Problem statement – In this**,** we have to predict the number of positive and negative review based on sentiment by using different classification model.

Sentiment Analysis – sentiment analysis is the use of natural language processing (nlp), text mining, and computational to recognize extract and understand the emotional inclination present in the text. With the widespread publicity of review, blogs, ratings, recommendation, and feedback, online opinion has turned into a goldmine for business looking to capture the market with their product. Recognize the new opportunities and manage reputation and brand name. Sentiment analysis use for almost every sector and is widely applicable for market research, customer feedback, brand monitoring, voice of employs, and social media monitoring.

Data set Description – In this data set we have use 50k. 25k for training purpose and 25k for testing purpose. Hence this is a binary classification problem based on sentiment analysis.

Cleaning and text processing- some operation perform in cleaning and text processing

1. Remove the special character.
2. Convert the text into lower case
3. Tokenize the sentence
4. Removing the stop words
5. Steaming the words
6. Join the steamed word
7. Building the corpus

Natural language toolkit - The natural language toolkit is a platform use for building a python programs that work with human language data for applying in statistical natural language processing.

It includes text processing library for tokenizing, parsing, classification, steaming, tagging and semantic.

The natural language toolkit includes more than 50 corpora and lexical sources such as the Penn Treebank Corpus, Open Multilingual WordNet, Problem Report Corpus, and Lins Dependency Thesaurus.

Porter steamer – The steaming is the procedure of reduced word to its word stem that affix to suffix and prefix or to the root of words known as lemma. For examples words such as “likes”, “liked”, “likely” and “liking” will be reduce like after steaming.

Stop words - word which are generally filter out of before processing a natural language are called stop words . some most common word in any language (like, article, preposition, pronouns, conjuction etc) and does not add much information to the text. Some example of few stop words in English are “the”, “a”, “an”, “so”, “what”.

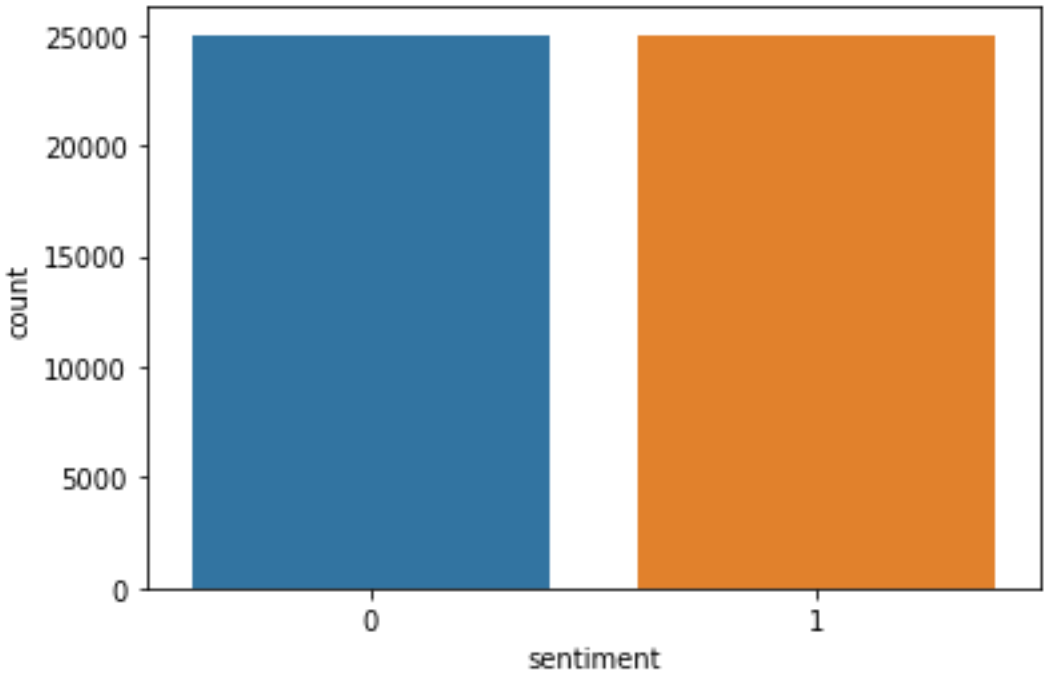
Corpus - The collection of document is called corpus.

Tokenization and lemmatization

Tokenization – Tokenization is the procedure of breakdown the sentence into word called tokens.

