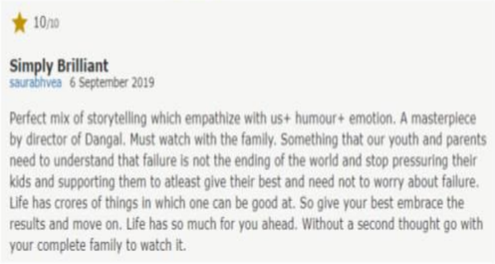
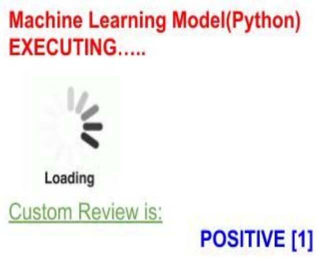
CLUSTER INNOVATION CENTRE (CIC)

(UNIVERSITY OF DELHI)

3rd Floor, Rugby Sevens Building, University Stadium
G. C. Narang Road, University of Delhi, Delhi-110007

PROJECT SUMMARY

Project Title	IMDB Movie Review Sentiment Analysis
Project Mentors	Prof. Shobha Bagai
Project Members	Hrithik Kumar Verma (11817) Kavya Jaiswal (11819) Ritik Arya (11829)
Abstract	Sentiment Analysis, which is otherwise called Opinion Mining,a field inside Natural Language Processing (NLP) utilizing Machine Learning algorithms that showcases the writer's emotions and extract opinions from within the text. Presently, the arena of sentiment analysis is very vast having many real-life applications such as social media monitoring, movie's character predictions, marketing analysis propaganda, customer reviews and product analytics. Sentiment analysis is a field developing day by day and pertaining to great interests nowadays. Here, we have demonstrated the IMDB movie reviews' sentiments and their analysis as positive review or a negative review. As the growing data size is advancing in many domains, it has become indispensable and imperative to extract opinions, feelings, attitudes from review sites, particularly, in our case,movie reviews.
Summary	A dataset of 50,000 movie reviews taken from IMDb for the review analysis. The data was taken and compiled from IMDB. The dataset contains 50,000 reviews split equally into 25,000 train and 25,000 test sets. Thus the reviews have an equal share of positive and negative labels (25,000 positive and 25,000 negative). An additional 50,000 unlabeled documents is added for unsupervised learning that is in validation dataset. IMDb rates the movies on a scale of one to ten based on the reviews provided by the users. The labeled review of anything with <= 5 stars is marked as negative and anything with > 5 stars are marked as positive. Thus reviews with neutral sentiments or opinions are not considered in the unsupervised set, reviews of any rating are included and there are an even number of reviews > 5 and <= 5.There are two top-level directories [train/ test/] corresponding to the training and test sets. Each contains [positive/ negative/] directories for the reviews with binary labels positive and negative. In addition to the review text files, we include already-tokenized bag of words (BoW) features that were used in our project.

	The preprocessing of data includes the following: ‣Cleaning the data ‣Stemming ‣Negation handling ‣Term frequency-Inverse Document frequency ‣Word Embedding ‣ Performing Tokenization & Segmentation,Noise Removal (Remove stop words),Lemmatization & Normalization. Various Machine learning models are used and the corresponding libraries performing those tasks in python are imported and implemented.
Result Images	 ★ 10/10 Simply Brilliant saurabhvea 6 September 2019 Perfect mix of storytelling which empathize with us+ humour+ emotion. A masterpiece by director of Dangal. Must watch with the family. Something that our youth and parents need to understand that failure is not the ending of the world and stop pressuring their kids and supporting them to atleast give their best and need not to worry about failure. Life has crores of things in which one can be good at. So give your best embrace the results and move on. Life has so much for you ahead. Without a second thought go with your complete family to watch it. Machine Learning Model(Python) EXECUTING..... Loading Custom Review is: POSITIVE [1]
Result and Conclusion	Using the above machine learning models efficiently and implementing python code for the same, we have successfully hovered the way of Sentiment Analysis and Prediction of IMDB movie reviews. The training set has two columns-movie review and their associated labels of 1for POSITIVE review and 0 for NEGATIVE review.In the FEATURE ENGINEERING section, we take the preprocessed texts as input and calculate their TF-IDF thus vectorizing the review arrays. We retain 10000 features per text in deep learning section. The EPOCH value is 30. Next, we take the 10,000 dimensional tf idf as input, and keep the 2000 dimensions that correlate the most with our sentiment target. The

	corresponding words,then make sense.We had chosen a very simple dense neural network using tensorflow python module, performing binary classification. Then we trained the model and validated it.. We first centralized the probabilities and predictions with the original train and validation data frames and thus measured the accuracy of the predictions afterwards. The accuracy of our machine learning model is 88% approximately due to the huge dataset taken.
Future Prospects	We have not considered the reviews which showcases neutral emotions in the dataset. The positive reviews and negative reviews has been demonstrated exclusively.We are looking forward to include complex words and words with mixed sentiments provided instantly by the user in future amendments of our project.