Group 15
Lavish Gulati - 170101082
Vakul Gupta - 170101076
Kartik Gupta - 170101030
Chirag Gupta - 170101019

CS565 - Intelligent Systems and Interfaces - Problem Description

1 - PROBLEM STATEMENT

To analyze user reviews on IMDB rated movies using sentiment analysis so that the model achieves the following:

1. Given a user review, it predicts the user rating on that review.
2. Given a corpus of user reviews on a specific movie, it predicts the movie rating.
3. Given a user rating and a movie, it is able to generate a user review.

NOTE: Task 3 could also be used as part of a chatbot implementation.

A part of this project would be to generate an extensive corpus containing the following IMDB data (movie names, movie ratings, user reviews and user ratings) and use it to train a model using feature extractions, representations, and classification techniques to achieve the aforementioned tasks.

The final deliverable would be a product which includes the self-generated dataset and the trained models, including an extensive analysis of the state-of-the-art techniques and an attempt to improve upon these techniques.

2 - MAJOR CHALLENGES

- **Corpus Generation**: A major challenge to IMDB movie review sentiment analysis is an absence of an extensive dataset with a wide array of movies (polarized and neutral) ranging from different timelines, genres and movie industries. To the best of our knowledge, we could only find a dataset [2] generated by Stanford containing a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. Hence, we are taking up this task to generate an extensive corpus and publish it for statistical and machine learning analysis.
- **Sentiment Analysis** itself offers a wide variety of challenges such as
  - **Irony and Sarcasm** - In sarcastic text, people express their negative sentiments using positive words. This fact allows sarcasm to easily cheat sentiment analysis models.
  - **Types of negations** - The original meaning of the words changes if a positive or negative word falls inside the scope of negation. In that case, the opposite polarity will be returned.
  - **Word ambiguity** - Impossibility to define polarity in advance because the polarity for some words is strongly dependent on the sentence context.
  - **Multipolarity** - Sometimes, a given sentence or document—or whatever unit of text we would like to analyze — will exhibit multipolarity. In these cases, having only the total result of the analysis can be misleading, very much like how an average can sometimes hide valuable information about all the numbers that went into it.

3 - BRIEF REVIEW OF EXISTING MODELS

On existing datasets, there is some work done by researchers at Stanford University wherein they used unsupervised learning to cluster the words with close semantics and created word vectors. They ran various classification models on these word vectors to understand the polarity of the reviews. This approach is particularly useful in cases when the data has rich sentiment content and is prone to subjectivity in the semantic affinity of the words and their intended meanings.

Apart from the above, a lot of work has been done by Bo Pang [5] and Peter Turnkey [6] towards polarity detection of movie reviews and product reviews. They have also worked on creating a multi-class classification of the review and predicting the reviewer rating of the movie/product.

These works discussed the use of Random Forest classifier and SVMs for the classification of reviews and also on the use of various feature extraction techniques. One major point to be noted in these papers was exclusion of a neutral category in classification under the assumption that neutral texts lie close to the boundary of the binary classifiers and are disproportionately hard to classify.

4 - PROPOSED DIRECTION

To generate the corpus for our project, we would be using tools like Selenium to automate web handling and BeautifulSoup to parse and extract reviews from the IMDB website. We also aim to use multiprogramming techniques to accelerate the data extraction process.

As a part of this project, we aim to study several feature extraction techniques used in text mining e.g. keyword spotting, lexical affinity and statistical methods and understand their relevance to our problem. In addition to feature extraction, we also look into different classification techniques and explore how well they perform for different kinds of feature representations (Bag of Words, n-gram modelling, TF-IDF modelling etc). We will finally draw a conclusion regarding which combination of feature representations and classification techniques (like Naive Bayes, Random Forest, KNN etc) are most accurate for the current predictive tasks.

5 - RELEVANT REFERENCES

[1] Sentiment Analysis – Wikipedia – https://en.wikipedia.org/wiki/Sentiment_analysis

[2] Large Movie Review Dataset – http://ai.stanford.edu/~amaas/data/sentiment/

[3] Andrew L Mass, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng and Christopher Potts (2011). Learning Word Vectors for Sentiment Analysis

[4] Internet Movie Database – http://www.imdb.com/

[5] Pang, Bo; Lee, Lillian; Vaithyanathan, Shivakumar (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[6] Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics.