STATISTICAL AND SENTIMENT ANALYSIS OF CONSUMER PRODUCT REVIEWS

Zeenia Singla[1], Sukhchandan Randhawa[2], and Sushma Jain[3]

[1,2,3]Department of Computer Science and Engineering, Thapar University, Patiala-147004

***Abstract*—Big Data commerce has given a big leap to e-commerce. It has opened up the avenues to smarter and informed decision making for large industries as well as the consumers. Online reviews on e-commerce giants like Amazon, Flipkart are one such paradigm which can be used to arrive at more profitable decisions. They are not only beneficial for the consumers but also for the product manufacturers. Online reviews have the potential to provide an insight to the buyers about the product like its quality, performance and recommendations; thereby providing a clear picture of the product to the future buyers. The usefulness of online reviews for manufacturers to realize customer requirements by analyzing helpful reviews is one such unrealized potential. Both positive and negative reviews play a big role in determining the customer requirements and extracting consumer‘s feedback about the product faster. Sentiment Analysis is a computational study to extract subjective information from the text. In this research, data analysis of a large set of online reviews for mobile phones is conducted. We have not only classified the text into positive and negative sentiment but have also included sentiments of anger, anticipation, disgust, fear, joy, sadness, surprise and trust. This delineated classification of reviews is helpful to evaluate the product holistically, enabling better-decision making for consumers.**

***Index Terms*—Data Analysis, Big Data, Text Mining, Text Classification, Sentiment Analysis, Online Reviews.**

I. INTRODUCTION

Due to the rapid growth of electronic commerce, online reviews have replaced the traditional “word-of-mouth” and have been playing a vital role in influencing the consumer's buying patterns and sales of a product. Reviews act as a trust-building platform for the consumers where by judging the previous buyers’ experience they are able to make informed decisions. From the manufacturer's point of view, helpful online reviews are crucial to mine customer requirements for improving a product or designing a new product. By capturing relevant online reviews, manufacturers can adhere to the customer requirements in the target market. Manufacturers also get an insight to the competitive market and ongoing trends influencing their marketing decisions as well.

Retail websites like Amazon.com offer different options to the reviewers for writing their reviews. For instance, the user can provide rating in the form of numerical stars (usually ranging from 1 to 5 stars) or open-ended customer-authored comments about the product. The presence of online reviews on a website is believed to increase the user credibility, attract consumer visits, augment hit ratio and time spent on the site. The discovery platforms like Zomato and Trivago are booming just on the basis of user reviews provided on restaurants and hotels. Reliable customer reviews build a trust factor among the novice users and help to enlarge the customer base. Both positive and negative reviews help the consumers and the manufacturers. Manufacturers can take negative feedback constructively and can know about the areas that they need to work upon to improve their product or service.

In this research, unstructured data is taken into context which will be filtered to remove noisy data and pre-processed to evaluate sentiment of the mobile phone reviews. The proposed work will help future buyers to make better decisions on the basis of analysis of feedback received by a particular smartphone brand. It will also allow manufacturers to meet consumer expectations better on the basis of feedback received.

II. RELATED WORK

Data Analytics has enabled users to unravel the hidden patterns in data. Big data provides insight on consumer behavior [1] which can be used to make informed decisions. An average consumer is generating both structured and unstructured data which is transforming market decision making. Big Data so generated is defined using three dimensions: Volume, Velocity and Variety. The volume and the relentless rapidity at which data is being generated every day is exceeding the computing capacity of many IT departments. Two more Vs that play an important role in explaining big data are: Veracity and Value. Veracity adds to the noise and abnormality in data that degrades the quality of data in question. By filtering the irrelevant data, remaining data can be utilized to provide valuable business insights. Big Data has enabled businesses to flourish and improvise on the basis of evidence rather than intuition. It aids in gaining insights on better targeted social influencer marketing, segmentation of customer base, recognition of sales and marketing opportunities, detection of fraud, quantification of risks, better planning and forecasting, understanding consumer behavior, etc. [2].