Lemmatization – lemmatization help the reduce to its common base root word. And help the linguistic analysis of the word.

Visualization the target feature



Fig(1)

In the above graph 0 is represent the negative review and 1 represent the Positive review in fig(1).

Bag of word model

The NLP (Natural language processing) model only understand the numerical value. Hence we need to convert textual data to a numerical value. The bag of word is a very simple model it convert a sentence into a bag of word with no meaning and it convert the sentence into a fixed length vector of number. Each word has been assigned unique number with the count of the number of occurrence of the word. We only focus on the representation of the word not the order of the word.

Count Vectorizer

Count vectorizer tokenization (tokenization means breakdown a sentence into word by performing preprocessing task hence converting all word to lower case, removing special character etc.

An encoding vector is returned the length of the entire vocabulary (all words) and integer count for the number of times each word occurs in the sentence.

Model Evaluation

In a machine learning the evaluation of a metrics to understand the performance as well as strengths and weakness of a machine learning model. Model evolutional is most important for evaluating the model performance in early research stages and also play a role in model monitoring.

Accuracy – Accuracy is evaluating the number of correct prediction divide the total number of correct prediction, and the formula of accuracy is –

Accuracy = number of true prediction / total number of true prediction

Accuracy = TP + TN / TP+TN+FP+FN

Where TP = True Positive, FP = False Positive, TN=True Negative, FN=False Negative

Precession – The precession is to summarize the ratio between the numbers of positive samples correctly classified to the total number of samples classified as a positive. The formula of precession.

Precession - true Positive / True Positive + False Positive

Precession – TP /TP+FP

Recall – Recall is summarized by the number of positive samples divide by the total number of positive samples and negative samples. The recall measure the model detect the positive samples. The formula of recall is -

Recall = True Positive / True Positive + False Negative

Recall = TP / TP + FN

Where TP = True Positive and FN=False Negative

F1-Score - F1 score combine the precession and recall metrics into single metric. F1 score has been designed to work well on imbalanced data. The formula of F1-score is -

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Model Building

We are build our sentiment classification model. But we need to select a supervised classification model that satisfied our requirement model.

In this, we have use multiple machine learning algorithm like logistic regression accuracy is 0.88, random forest classifier accuracy is 0.85, multinomial naïve Bayes model accuracy is 0.86

and support vector machine model accuracy is 0.87. Compared all the supervised machine learning algorithm our logistic regression model give the high performance compare to another model.

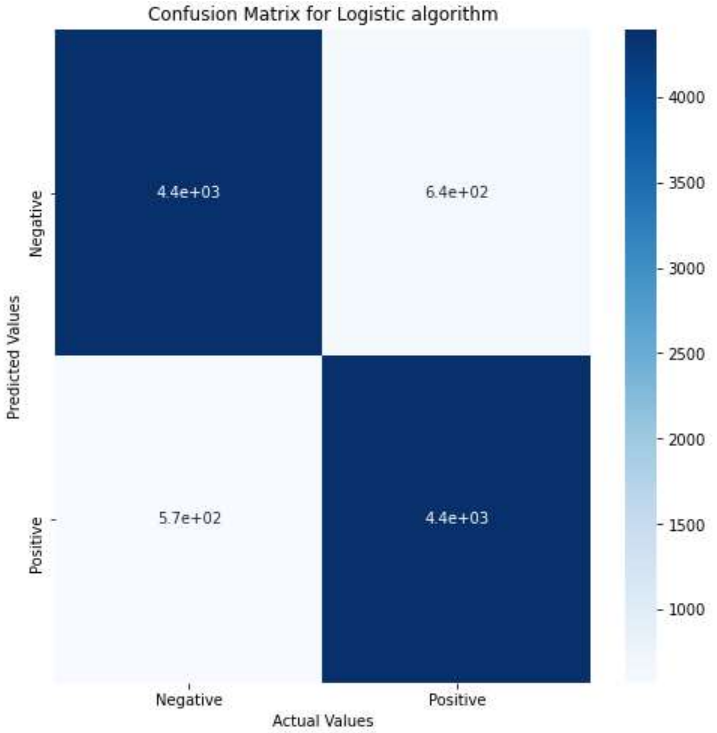
Classification report of Logistic Regression model

```
from sklearn.metrics import classification_report
print(classification_report(test_y,y_pred))
```

