Movie Reviews Sentiment Analysis: Milestone Report

### Problem statement:

### Sentiment relates to the meaning of a word or sequence of words and is usually associated with an opinion or emotion. And analysis? Well, this is the process of looking at data and making inferences; in this case, using machine learning to learn and predict whether a movie review is positive or negative. Companies use sentiment analysis in a variety of settings, particularly for marketing purposes. Uses include social media monitoring, brand monitoring, customer feedback, customer service and market research (“Sentiment Analysis”).

### Data Wrangling:

### We’ll be using the [IMDB movie dataset](https://www.kaggle.com/c/word2vec-nlp-tutorial/data) which has 25,000 labelled reviews for training and 25,000 reviews for testing.

### Here the features are the words used in the review and it’s important to clean the reviews.

### We use regular expressions in order to remove special characters and unwanted data. Also we would fetch the reviews from the html by using beautiful soup in case the review is referred to html.

### We need to remove the stop words from the reviews as they don’t hold any value in factorizing the sentiment and would cause noise in the models due to their high frequency

### We generally apply lemmatization or stemming in NLP in order to bring the words to their root format. We will be using Porter Stemmer and also WordNet Lemmatize in deep Learning model to derive words to their root format

### Exploratory Data Analysis:

### Creating word cloud is the one of the simplest ways to understand the words used for positive and negative reviews

### Word Cloud for positive reviews:

### C:\Users\611242028\Desktop\download.png

### Word Cloud for Negative reviews:

### C:\Users\611242028\Desktop\download.png

### Creating the ML model:

### Before building up the model it’s important to convert the NLP text to numerical data for the model to process.

### We have used both count vectorizer and Term Frequency-Inverse Document Frequency (td-idf) techniques and applied on Models where td-idf scored similar in accuracy although it’s much better technique for numerical conversion.

### We started to apply on regular machine learning models:

|  |  |
| --- | --- |
| ML Model | Accuracy |
| Logistic Regression | 73.8 |
| Support Vector Machines | 49.2 |
| Multinomial Naïve Bayes | 74.8 |

### We can observed that both logistic regression and multinomial naive Bayes model performing well compared to linear support vector machines.

### We can improve the performance by using neural networks models in deep learning. In this project we will using Long Short Term Memory (LSTM model) using Keras library.

### Below is the bi-directional LSTM model summary:

### 

### We have used 3 epochs while training and attained accuracy of 86% which shows the strength of deep learning.

### Conclusion:

### We have achieved much better accuracy with LSTM model. But we can achieve better results by applying more epochs which require good system configuration

### Also we can have a go at Pytorch framework to see if it yields better results than Keras