An Integrated Framework for Opinion Mining and Summarization of Hotel Reviews

KMA De Abrew[1,#], PPNV Kumara[1], and DU Vidanagama[1]

[1]*Faculty of Computing,General Sir John Kotelawala Defence University, Sri Lanka*
[#]<madhuri.abrew@gmail.com>

Abstract— Customer opinions hold a very important place in businesses, especially for companies and hotels. In last years, opinions have become more important due to global Internet usage as opinions pool. The opinion mining is one of the most popular topics in Text mining and Natural Language Processing. Unfortunately, looking through customer reviews and pulling out information to improve their service is a difficult work due to the large number of existing reviews. It is identified that there is no such an automated system to mine and extract information from consumer opinions in Sri Lanka for hotel industry. The proposed system focuses on the problem of opinion mining, particularly applied to hotel reviews and targets to develop a system that is capable of extracting and classifying different opinions automatically, based on the sentiments polarity. The objectives of this research are: Study the existing problem of opinion mining of Hotel Reviews in Sri Lanka, study the technologies and tools to solve the problem, develop a prototype to solve the identified problem of opinion mining and to test and evaluate the proposed solution. The proposed solution presents a system to mine client opinions, classify them as positive, negative or neutral. This system mine the reviews and summarize them for the ease of decision making.The system uses Sentiment Word Net based method for opinion mining from hotel reviews and sentence relevance score based method for opinion summarization of hotel reviews. The proposed system will save the time of consumers by providing the summarized opinions.Also this website allow users to search for hotels based on category and the location according to the ratings.In addition this will help the higher authorities by making decisions and improve the services based on opinions of customers.

Keywords—**Opinion mining, customer feedback, sentiment classification**

I. INTRODUCTION

The consumers rapidly use Internet for searching and buying products and services. People are interested in using the Internet, because it is much easier than traditional methods. Most of the important information available on the web can be found as the opinions expressed by users, such as through customer product and service reviews. These customer reviews play a major role when making decisions about the purchasing or selecting the products and services(Al-Abdullatif and Kotb, 2014). These reviews provide the consumers with quick access to information which are related to the services and help them to make better decisions faster than ever before. It is difficult for a customer to read all reviews and make a decision due to the availability of the large number of reviews. Furthermore, many reviews are long and have only a few sentences containing opinions about the particular product or service. This makes it hard for a potential customer to read them to make a decision on whether to select that product/service or not(Sixto et al., 2013). If he/she only reads few reviews, he/she may get a biased view (Ding et al., 2008). Customer reviews are the bridge between sellers and buyers. Trustworthiness of online reviews is growing and often, positive customer reviews increase the trust on that product or service. It is important to identify the user opinions and classify them according to the mining techniques. Opinion Mining is a combination of Natural Language Processing (NLP) and Information Extraction (IE) which aims to obtain opinions of the reviewer expressed in positive or negative comments by analysing large number of reviews. The main task of sentiment analysis is to classify the reviews and determine its polarity. Polarity is expressed as positive, negative or neutral. This research focuses on mining and summarization of consumer reviews on hotels in Sri Lanka. The main objectives of this research are: (1) Study the problem of mining and summarization of consumer opinions in hotel industry in Sri Lanka (2) Study the techniques and tools to solve the problem (3) Develop a conceptual model to solve the problem. The research is based on unsupervised approach that determine the sentiment orientation of customer reviews. Sentiment orientation determines the polarity of the reviews and it classifies the reviews as negative, positive or neutral. Sentiment Word Net based method will be used for opinion mining from hotel reviews and sentence relevance score based method will be used for opinion summarization of hotel reviews (Kim and Hovy, 2004). WordNet will be also used to automatically acquire emotion related words (Angulakshmi and ManickaChezian, 2014). This approach

helps the users in decision making by providing the summary of total number of positive and negative reviews and the hotel ratings.

