Extraction and Summarization of Reviews using Lexicon based Approach

A A Sheikh1, T Arif1, M B Malik2, S I Bhat1

Department of Information Technology BGSB University Rajouri, J&K, India
Department of Computer Science BGSB University Rajouri, J&K, India
E-mail: sheikhaijaz@bgsbu.ac.in

Abstract. Online service providers have enabled users to freely comment in the form of review on an online items. These large scale of reviews contains valuable information and poses the need of automatic summarization. In this paper the extraction and summarization of reviews expressed by customers on Samsung mobiles has been carried out. The proposed approach extracts the Samsung reviews using custom build scraper from Flipkart and classify them as positive, negative and neutral using Lexicon based approach (SentiWordNet).

1. Introduction
Before the arrival of online shopping platforms we asked our friends, family members or experts to recommend about something that we wanted to buy such as books, watches, clothes, electronic products, etc. Nowadays, we easily get an idea about others opinions as a form of review or ratings which are available through their sites. Reviews have helped many users to increase their information about various items and selects the ones that suits their requirements. A recent studies shows that reviews about items on online shopping websites are the second reliable source of information after family or friends recommendations [1].

People are becoming comfortable with online shopping services in response they collectively generate huge amount of opinions as a comments or feedback about products [2]. In order to enhance their experience and satisfaction it has become a common practice for service providers to enable their customers to rate or express review on an item that they buy. These reviews signifies user likes and dislikes about a specific item in the form of text [3]. Therefore, a positive review on an item is like a guide that builds users positive mind for purchasing the same item. While as negative review divert his/her attention towards any other item. In addition, negative reviews also makes vendor aware about current faults of their item. An online user’s read reviews before they come to conclusion what to buy, but the huge number of reviews on an item creates a puzzling for them to read it all. In this perspective it’s important to have a system for collection, processing and capable of making it relevant for users. Sentiment analysis also called in various literatures as opinion mining [4], studies the identification and classification of reviews expressed in text through the use of different approaches based on machine or lexicon [5].

Opinion mining is the study of classification of reviews given by customers towards various online entities like items, services, industries, public, etc. the basic task of opinion mining algorithms is to categorize these reviews in positive or negative polarity. If the opinions discovered from reviews are with positive polarity, it simply indicates reviews are in favor of an item otherwise against. In this
paper an approach is proposed for both extraction and classification of reviews. The approach is divided into two modules data extraction for the collection of Samsung mobile reviews and lexicon based classifier for the classification of those extracted reviews. The data extractor is build using the CSS selectors in Python programming language for the extraction of reviews as well as metadata from website. The classifier for classification of reviews is based on Lexicon-based approach. The lexicon approach deals with textual content that generally occurs in reviews. This approach uses a corpus with positive and negative words to match the sentiment words of review to evaluate their orientation and degree.

The rest of the paper is organized as follows: Section 2 discusses the related work in the field of opinion mining using both supervised and unsupervised approaches. The architecture of the proposed system is given in Section 3. In Section 4 all the results of data extraction and opinion mining module are given followed by the conclusion and future work in Section 5.

2. Related Work

Research approaches in opinion mining are based on bag of words, lexicon based, machine learning or hybrid techniques. The Bag of words are using positive and negative words [4], while as lexicon techniques are using POS (Part of Speech Tagging) for identification and lexicon dictionary like SentiWordNet [5]. Most of the studies are built on machine learning are using NB (Naïve Bayesian) and SVM (Support Vector Machine) technique for classification of opinions. NB technique use likelihood perceptions bases on Bayes theorem [6]. While as SVM is a machine learning technique based on patterns in opinions [7]. There are also some approaches built on combination of multiple techniques, called as hybrid technique [8] [9] [10] Proposed system for reviews extraction and classification. The hotel reviews were extracted from TripAdvisor website with the help of web crawler based on “CSS selectors”. For sentiment classification of reviews a lexicon based dictionary SentiWordNet was used. The reviews were classified as positive negative and neutral. Another approach using SentiWordNet was proposed by [11] for classification of 500 news headlines. The set of national and global news headlines were used for the study. All the headlines were classified into positive and negative orientation with their sentiment score. To evaluate the accuracy of proposed approach comparison was made between proposed approach and manual labelled classification. The sentiment polarity and score of music lyrics were also computed by [12] with the help of SentiWordNet. The objective of study was a linguistic analysis of music lyrics, whether or not the particular song is appropriate for individual, only by classifying it into positive and negative polarity as per content within it.

