**CE807 – Assignment 2 – Sentiment Analysis**

**School of Computer Science and Electronic Engineering – University of Essex**

Task: 1

Name:

Siddharth Shyamsunder – 1802772

Pranavi Vasa – 1806152

Lecturer: Prof. Ansgar Scherp

Submission Date: 23rd April 2019

Contents

[INTRODUCTION: 3](#_Toc6865908)

[BASIC TERMS: 3](#_Toc6865909)

[BACKGROUND: 4](#_Toc6865910)

[CRITICAL DISCUSSION: 5](#_Toc6865911)

[CONCLUSION: 5](#_Toc6865912)

[REFERENCES 5](#_Toc6865913)

# INTRODUCTION:

Scientific emotion mapping, also termed as Sentimental Analysis or Opinion Mining, is one of the most sought after areas in this technological era. Most markets, especially in the area of sales, e-commerce, Social Media Branding, especially profit their businesses through analyzing and paying a more keener focus on the areas of emotional behavior of their customers. Giving an example of an ecommerce industry, when a user purchases, items of a similar type, say for example a particular brand of shoe. Then it is more likely that he/she may buy another product say track pants from the same brand to match his outfit. In this scenario, seeing a user’s review on a particular brand, can decide, what the customer is more likely to buy further on. If the same customer, purchases a shoe from the band and gives a negative reviews. Then it is not likely that he/she might buy another product from the same band.

In this way, through estimating, the behavioural patterns and opinions from a customer, an organization can be greatly benefitted in identifying how, they have to distribute and recommend products amongst customer, when products must be manufactured and sold in the market to yield highest returns, and in the case of the film industry, which we target upon understanding, what movies, may have to run in the following weeks, based on positive or negative reviews on the customers, also decide if any of the service counters, food, shopping etc, in the theater multiplex, needs to be upgraded, based on the customer’s negative reviews.

In the case of Sentiment Analysis using Machine Learning Techniques, our dataset, would include a training set, wherein, we would, have the opinios of users, as one of our features, and as a target tothis opinion, the corresponding sentiment, Our end, goal is to analyze the opinions from the users, curtailing to what sentiment, they reflect upon, in the case of a movie review, the star ratings, 0-4, that is given bya customer, to judge the quality of a film being showcased, and then make use of this model, to identify the algorithm, that would allow us to have this model be able to successfully evaluate the sentiments as showcased on unseen reviews.

# BASIC TERMS:

Opinion Mining[1][2]: Opinion Mining commonly termed as Sentimental Analysis or Emotional Artificial Intelligence is the term that is used by natural language and computational linguists, to make use of natural language processing techniques, text analytics, and biometrics to identify patterns in human behavior and feedback, to extract opinions from customers, to quantify such extracted information to make meaningful information to base decisions that would ameliorate or destroy the profitability of a business or user satisfaction.

Sentence Level Opinion Mining[1][3]: Sentence Level Opinion Mining, is one of the sub categories of Opinion based Sentimental Analysis. In Sentence level opinion mining, each of the sentences in a feedback acts as an entity to identify, the sentiment being placed on that sentence. Once a sentence has been identified as per a particular Sentiment value, further sentences, are equally analyzed and on this respect, the entire document is summarized into the apt Sentiment value by averaging the sentence.

Apart from Sentence Level Opinion Mining, Sentiment Analysis can also have the Document Level Opinion Mining, and Aspect Level Opinion Mining, paying more emphasis on the entire document or a particular aspect respectively.

Naïve Based Learning[1]: Naive Bayes classifier is one of the most commonly used and simplest and Supervised Learning Classifier. The Naive Bayes model decides the probability of the class not based on the arrangement of the words but rather the distribution of words on the entire document.

# BACKGROUND:

For optimal performance for identifying sentiments through means of Sentence level opinion mining, researchers have proposed, 2 methods of learning approaches, the Supervised Learning Approach and the Dictionary based Approach. Considering, we are making use of class labels to identify Sentiment Analysis, Supervised Learning plays a major role in identifying Sentiment levels for unseen values. Researchers have made use of different Supervised Learning approaches like the Decision Tree Classifier[4], Probabilistic Classifier[1], Linear Classifier[5] and the Rule based Classifier[6].