Sentiment analysis, also known as opinion mining, means identifying the sentiments of the users on the basis of positive, negative and neutral connotations. Opinion mining can be classified into three different levels: document level, sentence level and phrase level [3]. A lot of prior research has been done in this field where words and phrases have been classified with prior positive or negative polarity [4]. This prior classification is helpful in many cases but when contextual polarity comes into the picture, the meaning derived from positive or negative polarity can be entirely different. For example, the word ‘amazing’ has a prior positive polarity and the word ‘degrade’ has a prior negative polarity. However, they may be used with negation words like ‘not’ that change the context completely and sometimes phrases containing negation words intensify rather than changing the polarity. For instance, the product delivered was not only good but amazing in terms of looks. This contextual polarity of the phrases was taken into consideration in [5] and ambiguity was removed. [6] used a refined method to establish contextual polarity of phrases by using subjective detection that compressed reviews while still maintaining the intended polarity. Delineated study has been conducted on tweets available on Twitter, movie reviews to build the grounds on sentiment analysis and opinion mining. A sentiment classifier has been built to categorize positive, negative and neutral sentiments not only in English but also for other languages using corpus from Twitter [7]. [8] determined the polarity of smartphone product reviews only on the basis of positive and negative orientation of the review. [9] established a system using support vector machine where sentiment analysis is carried out by taking into consideration sarcasm, grammatical errors and spam detection.

Through Statistical and Sentiment Analysis of Consumer Product Reviews (SACP), we have calculated sentence level sentiment orientation of the reviews into ten different sentiments viz. anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive and negative; providing a much more detailed feedback to the consumer than the previous systems. The organization of paper is as follows: Section III explicates the dataset used and the approach followed to conduct the analysis. Section IV represents the statistical analysis carried out between different features to determine relationship between them. Section V demonstrates the sentiment of the text of the top three brands. Finally, section VI concludes the proposed work and describes its future scope.

III. Framework

The SACP framework will work in two different modules as shown in figure 1.

A. Data set and its Features

The first module includes data collection and preprocessing of data. A large sample of online reviews is collected from the e-commerce giant Amazon.com. The data set consists of over 400,000 reviews for approximately 4500 mobile phones. It includes six features as explained in table I.

TABLE I: Features included in the Data Set

Feature	Description
Product Name	Model name of the mobile phone.
Brand Name	Brand associated with the mobile phone.
Price	Price of the mobile set in dollars.
Rating	User rating between 1 to 5.
Reviews	User reviews provided for every mobile phone.
Review Votes	Number of people who found the review helpful.

B. Approach

The approach followed by the proposed SACP framework is described in Figure 1. Initially, the experimental data is collected from an e-commerce website Amazon.com. Each data set is in the Comma Separated Values (CSV) file format and available as supplement. In the second step, data is pre-processed to remove stop words, punctuation marks, whitespaces, digits and special symbols. ’tm’ package [10] is employed for text mining. In the third step, feature selection is done to extract relevant features from the data set. The given data set consists of only six features. All these features are relevant for making the required prediction, so feature selection is not required in the proposed work. In the fourth step, statistical analysis of the data set is conducted to examine correlation between different features. In the fifth step, sentiment analysis is carried out to estimate the sentiment of the text involving feelings of anger, anticipation, disgust, fear, joy, sadness, surprise, trust, along with polarity of the text: positive or negative.

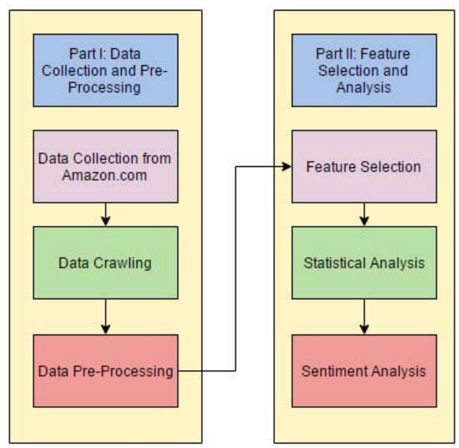

Fig. 1: Methodology Used

IV. Statistical Analysis of Data Set

Different features are compared and plotted to understand the relationship between different features. The different outcomes of running the statistical analysis are listed below.

A. Number of Review Counts by Brand

After analysis, the count of number of reviews is determined for each brand as represented in figure 2. The bar chart exhibits that Samsung, BLU and Apple are the top three brands having the highest number of reviews. Brands HTC and CNPGD have received the minimum reviews from the consumers. Thus, one can conclude that the brands

Samsung, BLU and Apple have the highest customer base among all the other brands.

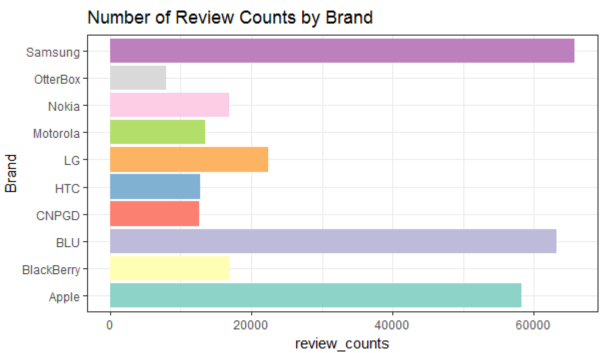

Fig. 2: Number of Review Counts by Brand

B. Rating Distribution by Brand

Rating distribution corresponding to each brand is found out to determine the brands receiving highest and lowest rating as depicted in Figure 3. User rating for each review is provided on a scale of 1 to 5 where 1 to 2 depict negative rating, 3 is neutral and 4 to 5 depict positive rating. The top three brands having highest rating are Samsung, Apple and BLU. The negative ratings of these three brands is almost the same. It can be concluded that Samsung is the most favoured brand and OtterBox is the least in terms of rating. CNPGD brand has almost equal rating of 1, 3 and 5 respectively showcasing that CNPGD brand is an average brand.