Besides, usage of SentiWordNet several authors in their study have used Bag of Words for classification of reviews, paragraphs and documents. The Bag of Words is simply classifying reviews based of two dictionaries, a bag of positive and negative words like “good”, “bad”, “excellent”, etc. The review that contains maximum positive words are considered as positive review, while as review with maximum negative reviews will be negative. The score of these reviews will be calculated by simply calculating total number of positive minus negative reviews. [13] Proposed an approach to classify food reviews extracted from different sites. A bag of positive and negative words had been used to classify the reviews. Review that was having highest score was rated first and so on. One more approach based on the same technique was proposed by [14] here six features “Camera”, “Battery”, “Screen”, “Sounds”, “Design”, and “Hardware, Software performance” of three mobile devices “Samsung S6”, “Samsung A8” and “Samsung J7” were taken. The positive and negative words weight about features of the above captioned mobiles contained within the reviews were calculated. The review that contains highest number of sentiment words with score was ranked first and so on. The same approach was used by [15] to calculate the sentiment score and orientation of tweets. The tweets regarding “Samsung electronics devices were extracted from twitter. The study was done to analyze the tweets orientation, whether the given number of tweets are positive or negative. This bag of words shows no concept for negation handling and variance shifter, if the review contains negation words like “not bad” the approach will consider these reviews as a negative but, it’s totally opposite.
However, to overcome this issue, machine learning approaches based on several different techniques have been proposed in studies. [6] Proposed an approach to classify Bangla reviews. The Bangla corpus for opinion mining was not present so they use Amazon reviews and translated them into Bangla through the use of Google translator. Later on the reviews of both the datasets (English and Bangla reviews) were classified into weak, steady and strong orientation using well known machine learning technique Naïve Bayes. To check accuracy of proposed approach the comparison was made between orientations of both datasets, which results a same and good accuracy. Another machine learning technique Support Vector Machine was used by [16] for the study of opinion mining of Bangla reviews. The Amazon watches online reviews were translated into Bangla reviews. To identify the orientation of both English and Bangla reviews Support Vector Machine and Naïve Bayes techniques were used. The classification through SVM shows better results as compared to NB. Combination of two traditional approaches results into a new hybrid approach. A hybrid approach proposed by [17] the approach was built on the combination of both Support Vector Machine and Particle Swarm Optimization to classify movie reviews into positive and negative. The results showed that SVM with PSO gives better accuracy compared to SVM or PSO approach.

3. Architecture of the Proposed Approach
The approach performs the work in two main modules Data extraction from Ecommerce and Opinion mining using Lexicon based technique. Figure 1. Depicts an overview of the architecture of proposed approach.

![Figure 1: Architecture of the Proposed Approach](image)

3.1 Data Extraction
Scraper also termed as a “Spider” or “Bot” is a program designed in any programming language that automatically browse the webpages of website [18]. The scraper starts from seed URL and visits all other linked pages. It use “CSS Selectors” for identification of relevant information from those webpages. Scraper recursively transverse webpages identify relevant information and then using extraction function it extracts and save them. For identification, extraction, and collection of Samsung reviews from Flipkart website a Review Scraper was designed in Python with some useful extracting libraries. The Review Scraper use “Selenium” [19] library to visits the Flipkart website in automate manner. The information from visited webpages like ratings, reviews, User-ID, Item-ID, etc. are extracted using “BeautifulSoup” [20], and actually BeautifulSoup use the concept of “CSS Selectors” because all the elements have similar “CSS Selectors” for all the web pages. Another library “Pandas” [21] is used for the storage of extraction information in an organized manner. Figure 2 given below depicts algorithm for the working of data extraction.
3.2 Opinion Mining
The information obtained through data extraction module might contain non-relevant information. An elementary cleaning was achieved through NLP methods in python to remove all useless data like HTML tags, stopwords, etc. the most common stop words within reviews are “a”, “the”, “this”, etc. these hardly mention anything about sentiments. Because of their most frequency in the review, they must be draw out from the reviews. For example the review “the design is good” will be “design good” after stopwords removal. After cleaning the reviews they must be passed to the POS Tagger [12]. The part-of-speech tagger is used to tag all words of review with their correlate part-of-speech tag. Tagging of words is significant to recognize features on which users have expressed their feelings. Features are especially nouns while as feelings are adjectives. The POS tagger tags all the reviews which were given to it. All the reviews with their correlated tags are passed to the Lexicon dictionary for the computation of weight and polarity. Figure 3 given below depicts how orientation and sentiment score is calculated.

4. Results and Analysis
4.1 Data Extraction Results
Using a custom made scraper 37034 reviews were extracted from Flipkart websites. These reviews were posted by 24397 users towards 101 Samsung mobile devices. Each review contain following information :- Item name, Item ID, User name, User ID, Review, Rating, Review title, Upvote, Downvote and Date. Figure 4 (a) given below shows no of reviews
extracted from year 2014 to 2017 and figure 4 (b) depicts distribution of numerical ratings from 1 to 5 scale of extracted data.