In the Dictionary based Approach[7] a small set of sentiment words also called as seed words are gathered. These words already have their positive or negative orientation known to the machine Learning consultant. The seed word set expands then to the synonyms and antonyms of the words already preexisting in the seed word list through online dictionaries. This process continues till no new words are found. In the end, excess words are manually cleaned.

The decision tree[4] provides more of a hierarchical approach to sentimental analysis, in this method, decision at each stage of the type of phrase forms the basis of the decision to proceed onto the next stage. The condition acts upon the presence or absence of particular words. In the case of movie reviews, the words opinions are preprocessed based on the importance of the review by making use of the inverse document frequency, and then conditions are applied. We make use of linear classifiers[5], as they work best for textual information, where text is sparsely distributed. Support Vector Machine Classification along with the domain specific knowledge is an ideal algorithm used to address Aspect based opinion mining.

The Probabilistic classifier or the naïve bayes classifier[1] works on a simplistic approach of identifying the probability of words through distribution of the words in the document rather than identifying the position of the words. It makes use of the Bayes theorem to identify if a sentence has a negative sentiment or a positive sentiment. In the Rule based Learning model[6], the data is governed by a set of rules. The right hand side shows the class labels, whereas the left hand side shows the feature space and its condition in the DNF form.

# CRITICAL DISCUSSION:

From the above provided methods, suggested by researchers, it can be arguably considered, that the Dictionary based Approach[7], is a progressively scalar process, wherein the process continues taking up synonyms and antonyms of all the words in the seed words set until, no word is left. As far as this method will be automated, there is risk of a lot of excess baggage of words, which may further not be properly cleaned during manual inspection. Once again, this also depends on the efficiency of the manual inspection.

The decision tree algorithm[4] is based on the choice of condition of absence of words in a document. This method would prove most inefficient, in the case of sarcastic or sentiments without sentiment words. Same would be the case for interrogative sentences.

In Linear Classifier Model[5], the model finds, correlation or patterns in phrases, to properly distinguish the works of one class from that of another, this would require not only making use of a machine learning algorithm, but also understanding the domain specific knowledge required to understand the similarity between words in a sentence to express a sentiment. Using this method of Supervised Learning to identify sentiments, would be rather expensive due to the inclusion of domain specific knowledge, this method was also proven to be avoided.

The rule based learning approach[6] divides the data into a set of rules, one of the major disadvantages of making use of this method is thid method is almost very difficult to make use for in sparse data. Once again like the Linear Learning classifier, works on the identification of similarity in the word, making it an expensive process to consider.

In the end, a method like Naïve Bayes or Probabilistic Learning approach[1] would give the highest value in terms of accuracy, as this method does not include sparsity of data, rather the importance of the data in the document. This can be achieved, by properly preprocessing data using Inverse Domain Frequency or Stemming techniques, and minimul support threshold is required to handle frequency of words in the document.

# CONCLUSION:

Although a variety of methods can be used to identify the sentimentality of words extracted after an opinion mining approach, a naïve bayes approach would be the best suitable in terms of high performance, in identifying the sentimentality of words, as naïve bayes algorithms aresimilistic and commonly used in text analytics, and do not work on the relationship or similarity between 2 different words.

# REFERENCES

[1] Vidisha M. Pradhan, Jay Vala and Prem Balani, “A Survey on Sentiment Analysis Algorithms for Opinion Mining”, International Journal of Computer Applications (0975 – 8887) Volume 133 – No.9, January 2016.

[2] Sentiment Analysis, [Online]. Available <https://en.wikipedia.org/wiki/Sentiment_analysis> [Accessed: April 7th 2019].

[3] G.Vinodhini, R.M.Chandrasekaran, "Sentiment Analysis And Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, June 2012.

[4] Jeevanandam Jotheeswaran, Dr. Y. S. Kumaraswamy, "Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure", Journal of Theoretical and Applied Information Technology, 2013.

[5] Chetashri Bhadane,Hardi Dalal, Heenal Doshi, "Sentiment Analysis: Measuring Opinions", Science Direct, 2015.

[6] Kai Gao, Hua Xu, Jiushuo Wanga, "A Rule-Based Approach To Emotion Cause Detection For Chinese Micro-Blogs", ELSEVIER, 2015.

[7] Richa Sharma, Shweta Nigam, Rekha Jain, "Polarity Detection at Sentence Level", International Journal of Computer Applications, Volume 86- No 11, 2014.