A SURVEY ON TEXT CLASSIFICATIONS IN REVIEW ANALYSIS

AATHIRA R.
P.G. Student, Department of Computer Science & Engineering,
MES College of Engineering, Kuttippuram, Kerala, India

VIJU P. POONTHOTTAM
Assistant Professor, Department of Computer Science & Engineering,
MES College of Engineering, Kuttippuram, Kerala, India

Abstract - With the swift development of digital photography and the wide-spread popularity of social networks, it has become more common that people share their life experiences and opinions using images and videos together with text. Many kinds of research on the analysis of multimedia are been greatly motivated and promoted by the rapidly growing online social media data, which has, in turn, improved the social media by increasing its usage for wide applications, ranging from marketing to election forecasts. Analyzing reviews available on the internet has been helpful for businesses to understand the public opinion and to make necessary improvements. The review analysis undergoes various steps- data acquisition, pre-processing, feature extraction, classification, and summarization. Text classification phase, the most important among them can be implemented by different approaches. This paper provides an overview of the three different approaches for text classification and conducts a comparative study to find the best among them.

Keywords - Text Classification; Review analysis; SentiWordNet classifier; Senti-Lexicon Algorithm; Naïve Bayes Classifier.

I. INTRODUCTION

Whenever a person has to make a choice, an important part of the decision-making procedure is to know the approach of other people towards the same and be aware of their experience. Advice may be attained from friends, relatives or an expert in the field. The internet can be considered as a big source of opinion evaluation with a huge amount of review information available online. Before buying a product an individual always checks the comments, star ratings, likes and customer review it has received. Opinion mining has made a valid position on every subject around us from customer reviews to political views. But, the key challenge for a user today is that the review data available is so large that it is nearly impossible for a person to read and comprehend the large number of reviews available. It is obvious that the user will not be able to read all the reviews and may miss out some critical reviews that are critical to his/her needs. The discovering, analyzing and filtering of the information in the reviews is considered as a difficult job. To deal with this, a sentimental analysis is used.

Sentimental analysis [1] is a technique of processing natural language to evaluate the attitude of a person about a specific subject, product or topic. It is also called subjectivity analysis or review mining. Sentiment analysis is applied to reviews and social media for various applications, ranging from customer service to marketing. Opinion mining is the process of identifying user's opinion about a particular entity from reviews. It involves categorization of the various opinions into positive or negative polarity. Opinion summarization is the process of representation of review information in a short and summarized form. It also involves selecting important aspects and representing related opinions from various reviews. There are many applications of opinion mining such as decision making, recommendation systems, feedback analysis etc. It is one of the popular research areas in text mining and natural language processing.

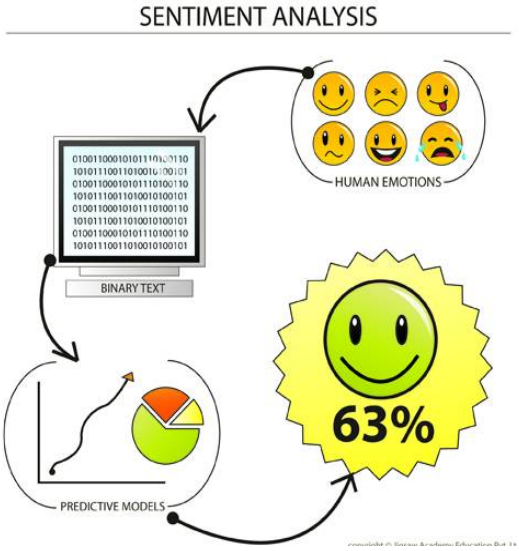

Figure 1. Sentimental Analysis [2]

II. FUNDAMENTALS IN REVIEW ANALYSIS

A basic review analysis process goes through several steps such as acquisition, pre-processing, feature extraction, classification, and summarization as shown in fig. 2.

Data acquisition deals with collecting data or reviews from various sites for the analysis process.

Pre-processing includes sentence separation, tokenization, special character removal, stemming, stop words removal, POS tagging etc. which avoids extra overhead during further processing.

In feature, extraction keywords are mapped into its respective aspects.

Classification phase groups text data into different classes.

Summarization phase generates a brief report after analysis.

Among all these steps the most important phase is the classification phase as it adds to the total outcome. There are mainly three important classification techniques used, which are SentiWordNet, Senti-Lexicon algorithm, and Naïve Bayes classifier.

