International Journal of Engineering & Technology, 7 (2.33) (2018) 1007-1009

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

An efficient approach towards review analysis using NLP and Watson

Aathira R.[1] *, Viju P. Poonthottam [2]

[1] *P. G. Student, MES College of Engineering, Kuttippuram, Kerala, India*
[2] *Assistant Professor, MES College of Engineering, Kuttippuram, Kerala, India*
**Corresponding author E-mail: aathi0106@gmail.com*

Abstract

A large collection of data is available over the internet which can be used to generate the needful relevant information according to individual needs. Even though the information given about an instance is enough to make a summary about it, the opinions and reviews updated by individuals regarding the instance give a clear idea of what the conclusion made by humankind is. So, analyzing reviews by Sentimental analysis help to identify the human opinion about an instance. Along with the reviews, the images uploaded by users showcase the real-time situations without any edits and can lead to a more specific conclusion. Images are analyzed with the tags generated from them using IBM WATSON. Therefore, taking in consideration both images and reviews will generate a well precise report about the instance. The review analysis is done by Naive Bayes Classifier which is considered as the best choice for text classification.

Keywords: *IBM WATSON; Naive Bayes Classifier; Review Analysis.*

1. Introduction

People now started sharing their life experiences and opinions over social networks which has an increase in its popularity. Research in the large-scale multimedia analysis is been promoted and motivated by the aggressively growing online data. Nowadays people do a lot of research before taking any decisions. The decision making of a person is dependent on the reviews and experience of other people about it and this can be considered as a significant piece of decision making procedure. The advice can be from relatives, friends or an expert in the field. Emotional level analysis has benefited from several applications such as customer service to marketing. With the enormous development in the generation of review information online, the people have now started to consider the internet as a big source of opinion evaluation. Review information of each instance is scattered all around the internet. One of the key challenges faced by the user today is that the review data is so large that it is nearly unrealistic for a person to read and understand a large number of reviews available. It is obvious that the user may not be able to read all the reviews and might miss out some reviews that are critical to his/her needs. The discovering, analyzing and cleaning of the information on the opinion sites are considered as a frightening job which can be solved by natural language processing.

Natural language processing NLP [1] is a way for computers to derive, analyze and understand the meaning of human language in a smart and useful way. It can be used to analyze text, allowing machines to understand how humans speak. This human-computer interaction has stimulated several real-world applications like sentimental analysis, automatic text summarization, topic extraction, POS tagging and more. Sentimental analysis [2] is used to identify the attitude of a person with respect to some topic. It can be defined as a technique for processing natural language so as to evaluate the position, sensitivity or evaluation of the people about a specific subject, product or topic. It is also called subjectivity analysis or review mining. Sentiment analysis is largely applied to reviews and social media for various kinds of applications, ranging from customer service to marketing to determine whether the writer's attitude towards it is positive, negative or neutral. Thus, it can be referred to opinion mining.

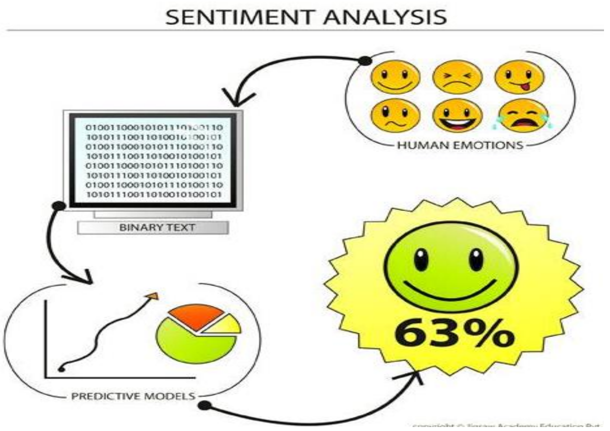

Fig. 1: Sentimental Analysis.

Opinion mining and sentimental analysis, both are almost similar the only difference being that opinion mining extracts and analyses the people's opinion about an entity while sentimental analysis search for sentiment words or expression in a text and then analyze it. The accuracy of the sentimental analysis depends on how well it sides with the human evaluations.

As mentioned the main disadvantage faced by a user is to read and comprehend thousands of reviews about an instance to create a clear summary of it. Therefore, a proposed system is developed using WATSON and Naïve Bayes classifier for the generation of a brief report from thousands of reviews.

2. Related work

Different review analysis systems are implemented using different techniques. Review analysis process mainly goes through five major steps. They are depicted in Fig. 2 and are explained as below.

a) Data Acquisition: Data acquisition refers to acquiring the user reviews from one or many sources. These sources of reviews should be reliable and sufficient in number.
b) Pre-processing: Preprocessing refers to cleaning of textual data in order to avoid excess processing overhead in further processing. Pre-processing involves various steps such as sentence separation, tokenization, special character removal, stop words removal, stemming, POS tagging etc.
c) Feature Extraction: Opinion to aspect mapping was done by maintaining a set of aspect related words. Whenever these keywords were encountered they were simply mapped to the respective aspect.
d) Classification: Text data is classified into various classes.
e) Summarization: A short and summarized report is generated about these reviews.

2.1. Implicit sentimental analysis of user reviews using senti word net [4]

The main motive of the system is to develop an opinion mining application with improved accuracy by following an implicit approach. SentiWordNet dictionary is been used for scoring opinion words. Each word is given a positive or negative score. SentiWordNet is a lexical approach based on the WordNet dictionary, a lexical database for the English language. Synsets are formed by grouping synonyms of English which provide short definitions and usage examples and records the number of relations between each word and its synonym. Thus, SentiWordNet is a combination of words and their thesaurus. Computing polarity of a word and calculating its average is one of the easiest ways in SentiWordNet classification.