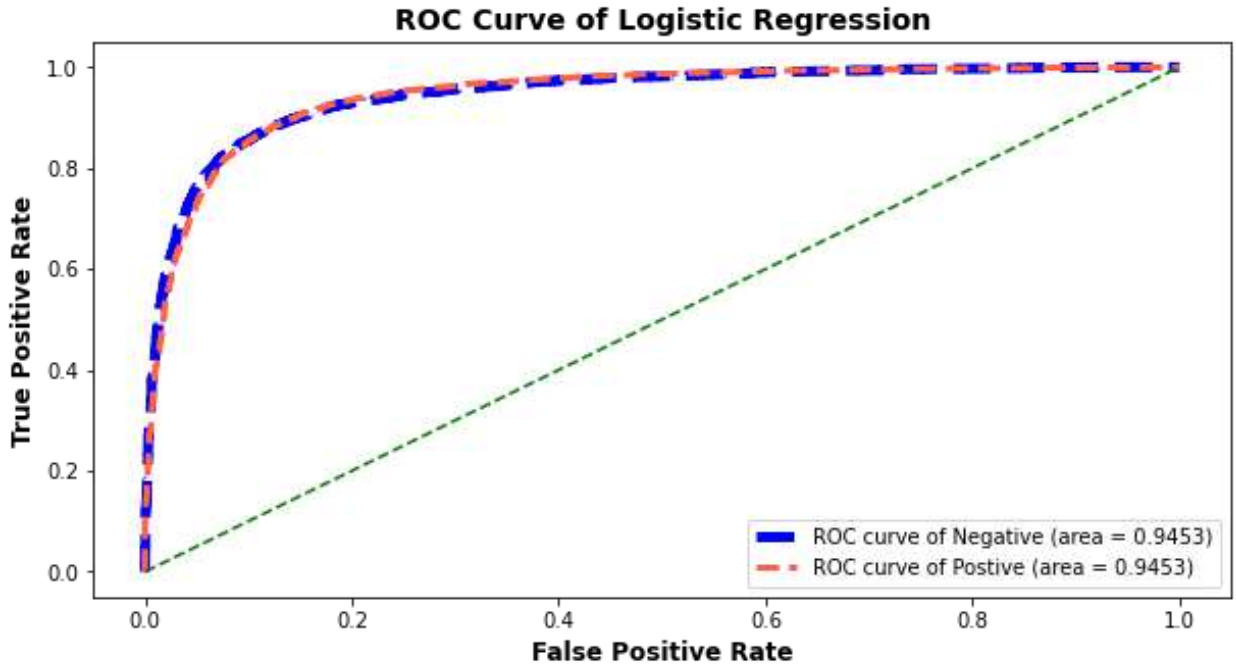
```
              precision    recall  f1-score   support

           0       0.88      0.87      0.88      5034
           1       0.87      0.88      0.88      4966

    accuracy                           0.88     10000
   macro avg       0.88      0.88      0.88     10000
weighted avg       0.88      0.88      0.88     10000
```

Confusion matrix for logistic regression model

```
import matplotlib.pyplot as plt
plt.figure(figsize=(8,8))
sns.heatmap(data=lr_cm, annot=True, cmap='Blues', xticklabels=['Negative', 'Positive'], yticklabels=['Negative', 'Positive'])
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.title("Confusion Matrix for Logistic algorithm")
plt.show()
```



ROC Curve Analysis for logistic regression model




```
AUC Score is:  {'Negative': 0.9453297120458849, 'Postive': 0.9453297120458
85}
```

Test the Logistic regression model

```
def predict_sentiment(review):
    review = re.sub(pattern='[^a-zA-Z]',repl=' ', string=review)
    review = review.lower()
    review_words = review.split()
    review_words = [word for word in review_words if not word in set(stopwords.words('english'))]
    ps = PorterStemmer()
    final_review = [ps.stem(word) for word in review_words]
    final_review = ' '.join(final_review)
    temp = cv.transform([final_review]).toarray()
    return lr.predict(temp)
```

```
#Predicting values
review = 'Do not read this, if you think about watching that movie, although it would be a waste of time'
if predict_sentiment(review):
    print("This is a Positive Review")
else:
    print("This is a Negative Review")
```

```
This is a Negative Review
```

```
#Predicting values
review = 'is such a great musical because it deftly blends the contrasting styles of film and stage. During a dazzling opening se
if predict_sentiment(review):
    print("This is a Positive Review")
else:
    print("This is a Negative Review")
```

```
This is a Positive Review
```