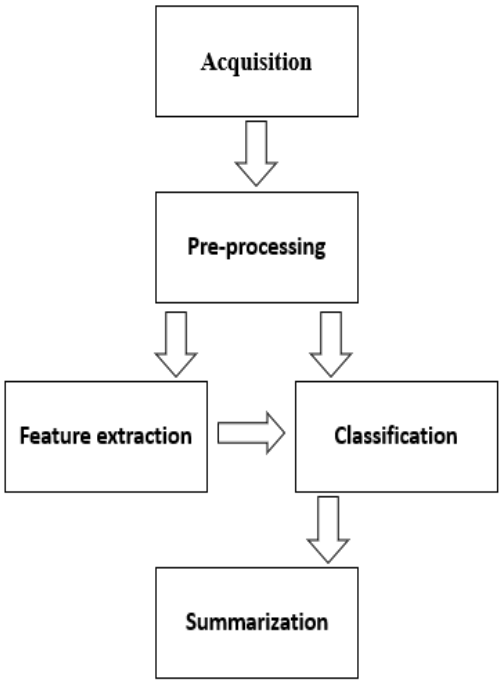

Figure 2. Steps in Analysis

III. DIFFERENT TECHNIQUES OF TEXT CLASSIFICATION

a) SentiWordNet Classifier

SentiWordNet classifier [2] is a lexical approach based on the WordNet dictionary. WordNet is a lexical database for the English language which groups English words into sets of synonyms called synsets. It provides short definitions and usage examples and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. One of the most straightforward approaches is to use SentiWordNet to compute the polarity of the words and average that value.

SentiWordNet = WordNet + Sentiment Information

For each wordnet synsets, the following information is available in SentiWordNet

Positive score Pos(s)

Negative score Neg(s)

Objective score Obj(s)

It is found out that the summation of these three scores equals 1.

$$Pos(s) + Neg(s) + Obj(s) = 1$$

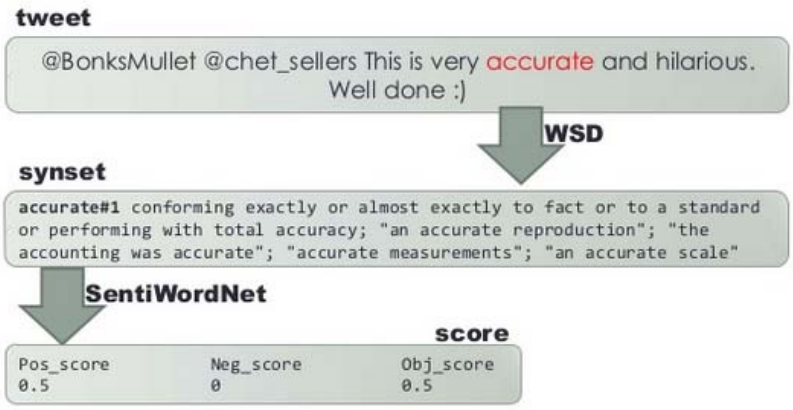

Figure 2. Example of SentiWordNet [4]

- In the given example in fig. 2, it shows how a tweet undergoes analysis to produce a score at the final stage of analysis. Before the analysis, it goes through the WSD phase i.e., Word Sense Disambiguation which means finding out the correct meaning of the word used in this context. So, the accurate synonym of a word is found out in the SentiWordNet analysis.
- Advantages and Disadvantages
  - Advantages

    Usage of synsets has offered different sentiment score for each sense of one word.
  - Disadvantages

    Can misinterpret the sentence with underlying meaning or feeling, which is difficult to comprehend, for example, sarcastic sentences.

a) Senti-Lexicon Algorithm

In Senti-Lexicon algorithm [5], for a given set of words positive and negative scores are calculated. If the sentence contains a negation word such as not, no, wasn't etc. then, the final Score value is reversed and the orientation flips.

Senti-Lexicon Algorithm

Input: D {review data}; PWord{positive words lexicon}; NWord{negative words lexicon}; PEmoticon{ positive emoticon lexicon}; Nemoticon{negative emoticon lexicon}; NegationW{negation words lexicon}

Variables: PWordScore{positive word score}; NWordScore{negative word score}; PosEmoScore{positive emoticon score}; NegEmoScore{negative emoticon score}

Output: Score {Final sentiment Score}; Sentiment {positive, negative, neutral}

Steps

(1) Pre-processing and data cleansing

(2) For each Review Sentence(S) in D do

Split D into separate words (W) and emoticons (E)

(3) For each W in S do

if (W == PWord) increment PWordScore

elseif (Word==NWord) increment NWordScore

end for

(4) For each E in S

if (E== PEmoticon) increment PosEmoScore

elseif (E==NEmoticon) increment NegEmoScore

end for

(5) Determine the total Score

Score= (PWordScore + PosEmoScore) -(NWordScore +NegEmoScore)

(6) Determine the role of negation in the review

if (NegationW==True) inverse the polarity of the Score variable

(7) Display Final Sentiment

if (Score >0) Sentiment=Positive

elseif (Score <0) Sentiment=Negative

else Sentiment=Neutral

end for

Result: The result of the analysis may be obtained as positive, negative or neutral review based on the Score values.