The paper is organized as follows. Section 2 describes the background study. Section 3 describes the proposed framework. Section 4 presents the approach to the .Section 5 presents discussion about the framework together with directions in the future.

II. LITERATURE REVIEW

Existing researches in area of opinion mining are discussed below. One of the most prominent work was done by Saensuk et al. (2015). They proposed a method for mining opinions on smart-phone reviews written in Thai. The method summarized positive and negative polarity of each feature of smart-phones. They collected reviews from smart-phone pages on Facebook. After that they performed word segmentation using a Thai segmentation technique. According to this work they collected words to create three dictionaries named as dictionary A , dictionary B and dictionary C. Dictionary A collected the most commonly misspelled words written on social networks and their spelled words and misspelled words are corrected (as to obtain consistent data). For polarity identification, segmented words of each sentence are compared to polarity words in dictionary B to identify positive or negative words. Feature identification is performed by using the third dictionary (dictionary C).Segmented words of each sentence were compared to feature dictionary C. In this study, the opinion mining process consists of the data pre-processing and the opinion classification using feature-based. From the experimental result, it was shown that the proposed method gave 70.17% of accuracy(Saensuk et al., 2015)

Sharma et al. (2014) proposed a document based opinion mining system that classify the documents as positive, negative and neutral. Negation was also handled in the proposed system and it involved NLP too. The unsupervised dictionary based technique was used in this system. WordNet was used as a dictionary to decide the opinion words and their synonyms and antonyms. The system categorized each document as positive, negative and neutral and in the output it separately presents the total number of positive, negative and neutral number of documents(Sharma et al., 2014). For decision making the output generated by the system would be helpful for the users. They could easily classify how many positive and negative documents are presented. The polarity of the given document is determined on the basis of the majority of opinion words.This experiment was conducted using reviews of movies which showed the effectiveness of the system. 66% accuracy was achieved for this movie review domain.

Haruechaiyasak et al., (2010) proposed a framework for constructing Thai language resource for feature-based opinion mining. The feature-based opinion mining basically depends on the use of two main lexicons, features and polar words. (Haruechaiyasak et al., 2010). Extracting features and polar words from opinionated texts was based on syntactic pattern analysis according to their approach. The evaluation was performed with a case study on hotel reviews. The process started with a corpus which was tagged based on two lexicon types. From the tagged corpus, they constructed patterns and lexicons. The pattern construction was performed by collecting text segments which contain both features and polar words. The lexicons were used for performing the feature-based opinion mining task such as classifying and summarizing the reviews as positive and negative based on their various features. The suggested method has shown to be very effective in most cases.

Another most prominent work was done by Hu and Liu (2004) aimed to mine and summarize all the customer reviews of a product. These researchers only mined the features of the product on which the customers had expressed their opinions whether the opinions were positive or negative. This summarization task was different from traditional text summarization because they did not summarize the reviews by selecting a subset or rewrote some of the original sentences from the reviews to capture the main points as in the classic text summarization.In this research their task was performed in three steps: (1) mining product features that had been commented on by customers; (2) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative; (3) summarizing the results. Here, features broadly meant product attributes and functions. The system first crawled all the reviews, and put them in the review database. Then found those frequent features that many people had expressed their opinions on. Then, the opinion words were extracted using the resulting frequent features, and semantic orientations of the opinion words were identified with the help of WordNet. Using the extracted opinion words, the system then found those infrequent features. In the last two steps, the orientation of each opinion sentence was identified and a final summary was produced. This system used POS tagging (part-of-speech tagging) from natural language processing, which helped to find opinion features (Hu and Liu, 2004).

