**Task: Predicting sentiment of a text and returning a text with similar sentiment**

|  |  |
| --- | --- |
|  | **Predicting sentiment of a text and returning a text with similar sentiment**  Submitted By:  Rajesh Shrestha  Santosh Khanal |

Data Source: IMDB-Dataset-of-50k-movie Reviews

**Activity:**

First of all we do research on the different data sources and methodology. We downloaded one of the dataset from the kaggle name IMDB-Dataset-of-50k-movie Reviews. Then we started the project in kaggle and added the available dataset from the data source. A datasets contain two columns i.e review and sentiment. Then read the uploaded csv file and list all the reviews and sentiment by using below code:

A screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Checked all the data with null values

A screenshot of a cell phone

Description automatically generatedWe count the number of reviews with positive and negative sentiments in the dataset and found that there are 25000 review with positive sentiments and 25000 review with negative sentiments.

A screenshot of a cell phone

Description automatically generated

After that we performed some text processing such as removing text in square brackets, removing links, special character removal, changing texts to lowercase and removing all the words containing numbers and applied pre-processing to all the review data. A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

Now labelling all the positive sentiments with 1 and negative sentiments with 0.

A screenshot of a social media post

Description automatically generated

We then split the entire data set into training data and testing data where the training data comprises 80% of the original dataset and the testing data was 20%. We train the model on the training dataset and then use the testing dataset to determine the accuracy of the prediction.

We used a vectorizer for feature extraction of the text as well.

A screenshot of a social media post

Description automatically generated

We are using logistic regression to train the model. We fit the training dataset into the logistic regression instance and then predict the sentiment for the testing dataset.

A screenshot of a cell phone

Description automatically generated

To determine the accuracy score, we are using accuracy score from sklearn and for the model that we trained the accuracy score was 0.90 which is a really a good accuracy.

A screenshot of a cell phone

Description automatically generated

We are using the trained model to predict the sentiment of any custom text past to the model. Based on this we will develop a web app and inject our data on the database which will return the text with the similar sentiment from the available dataset on the basis of the positive and negative sentiment.

A screenshot of a cell phone

Description automatically generated