- Advantages: Usage of synsets had 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.

2.2. Sentimental analysis of movie review data using senti-lexicon algorithm [5]

The central aim of this work is to perform sentimental analysis on movie review data using Senti-Lexicon algorithm to find polarity of a review as positive, negative or neutral. Technologies are been adopted so that the proposed system can handle negation effect on the reviews and the role of emotions.

In Senti-Lexicon algorithm, for a given set of words calculate positive and negative scores. If the sentence contains a negation word such as not, no, wasnt etc. then, the final Score value is reversed and the orientation flips.

- Advantages: Simple, versatile and feasible.
- Disadvantages: Performance should be improved.

2.3. Summarization of customer reviews for a product on a website using natural language processing [6]

This work creates an android application that assists buyers in online shopping. It analyses and summarizes what other people have experienced about a product from their reviews. A Naïve Bayes classifier is used for review classification. Naive Bayes classifier particularly suits when the range of inputs is high. It states that value of a particular feature is independent of the value of any other feature. It is developed on the basis of Bayes Theorem which calculates the probability that something will happen, given that something else has already occurred. Bayes theorem gives the conditional probability. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The most likely class is the one with the highest probability.

- Advantages of Naive Bayes Classifier: Classifier is highly scalable. Only requires a small number of training data for parameter estimation. It is robust enough to ignore serious deficiencies in its underlying naive probability model. It is straightforward, uncomplicated and efficient for large datasets.
- Disadvantages: Naïve bayes calculate only the probability of something to occur therefore, a hundred percent confirmation on an obtained data cannot be done.

The three technologies are compared and depicted in Table 1.

- Performance: Senti-Lexicon has a low performance compared to SentiWordNet and Naive Bayes algorithm which has been proved to be its main disadvantage.
- Cost of computation: Naive Bayes Classification has low cost of computation since it need 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.

Table 1: Comparison

Techniques	Performance	Cost of computation	Accuracy
Senti Word Net	Low	High	High
Senti-lexicon Algorithm	High	High	High
Naive Bayes Classifier	High	Low	High

3. Problem definition

A given set of reviews are analysed using Naïve bayes classifier to generate a brief report. The existing systems do not analyze images given in different sites along with reviews. Images are important for analysis as it conveys the view or the current situation of a place. Existing systems analyses each and every review at present. All the irrelevant reviews are not avoided from the analysis as they do not add to the summarization which should be precise. Keywords required for the summarization is not identified for the existing systems.

4. Proposed system

Naïve Bayes classifier has proved to be the most preferred text classification technique among the techniques mentioned above. So, this classifier is adapted to analyze the reviews in this proposed system. The flow chart of the proposed system is as shown in Fig. 2.

The proposed system has three main stages : Data acquisition, Data analysis and Report generation. Data acquisition phase consists of gathering the images and reviews about an instance. Images collected about each instance undergoes analysis with IBM WATSON to generate tags. As everyone believes what they see more than what they hear to take a decision, so images can be considered a valuable piece of relevant information about an instance. Images analyzed using IBM WATSON identifies scenes, objects, faces, and other contents. Set of keywords describing images are returned as a response.

The Data analysing phase is where the tags and reviews are analysed using Naïve Bayes classifier to generate a brief report about the instance which is easy to read and comprehend for users.

- Advantages of Proposed System: A quick report about the instance is available which is easy to read and understand.

The proposed system is a method which analyses data more effectively.

5. Performance

The proposed system uses WATSON and Naïve Bayes classifier which works efficiently and securely. Naive Bayes classifier has proved to be the most efficient classifier with high accuracy and high performance. Proposed system takes into consideration the sarcastic sentences and images for analysis which add up to the total performance. This system reduces the workload of reading each and every review to get a summarization about an instance.

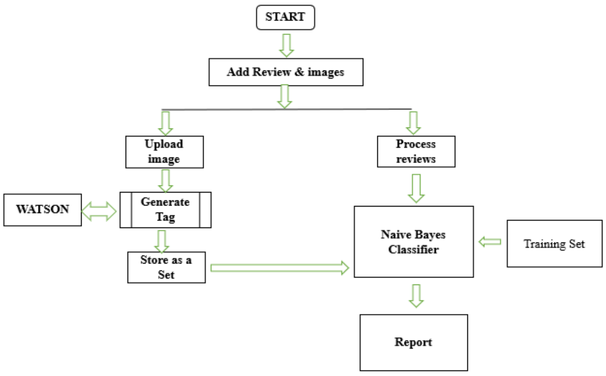

Fig 2: Flowchart of the Proposed System.

6. Conclusion

A system to analyse reviews are proposed using WATSON and Naïve Bayes classifier and performance analysis have shown that proposed system has high performance. It reduces user burden of reading and comprehending thousands of reviews to get a summary about an instance on internet

Acknowledgment

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] Introduction to Natural Language Processing (NLP) – Algorithmia Blog, [https://blog.algorithmia.com /introduction-natural-languageprocessing-nlp/].

[2] 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.

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

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

[5] Deebha Mumtaza, Bindiya Ahujab, "Sentiment Analysis of Movie Review Data Using Senti-Lexicon Algorithm", 2016 2nd International Conference on Applied and Theorectical Computing, 21-21 July 2016, Bangalore, India.

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