Report

Introduction :

Sentiment analysis is becoming more and more popular for businesses looking to understand the feelings of their customers because to developments in deep learning and other disciplines of technology.

Businesses now categorise words into positive, negative, and neutral groups by using statistical analysis, natural language processing, and text analysis to determine the sentiment of each word. The most successful businesses recognise the value of knowing their consumers' thoughts and how, what, and why they are saying things.

Sentiment analysis is the study of a customer's positive or negative textual sentiment. The efficiency of sentiment analysis algorithms has grown with recent advances in deep learning and machine learning. Sentiment Analysis is used for many purposes such as Brand Monitoring, Market Research and Analysis etc. In my report I’ve done sentiment Analysis on the SST2 dataset. The Stanford Sentiment Treebank is a corpus with fully labeled parse trees that allows for a complete analysis of the compositional effects of sentiment in language.

Literature Review :

Sentiment analysis is necessary for in-depth research in a number of real-world applications. for instance, through product analysis, ascertain which features or attributes of a product, in terms of quality, appeal to consumers.

The proposed framework for sentiment analysis based on an ensemble of LSTM and CNN models in one of the works Sentiment Analysis is Done Using Ensemble of CNN and Bi-LSTM Models. Glove embeddings are used to represent each word in reviews. The CNN and LSTM models are then given the embeddings to make predictions. The final predictions are then calculated by averaging the anticipated scores from the CNN and LSTM models.

There is also an existing work based on XLNet . XLNet is a generalised AR pretraining technique that combines the benefits of AE and AR approaches with a permutation language modelling goal. XLNet's neural architecture, which integrates Transformer-XL and meticulously designs the two-stream attention mechanism, is built to function perfectly with the AR goal. On a variety of tasks, XLNet significantly outperforms prior pretraining targets.

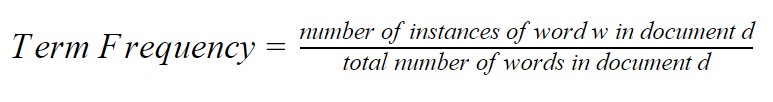
Methodology :

The Methodology mainly focuses on the text data preprocessing and feature Learning. Before getting into the data preprocessing it is essential to know what actually the SST2 Dataset consists of, SST stands for Stanford Sentiment Treebank. It is introduced by Pang and Lee (2005) and consists of 11,855 single sentences extracted from movie reviews. It was parsed with the Stanford parser and includes a total of 215,154 unique phrases from those parse trees, each annotated by 3 human judges. Binary classification experiments on full sentences (negative or somewhat negative vs somewhat positive or positive with neutral sentences discarded) refer to the dataset as SST-2 or SST binary.

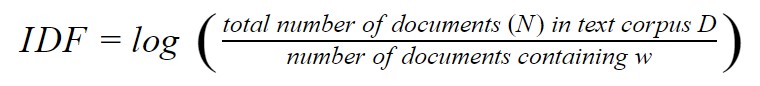
The training set of this dataset consists of 67.3k rows and three coloumns namely sentence,labels and idx. I’ve used the training set data only to train and test the model in the 80-20 ratio i.e 80% training data and 20% test data.

First it is required to check whether the data contains any null values or not if yes , then we have to resolve those in order to get a better accuracy of the model. Then the next step in data preprocessing would be removing html tags, urls, removing punctuations like . , ! $( ) \* % @, removing stopwords ( stopwords are words which are mostly repeated like ‘I’ ,‘you’, ‘we’ etc.. which doesn’t contribute anything to the meaning of the context.) , lower casing and then we have tokenize the sentences.

We can also some advanced techniques to improve the accuracy of the model such as stemming and lemmetizing words. These two methods return the word to its root form (for example if there is a word walking it’s root word would be walk etc. ).  
  
When we are done with the data processing we can split the data into training and testing dataset. But the text form of the data cannot be fed into the model directly it should be converted into some numeric values such that the model understands the data. So here Tf-Idf comes into picture. It is generally the multiplication of the probability of the Term frequency and and the inverse document frequency. It is to decide how relative is the particular word to the respective document.

The phrase term frequency refers to a word's frequency measure (w) in a document (text) d. It is computed by dividing the total number of words in document d by the number of instances of word w. Term frequency is a measure that shows how frequently a word appears in a document in relation to the total number of words in the document. There is always the same denominator. 

IDF provides a word's relevance as a numerical value. Inverse The total number of documents (N) in a text corpus D divided by the number of documents containing word w is the definition of document frequency of word w.



The product of TF and IDF is the TF-IDF. TF-IDF is usually one of the best metrics to determine if a term is significant to a text. It represents the importance of a word in a particular document.

Now, The data is ready to train the model, the model that I’ve trained is Naïve Bayes Classifier. When it comes to sentiment analysis categorization, Naive Bayes is a very straightforward set of probabilistic algorithms that determines the likelihood that a certain word or phrase should be regarded as positive or negative.

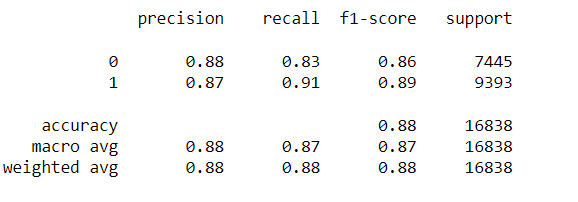
This is basically how the Bayes theorem operates. The likelihood that A, assuming B is true, is equal to the likelihood that B, assuming A is true, multiplied by the likelihood that A is true and divided by the likelihood that B is true:

P(A|B)=(P(B|A )\*P(A))/P(B)

When techniques like lemmatization, stopword removal, and [TF-IDF](https://monkeylearn.com/blog/what-is-tf-idf/) are implemented, Naive Bayes becomes more and more predictively accurate.

Evaluation :

The Evaluation metrics is done based on precision, recall, F1 score and Accuracy of the model.



Conclusion :

So, by choosing proper data processing techniques and some advance techniques like stemming and lemmetization and choosing the perfect model model for the dataset can achieve a better accuracy. Here, the achieved accuracy is 88% on the SST2 dataset.