Project Proposal

Sentiment Analysis of IMDB reviews

Dhruvil Parikh
(013772720)
dhruvilbhaveshbhai.parikh@sjsu.edu

Maulin Bodiwala
(013733590)
maulin.bodiwala@sjsu.edu

Aditya Patel
(012570857)
aditya.patel@sjsu.edu

Charles Patel
(012570870)
charleskumar.patel@sjsu.edu

***Abstract*— The advent of Web 2.0 has led to an increase in the amount of sentimental content available in the Web. Social network such as Facebook, Twitter, LinkedIn etc., Critic websites like IMDB, Rotten Tomatoes are rich in opinion data and thus Sentiment Analysis has gained a great attention due to the abundance of this ever growing opinion data. As these data is user generated and in text format it is very noisy and requires significant amount of pre-processing and then we will apply machine learning algorithms for classification between positive review and negative review.**

***Keywords—* Sentiment Analysis Machine Learning, Classification, Naïve Bayes, Logistic Regression, Support Vector Machine, Natural Language Processing**

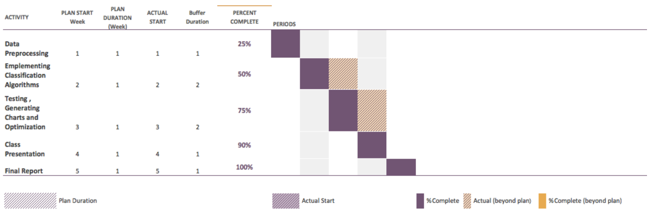
I. Data Set

We have gathered around 28000 raw movie review data without any preprocessing from IMDB website. These data is from various useres and nearly half of that contains positive review and other half contains negative review. As IMDB doesn't catagorise them as positive or negative we have chosen IMDB rating as our reference. IMDB rating is on the scale of 10 with 1 being the lowest and 10 being the heighest. The reviews with 8,9 or 10 as rating are positive reviews and the reviews with 1,2 and 3 are negative reviews.

II. Approach

The data collected by us is raw data straight from users which contains many stop words like (a, the, and etc.) we will pre process these data using NLTK (Natural Language Toolkit) and then feed it into different classifiers. Initially we will use Naïve Bayes, Logestic Regression, Support Vector Machine and K nearest Neighbours for classification. We will use Bag of words model for representating the data so our input vector will have the dimention equal to the length of dictonary. Our dictonary will contain all the words in all the reviews. We will be using MAP estimate instead of MLE because there is a high probablity that the testing data may contain many words which are never seen in training data.

III. Gantt Chart



ACTIVITY	PLAN START Week	PLAN DURATION (Week)	ACTUAL START	Buffer Duration	PERCENT COMPLETE
Data Preprocessing	1	1	1	1	25%
Emplementing Classification Algorithms	2	1	2	2	50%
Testing , Generating Charts and Optimization	3	1	3	2	75%
Class Presentation	4	1	4	1	90%
Final Report	5	1	5	1	100%

References

[1] IMDB website - https://www.imdb.com/

[2] Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies Mr. B. Narendra, Mr. K. Uday Sai, Mr. G. Rajesh, Mr. K. Hemanth, Mr. M. V. Chaitanya Teja, Mr. K. Deva Kumar

[3] Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier by Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose.

[4] Sentiment Analysis of Movie Review Comments by Kuat Yessenov, Sasa Misailovic

[5] Deep learning for sentiment analysis of movie reviews by Hadi Pouransari and Saman Ghili

[6] Link To Gantt Chart - https://drive.google.com/drive/folders/1jF2L1hau-MQ1UtVEXaYH_LRfSBsCbwY_?usp=sharing