![Figure 4: Extracted Reviews and Ratings](image)

### 4.2 Opinion Mining Results

Opinion mining also called as sentiment analysis [22] a technique determines interest of other person towards an item by mining knowledge from textual reviews. The main focus of opinion mining is to classify the polarity of reviews, whether the reviews are positive or negative towards an item with different intensity. To compute numerical score and polarity of the reviews a metric unit is required. SentiwordNet a lexicon database a part of natural language is employed. It employs the use if synsets of the Wordnet database [2] and assigns three sentiment score as objective, positive and negative score to each term of the sentence in a review. The review consists of different terms with different meaning, the lexicon dictionary takes each term along with its associated part of Speech Tag (POS) as an input. The reviews terms along with its corresponding part of speech tag were passed to lexicon set that generated a numeric score for each word. The sum-up of generated numerical score of all terms are in between -1 and 1, -1 indicates negative polarity while as 1 as positive.

To compute the intensity of numerical score and polarity of reviews, an unsupervised approach like Lexicon database was employed and implemented in python on all reviews of extracted dataset. After classification we found 31,758 number of reviews are positive, 4368 number of reviews are negative and 908 are neutral. Table 1 given below shows the results of reviews classification.

| Number of Reviews | Polarity |
|-------------------|----------|
| 31758             | Positive |
| 4368              | Negative |
| 908               | Neutral  |

To validate the opinion mining results, a comparison was made between score computed from reviews and numerical ratings. The comparison shows that our opinion mining produced a good accuracy results. Figure 5 given below depicts a difference between polarity of computed ratings and numerical ratings.
5. Conclusion
In this paper an approach is proposed for the extraction and summarization of opinions. Implementation of data extraction module produced a well ordered opinions (Reviews and Numerical Ratings) dataset. The opinion mining approach was applied on those reviews to produce the sentiment intensity and polarity of reviews. Our results shows that the proposed approach are promising and believe that online shopping recommendations based on those reviews will become increasingly significant as more and more people purchase and express their reviews on an online items. In the future, we will intent to improve opinion mining using a supervised approach and make an effort to build a predictive model.

6. References
[1] Chen L, Chen G and Wang F 2015 Recommender systems based on user reviews: the state of the art User Model. User-adapt. Interact. 25 99–154
[2] Kreutzer J and Witte N 2013 Opinion Mining Using SentiWordNet Lingfil 1–19
[3] Pang B and Lee L 2008 Opinion mining and sentiment analysis
[4] Liu B 2012 Sentiment Analysis and Opinion Mining vol 5
[5] Ohana B and Tierney B 2009 Sentiment classification of reviews using SentiWordNet Sch. Comput. 9th. IT T Conf. 13
[6] Hasan K M A, Sabuj M S and Afrin Z 2016 Opinion mining using Naïve Bayes 2015 IEEE Int. WIE Conf. Electr. Comput. Eng. WIECON-ECE 2015 511–4
[7] Keshav R, S A S, Goutham R and Naren J 2014 Content based Recommender System on Customer Reviews using Sentiment Classification Algorithms Int. J. Comput. Sci. Inf. Technol. 5 4782–7
[8] Basari A S H, Hussin B, Ananta I G P and Zeniarja J 2013 Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization Procedia Eng. 53 453–62
[9] Santhosh Kumar K L, Desai J and Majumdar J 2017 Opinion mining and sentiment analysis on online customer review 2016 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2016
[10] Tian X, Zhang L and Wei W 2017 The Design and Implementation of Automatic Grabbing Tool in Tripadvisor Proc. - 4th Int. Conf. Enterp. Syst. Adv. Enterp. Syst. ES 2016 76–80
[11] Agarwal A 2016 Opinion Mining of News Headlines using SentiWordNet 2016 Symp. Colossal Data Anal. Netw. 1–5
[12] Sharma V 2016 Sentiments Mining and Classification of Music Lyrics using SentiWordNet 1–6
[13] Anshuman, Rao S and Kakkar M 2017 A rating approach based on sentiment analysis Proc. 7th Int. Conf. Conflu. 2017 Cloud Comput. Data Sci. Eng. 557–62
[14] Rekha and Singh W 2017 Sentiment analysis of online mobile reviews Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2017 20–5
[15] Anwar Hridoy S A, Ekram M T, Islam M S, Ahmed F and Rahman R M 2015 Localized
twitter opinion mining using sentiment analysis *Decis. Anal.* **2** 8

[16] Sabuj M S, Afrin Z and Hasan K M A 2017 Opinion Mining Using Support Vector Machine with Web Based Diverse Data

[17] Samad A, Basari H, Hussin B, Pramudya I G and Zeniarja J 2013 Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization *Procedia Eng.* **53** 453–62

[18] Kim K S, Kim K Y, Lee K H, Kim T K and Cho W S 2012 Design and implementation of web crawler based on dynamic web collection cycle *Int. Conf. Inf. Netw.* 562–6

[19] Vykhur 2015 How to Scrape/Crawl Research Data Using Selenium WebDriver – and Java | Denvycom

[20] Beautiful Suop 2012 Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation

[21] Pandas 2011 pandas: powerful Python data analysis toolkit — pandas 0.24.0.dev0+1294.gca85a412 documentation

[22] Hu M and Liu B 2014 Mining and Summarizing Customer Reviews