Another study about Sentiment Analysis Based on Dictionary Approach which was done by Bhonde et al (2015). They have shown that using a dictionary based

approach to compile sentiment words is a famous approach because most dictionaries list synonyms and antonyms for each word. In this approach they proposed a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. Specially, this technique works as follows: First manually collect a small set of sentiment words (seeds) with known positive or negative orientations which is very easy. The algorithm then grows this set by searching in the WordNet or another online dictionary for their synonyms and antonyms. The latterly found words are added to the seed list. Then the next iteration begins. The iterative process ends when no more new words can be found. By this iterative process system can grow the seed list by adding new words. After the process completes, a manual inspection step was used to clean up the list.

III. CONCEPTUAL MODEL

The proposed framework is integrated the opinion mining and summarization. The unsupervised dictionary based technique will be used in this system. It uses Sentiment Word Net based method for opinion mining. WordNet is used as a dictionary to determine the opinion words and their synonyms and antonyms. Also this uses sentence relevance score based method for opinion summarization. This model will be going to implement as a web application where the usersare asked to provide reviews about hotels in Sri Lanka. The reviews will be in sentence form. The overview of the proposed framework is shown in Fig 1.

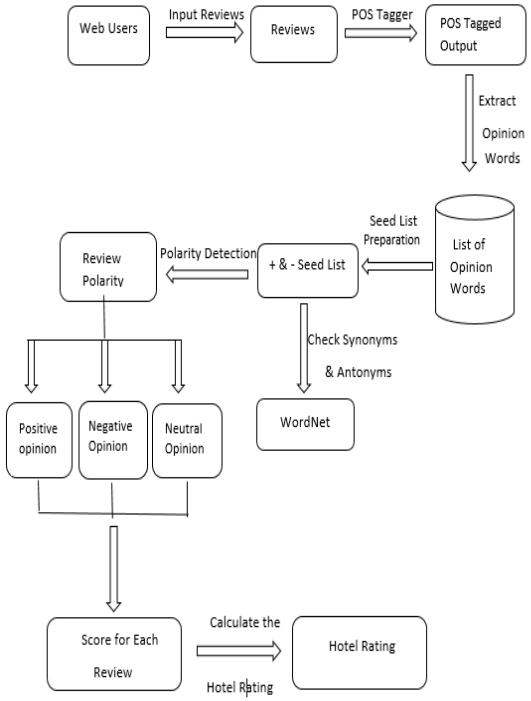

Figure 1. Conceptual Framework

A. Extracting Opinion Words and Seed List Preparation

Seed list initially contains some of the opinion words along with their polarity. From the taggedoutput all the opinion words will be extracted. The extracted opinion words will be matched with thewords stored in seed list .If the word is not found in the seed list then the synonyms aredetermined with the help of WordNet. Each synonym will be matched with words in the seed list. Ifany synonym is matched then extracted opinion word is stored with the same polarity in the seed list.If none of the synonym is matched then the antonym is determined from the WordNet and thesame process is repeated. If any antonym is matched then extract opinion word will be stored with the opposite polarity in the seed list. The seed list will grow every time whenever the synonyms or antonyms words are found in WordNet matches with seed list.

B. Polarity Detection and Classification

The polarity of the reviews are determined with the help of seed list and word net. Based on the majority of opinion words polarity of the reviews is determined. Opinions can be classified in to three groups such as:

- Positive Opinion - If thenumber of positive opinion words is more than the number of negative opinion wordsin the review.
- Negative Opinion - If thenumber of negative opinion words is more than the number of positive opinion wordsin the review.
- Neutral Opinion - If the numberof positive opinion words is equal tothe number of negative opinion words in the review.

If the opinion word is preceded by not, then the polarity of review is reversed.

C. Score Based Method

System gets the user reviews and system depicts the appropriate tags for each word in the sentence. This is the process where we determine the opinion words. After retrieving the opinion words, system will identify the polarity words. At the same time, the polarity of the words are reviewed and assigned the given score to the each polarity word in the seed list. Score is given in 1-5 range. According to that the system calculates the rating of the hotels.

