MET CS 777 – Big Data Analytics
Term Project Proposal
Muniba Shaikh
18/04/2022

Sentiment Analysis of IMDB dataset

Data Set Description:
IMDB dataset have 50K highly polar movie reviews for Text analytics. The dataset has two features: review and sentiment (positive and negative).

Data Set Link https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews?select=IMDB+Dataset.csv

Research Question:
My goal is to Perform text analytics on the IMDB dataset. I will build a classifier that can automatically determine whether a review about a movie is positive or negative by looking only at the review text.

Implementation:
I will use Spark's MLlib to build logistic regression and SVM classifiers for text analytics on train dataset and compare their accuracy and computation time. I will also use a feature selection method to reduce the problem's dimensionality.

Evaluation:
I will evaluate the model accuracy using the test dataset and calculate the performance metrics including F1-measure and confusion matrix.