Sentiment Analysis of Amazon Product Review Dataset, Project Proposal (ECE657a)

Group 29- Ganesh Rajasekar, Atish Telang Patil, Zeba Javedahamadkhan Vakil

University of Waterloo, Canada

1. Introduction

Sentimental Analysis is the process of computationally identifying and categorizing opinions from piece of text and determine whether the writer's attitude toward a particular product is positive, negative or neutral. It is very crucial for Industrial leaders like Amazon with their giant E-Commerce business to understand consumer behaviour and their sentiment to reach out better to them. With recent developments in the field of machine learning and text analytics this could be feasibly achieved. In this project we aim to predict the rating of a review given to a particular amazon baby product using supervised machine learning algorithms. Based on that we will compare the performance of each of the approaches used and comment on the best classifier for such applications.

2. Literature Review

As part of the Literature review and previous research work in the area of sentimental analysis we went through the Pang et al. [4] which was the earliest attempt to classify the document with sentiment instead of topics. Though they had mentioned that the traditional machine learning methods like Naive Bayes, maximum entropy classification, and support vector machines do not perform as well on sentiment classification as on traditional topic-based categorization. The following research papers contradicted those claims:

- **Sentiment Analysis of Yelp's Ratings Based on Text Reviews** [5] :
  In this project the author has tried to apply existing supervised learning algorithms to the Yelp dataset such as Naive Bayes, Perceptron and Multiclass SVM and compared their predictions with the actual ratings. They succesfully concluded that the binarized Naive Bayes combined with feature selection with stop words removed and stemming is the best in terms of sentiment analysis of such datasets. It is to be noted that the yelp dataset features closely resemble the amazon product review dataset that we are trying to address in this project.
- **Sentiment Analysis Using Product Review data** [1]
  The paper tackles the fundamental problem of sentimental analysis i.e. sentiment polarity categorization. The paper considers both the review-level and sentence-level categorization. In the study, data preprocessing is done by extracting all the subjective contents i.e. all the sentences which has at least one positive or negative word. Based on the POS tagging and the negative prefixes, negative phrases are then identified. After which the sentiment scores are computed. For training the classifiers, each training data entry is transformed to feature vector that has binary strings to represent tokens in order to overcome with the curse of dimensionality. Finally for estimation, 10-fold cross validation is applied where the sentiment score is used to identify the positive and negative classes. The sentiment score proves to be a strong feature achieving 0.73 F1 score for review-level categorization and 0.8 for sentence-level categorization. However, the paper addresses two limitations where it cannot perform well i.e. when F1 scores are very low and when there are implicit sentiments. For the categorization, the classification models used in the paper are: Naive Bayesian, Decision Trees and Support Vector Machine.
- **Sentiment Analysis on Large Scale Amazon Product Reviews** [2] :
  In this model they have adopted a supervised learning approach to polarize the unlabeled dataset. They have used active learning to label the data. The data preprocessing is done by tokenization, removing stop words and POS tagging. They have used mix of two kinds of approach for feature extraction. They finally measured the classification performance

based on Precision, Recall, F-measure and Accuracy parameters. Conclusively, it is observed according to the paper that SVM performs the best when large number of datasets are available.

3. Dataset [http://jmcauley.ucsd.edu/data/amazon/]

The Dataset we would be using is from the official amazon product data hosted in the link above and is provided by researchers from UCSD. This Review dataset contains the following features in JSON format:

- reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714
- reviewerName - name of the reviewer
- helpful - helpfulness rating of the review, e.g. 2/3
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)

As part of the data pre-processing and preparing the training data for classifiers we would be first splitting the reviews based on the polarity to positive and negative classes. Also we will use the NLTK in python for the purpose of lemmitizer and removal of stop words to clean the data.

4. Methodology

In this Project using the concepts learnt in the class, supervised learning algorithms and a bit of experimentation we will be applying the following approaches to perform sentimental analysis on the Amazon baby products review dataset. We will compare the approaches individually and analyze their performance metrics using parameters such as accuracy, precision, recall and F1-score. By implementing this project we will try to validate the claims presented in the papers above:

- K-Nearest Neighbour
- Support Vector Machine
- Artificial Neural Networks
- Naive Bayes Classifier
- Decision Trees
- Adaboosting

We would take the help of NLP tool kit and scikit-learn library in the anaconda distribution of python and may use Uwaterloo's GPU server if needed for reducing the computation time.

5. Challenges and Difficulties

The sentimental analysis scheme relies on the occurrences of the sentiment tokens, so detecting purely implicit sentiments would be a difficult task to implement. There could possibly be two examples for an implicit statement:

- The one in which the statement uses neutral words to describe a positive review which makes the judgement of sentiment polarity difficult. For example, sentence like "Item as described", which frequently appears in positive reviews but consists of only neutral words.
- In the other case, the use of irony (i.e. sarcasm) in the statements which changes positive sentiment into negative whereas negative or neutral sentiment might be changed to positive.

The other challenge worth tackling is the way how to treat comparisons in sentiment analysis [3]. For example:

- This is better than old books.
- This is better than nothing.

Here it is more likely to choose first one as positive and neutral for the second one. But to what context this has been said can make a difference. For example, in the first statement if the old books were considered useless in context, then the first statement turns out to be much similar to the second one. However, if no context is provided, these statements look different. To conclude we understand these are some of the problems that we might expect during the implementation of the project and would be more cautious while training our models to the dataset.

References

[1] Xing Fang and Justin Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):5, 2015.

[2] Tanjim Ul Haque, Nudrat Nawal Saber, and Faisal Muhammad Shah. Sentiment analysis on large scale amazon product reviews. In *Innovative Research and Development (ICIRD), 2018 IEEE International Conference on*, pages 1–6. IEEE, 2018.

[3] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[5] Qinxia Wang, X Wu, and Y Xu. Sentiment analysis of yelp's ratings based on text reviews, 2016.