D. POS Tagging

The user reviews are sent to the POS tagger. The POS tagger tags all the words of the reviews to their appropriate part of speech tag .POS tagging is essential to determine the opinionwords. It can be done manually or with the help of POS tagger. POS tagger is used here totag the reviews. Fig.2 shows an example of POS tagging.

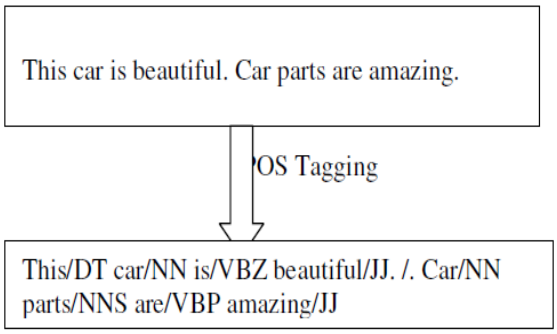

Figure 2. POS Tagging

IV. APPROACH

This framework will be implemented as a web application which will collect feedback reviews thatwill be posted by various users classified as negative, positive or neutral. The System will take reviewsfrom various users, based on the opinion, and state whether the posted hotel is good, bad, or worst. The user will be able to login to the system and givereviews about the particular hotel. The role of the administrator is to insert new hotels and add keywords in to the seed list. This systemwill be useful for those who are going to visit a new place and also useful for those who are often travellers.

The system will classify each review as positive, negative and neutral and presents the total number of positive, negative and neutral number of reviews separately. The output generated by the system will helpful for the users in decision making, which can easily identify how many positive and negative reviewswill be presented. The polarity of the given reviewwill be determined on the basis of the majority of opinion words.In the system there will be a web page where user has to select the category and the city to find the best hotel, and then the hotels will be appeared in descending order based on the hotel score calculation.

The hotel management also interests in user comments that automatically and systematically collects and summarizes the relevant information from the web would be advantageous and perhaps even more useful than the paper formats which nowadays many hotels use for gathering feedback from their guests. So the hotel management can get more benefits from this system. Most of the time travel agencies and hotel booking services often only publish scalar ratings, e.g. scores between 1 and 5. Such scores are not helpful for hotel managers as the numeric value does not provide information of what guests actually considered positive or objectionable. Also, the numeric scores are not comparable: when a 3-star hotel receives a higher score than a 4-star hotel that does not imply that the one is better than the other. For hotel managers and the consumers' textual user comments would be much more significant than the numeric scores since they would be interested to know what the consumers exactly commented on and how they thought of it. It will help others to decide which hotel is best to accommodate before they reach the place. Also it will help the hotel managers to manage their hotel standards accordingly.

A. System Design

This research will produce a web-based user interface that enables users put the reviews about the hotels and to find the best hotel for them in a simple way. This website will use a very simple method to present information to users and it will help users to make decisions very quickly and in an easy manner. It will help users to find the best hotel for them. It will contain two main pages for users: one for users to find the hotels according to the ranking and the other for users to put reviews about hotels. There will be also an administrator page for website administrator who controls and manages the whole website. Administrator can add new hotels to the website and control them. At the administrator page, the administrator can manage categories and hotels in different ways. At the user page, the user has to select the category and the city to find the best hotel, and then hotels will appear for them in descending order based on the hotel score calculation. The result table will contain the city where the hotel is located in and the hotel names with their rating percentages. So the users can decide the best hotel according to their requirements. Usually the hotel with the highest percentage is the best one. The user can also click on the hotel name to go to the hotel page.

V. EVALUATION

User reviews of the hotels will be used to perform the experiment. All the collected reviews will be applied to the proposed system which classifies the reviews as positive, negative and neutral. To compute how well the system classifies each reviews as compared to human decision, all the reviews will be manually classified and the corresponding opinion will be determined. This simply means we manually read all the reviews. For each review, we gives a polarity. The polarity of these reviews will compare with the polarity of reservation websites reviews to check the accuracy of the system. All the results generate by our system will be compared with the manually calculated results. This performance evaluation

will be done with a collection of hotel reviews obtained from various hotel reservation websites.

