent features which were needed in order to test the model with the saved data. These features were the vocabulary extracted from the TF-IDF vectorizer and the best Machine learning model (evaluated and selected based on the accuracy of different models developed).

Now we have developed 4 different algorithms for classification task (sentiment prediction). They were Naive Bayes, Linear SVM, Logistic regression and Random Forest classifier. These classifiers are the most popular ones for binary classification and therefore they were developed with the help of sklearn library from python. Table-1 shows the accuracy of all the models developed. It was found that the maximum accuracy was observed in the case of Logistic regression model whereas Linear SVM also performed equivalently well. The least performance was shown by the RF classifier.

Table-1

Classifier	Naive Bayes	Linear SVM	Logistic regression	Random Forest
Accuracy (in %)	84.9	87	88.2	79

Lastly the model was trained using Logistic regression and the classification report (consisting of the values of precision and recall) along with the confusion matrix plot were generated (see Fig. 4 and Fig. 5 respectively for confusion matrix and classification report).

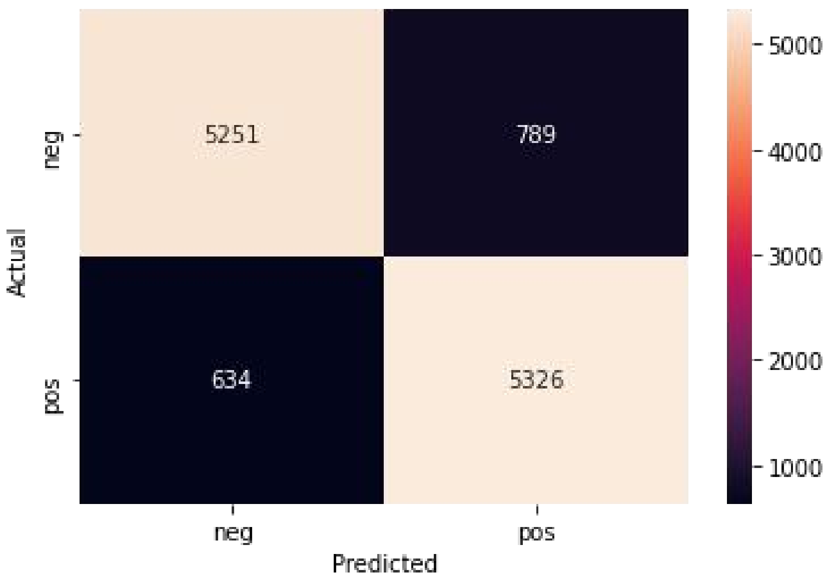
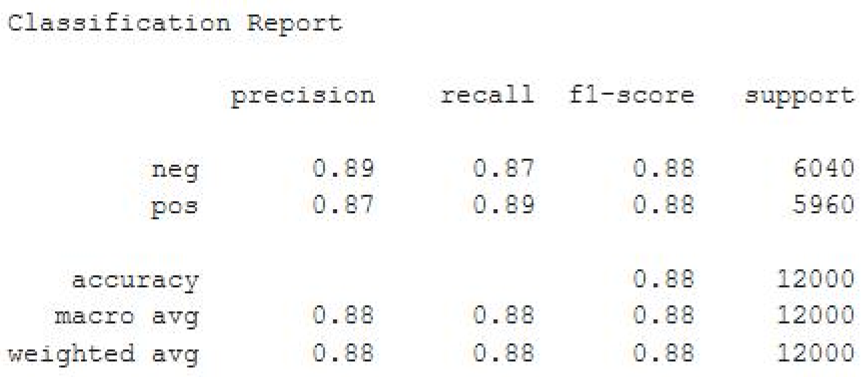


Fig. 4



```
Classification Report

              precision    recall  f1-score   support

         neg       0.89      0.87      0.88      6040
         pos       0.87      0.89      0.88      5960

    accuracy                           0.88     12000
   macro avg       0.88      0.88      0.88     12000
weighted avg       0.88      0.88      0.88     12000
```

Fig. 5

Lastly when the saved model was tested using a different script on a randomly selected 5000 different samples, the code ran without any errors and produced a decent accuracy with high values of precision and recall.