Movie reviews sentiment analysis using deep neural network

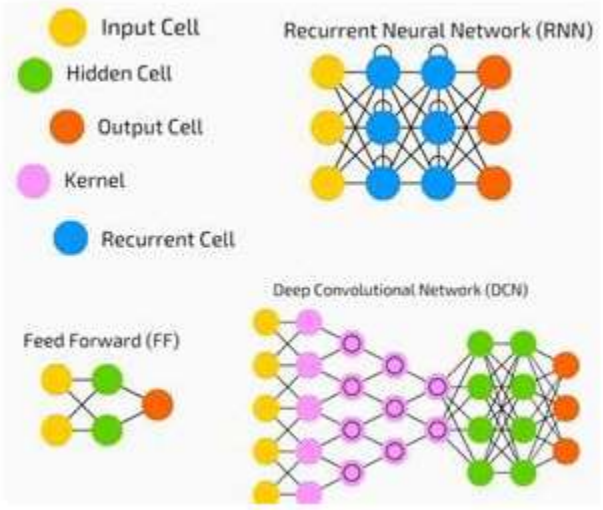
Aim – In this project we have use sentiment analysis using deep neural network.

Dataset Description – we have use 15k data set for training and 5k data set for testing in movie review data set.

Deep Neural Network – In the deep neural network is an artificial neural network with multiple hidden layer between the input and output layer. Equal shallow artificial neural network, Deep neural network can model complex nonlinear relationship.

The main aim of a neural network is to gain a set of input, perform progressively complex calculation on them and give output to solve real world problem like classification. We banned us to feed forwarded neural network.

We have an input and output and a flow of sequential data in a deep network.



Neural network are widely use in supervised learning and reinforcement learning problem. These network are based on a set of layers connected to each other.

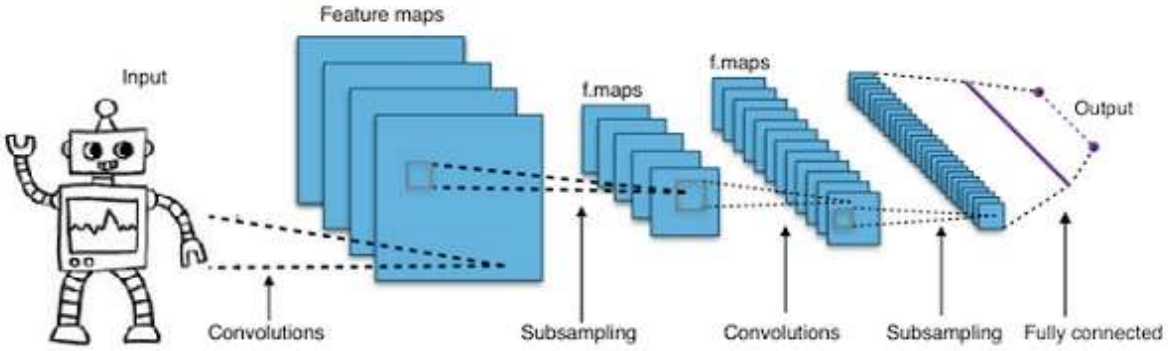
In the deep learning the number of hidden layer mostly nonlinear can be large say about 1000 layer.

Convolutional Deep neural network - we increase the number of layers in a neural network to makes it deeper. It increase the complexity of the network and allow us to model function that are enough complicated. The number of weight and biases will exponentially increase. As a matter of fact, learning such difficult problem can become impossible for normal neural network. Its lead to a solution the convolutional neural network.

Cnn are extensible use in computer vision have been applied also in acoustic modelling for automatic speech recognition.

The think behind convolutional neural networks is the idea of moving filter which passes through the image. It moving filter or convolutional applied to a certain neighborhood of nodes which for example may be pixels, hence the filter applied in 0.5 x the node value.

The noted researcher Yaan Lecun pioneered convolutional neural network. Facebook as facial recognition software uses these nets. The convolutional neural network are multilayer neural network. The layer are sometime up to 17 or more and assume the input data to be images.



The convolutional reduce the number of parameter that need to be tuned. CNN handle the high dimensional of raw images.

Model Evolution

We have use neural network with Convolutinal Neural network and find the training accuracy is 0.96 and testing accuracy is 0.87, in movie review data set.