VI.FUTURE WORK

The authors will plan to further improve and refine the techniques, and to deal with the outstanding problems identified. Also some efforts will be done to improve the techniques so that it will deal with the reviews contain relative clauses like not only-but also , neither-nor, either-or etc. There will be a module where user can book rooms of desired hotel through online. In future, this research will improve the online hotel booking process by providing customers to reserve hotels online. There will be another module which will the facilities provided by the hotels. Also in this research it is only used adjectives as indicators of opinion orientations of sentences. However the verbs and nouns can also be used for the purpose, e.g., "I like the feeling of the hotel", "I highly recommend this hotel to anyone". So it will be planned to concern about this issue in the future. Another fact is to study the strength of opinions. It is also very important thing. Some opinions are in a very strong manner and some are quite mild. So the fact of highlighting strong opinions (strongly like or dislike) can be very useful for users. This will be addressed as a further work.

VII. DISCUSSION AND CONCLUSION

Due to the availability of huge amount of user-generated content, such as booking websites, forums and blogs, opinion mining has become an interesting research area. Nowadays opinion mining is very important among the common people to a businessman, everyone who is dependent on the Web. This paper proposes a set of techniques for mining and summarizing hotel reviews based on data mining and natural language processing (NLP) methods. The opinions expressed on the web helps the users to determine which hotel is good for them and it helps the users to determine what the best hotel is for them. Also those reviews helpful to the hotel managers because they can get to know of what customers thinks about their hotels. So the objective of this paper is to determine the polarity of the hotel reviews and calculate the rating of hotel reviews.This system mine the reviews and summarize them for the ease of decision making .So it is necessary to mine this large number of reviews and classify them, so it is helpful for consumers to read and make decisions. In addition, this website allow users to search for hotels based on category and the location according to the ratings.

ACKNOWLEDGMENT

The authors like to express our great appreciation to all the staff members and colleagues of Faculty of Computing, KDU, for their valuable and precious time, which is generously and highly admired.

REFERENCES

Al-Abdullatif, W., Kotb, Y., 2014. Using Online Hotel Customer Reviews to Improve the Booking Process. Int. J. Comput. Appl. 97.

Angulakshmi, G., ManickaChezian, R., 2014. An analysis on opinion mining: techniques and tools. Int. J. Adv. Res. Comput. Commun. Eng. 3, 7483–7487.

Bhonde, R., Bhagwat, B., Ingulkar, S., Pande, A., n.d. Sentiment Analysis Based on Dictionary Approach.

Ding, X., Liu, B., Yu, P.S., 2008. A holistic lexicon-based approach to opinion mining, in: Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, pp. 231–240.

Haruechaiyasak, C., Kongthon, A., Palingoon, P., Sangkeettrakarn, C., 2010. Constructing thai opinion mining resource: A case study on hotel reviews, in: 8th Workshop on Asian Language Resources. pp. 64–71.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 168–177.

Kim, S.-M., Hovy, E., 2004. Determining the sentiment of opinions, in: Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics, p. 1367.

Saensuk, M., Songram, P., Chomphuwiset, P., 2015. Feature-Based Opinion Mining on Smart Phone Reviews. The Institute of Industrial Application Engineers, pp. 246–250. doi:10.12792/icisip2015.047

Sharma, R., Nigam, S., Jain, R., 2014. Opinion Mining of Movie Reviews At Document Level. Int. J. Inf. Theory 3, 13–21. doi:10.5121/ijit.2014.3302

Sixto, J., Almeida, A., López-de-Ipiña, D., 2013. Analysing Customers Sentiments: An Approach to Opinion Mining and Classification of Online Hotel Reviews, in: Natural Language Processing and Information Systems. Springer, pp. 359–362.