- Advantages and Disadvantages
  - Advantages

    Simple, versatile and feasible.
  - Disadvantages

    Performance should be improved.

b) Naïve Bayes Classifier

Naive Bayes classifier [5] particularly suites when the domain of inputs is high. It is based on Bayes Theorem which calculates the probability that something will happen, given that something else has already occurred. Bayes theorem gives the conditional probability. For example, a fruit is an apple if it is red, round about 10 cm in diameter. We then study each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the colour, roundness, and diameter features. It predicts membership probabilities for each class such as, the probability that given record or data point belongs to a particular class. The class having the highest probability is considered as the most likely class.

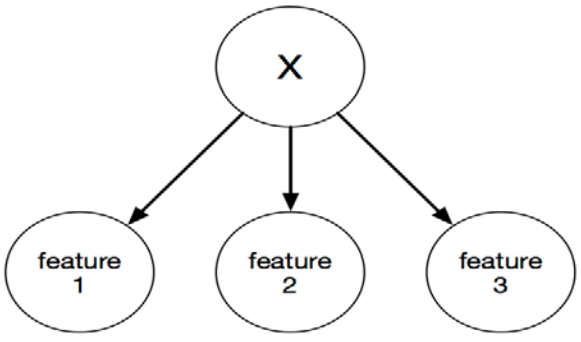

Figure 4. Naïve Bayes Theorem

- Advantages and Disadvantages
  - Advantages

    Highly scalable.

    Only requires a small number of training data.

    Robust enough to ignore serious deficiencies in its underlying naïve probability model.

    Straightforward, uncomplicated and efficient for large datasets.

- Disadvantages

  Naive Bayes can learn the importance of individual features but cannot determine the relationship between features.

IV. COMPARISON

A comparison between the various approaches of text classification is shown in table I. Three parameters are considered for comparison: performance, cost of computation and accuracy.

TABLE I. Comparison of various Text Classification Techniques

Approaches	Performance	Cost of computation	Accuracy
SentiWordNet	High	High	Low
Senti-Lexicon Algorithm	Low	High	High
Naive Bayes Classifier	High	Low	High

Performance:

Senti-Lexicon has low performance compared to SentiWordNet and Naive Bayes algorithm which has been proved to be its main disadvantage.

•Cost of computation:

Naive Bayes Classifier has a low cost of computation since it needs only a small training dataset for parameter estimation.

Accuracy:

SentiWordNet cannot identify sentence with underlying meaning for example, sarcastic sentences. So, it has low accuracy compared to Senti-Lexicon algorithm and Naive Bayes classifier.

V. CONCLUSION

The huge source of reviews available on the internet can be analysed for various applications ranging from customer service to marketing. Review analysis is done to get a summarized form of this enormous amount of reviews. Text classification, the most important phase of review analysis has various approaches by which it can be done. They are SentiWordNet, Senti-lexicon algorithm and Naive Bayes classifier. A comprehensive study of these approaches is conducted. These approaches are compared in terms of performance, cost of computation and accuracy, and it is concluded that the Naive Bayes classifier proved to be the most efficient among them.

VI. ACKNOWLEDGMENTS

We would like to express our sincere gratitude to all the staff members and students of Computer Science and Engineering Department of our college who helped us with their timely suggestions and support. We also express our sincere thanks to all friends who helped us throughout the successful completion of this work.

REFERENCES

[1] Xiaojiang Lei, Xueming Qian, Guoshuai Zhao, Rating Prediction Based on Social Sentiment From Textual Reviews , IEEE Transactions on Multimedia (Volume: 18, Issue: 9, Sept. 2016), Page(s): 1910 - 1921.

[2] Sentiment analysis, [https://viblo.asia/uploads/58039b5e-7d90-4165-9f0b-83fb77792318.jpg].

[3] Chaitali Chandankhede, Pratik Devle, \ISAR: Implicit Sentiment Analysis of User Reviews", 2016 International Conference on Computing Analytics and Security Trends (CAST), Pune, India. Dec 19-21, 2016.

[4] SentiwordNet,[https://image.slidesharecdn.com/20131001bigdataanalysistweetsentiment-131125063458-phpapp02/95/tutorial-of sentiment-analysis-18-638.jpg?cb=1385361545].

[5] Akkamahadevi R Hanni, Mayur M Patil, Priyadarshini M Patil, \Summarization of Customer Reviews for a Product on a website using Natural Language Processing", 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India.