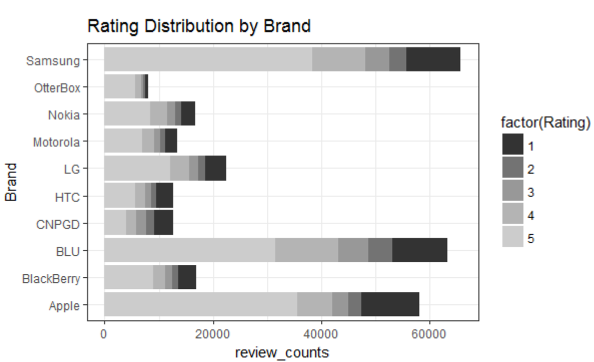

Fig. 3: Rating Distribution by Brand

C. Positive and Negative Reviews Distribution by Brand

Classification of reviews is done in terms of positive and negative reviews with respect to each brand. This analysis will influence the consumer buying patterns, as the consumer tends to choose the brand having maximum positive reviews. In figure 4, it can be observed that the brand Samsung has the highest share of positive reviews among all the brands. Brands BLU and Apple stand second and third in the stack. CNPGD brand has almost equal share of positive and negative reviews. Although, OtterBox brand has minimum number of reviews, yet its positive feedback is far greater than the negative feedback. This elucidates that even though mobile phones manufactured by OtterBox haven't been used by a large number of users, yet the consumers' experience has been fairly good with its handsets.

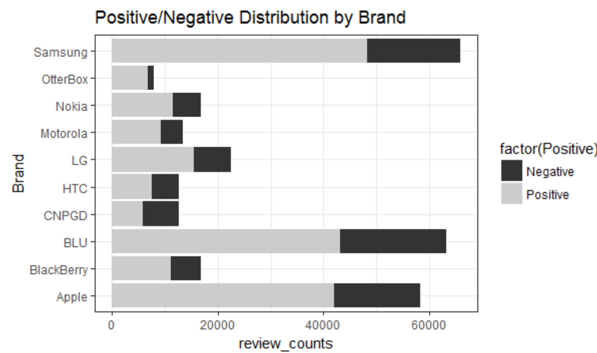

Fig. 4: Positive and Negative Reviews Distribution by Brand

D. Review Length

Review length is calculated to determine its relationship with other attributes. In this data set, there are over 400,000 reviews. Length of each review is calculated. Maximum length of a review is 29,624 characters. Mean length of all the reviews comes out to be 216.67 characters. The analysis of relationship between review length and dataset features like product rating and product price is presented below.

1) Review Length and Product Rating: Relationship between review length and product rating enabled us to determine whether detailed reviews affected product rating or not. Figure 5 elucidates that there is no correlation between the two.

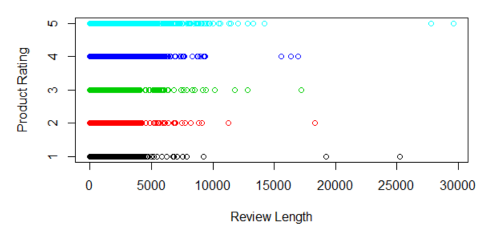

Fig. 5: Review Length and Product Rating

2) Review Length and Product Price: The figure 6 shows that with the increase in price, the length of the reviews does not increase. As the top models of different brands have high price, so one expects to have detailed reviews regarding these products. However, no such trend was observed.

E. Price and Rating

Higher the price of the product, higher the expectations of the consumers and better the quality. As seen in figure

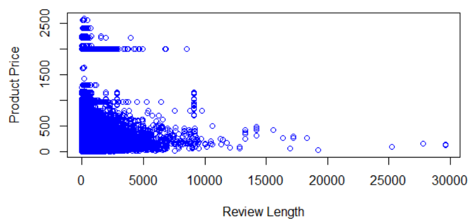

Fig. 6: Review Length and Product Price

Fig. 8: Word Cloud for Samsung Reviews

7, high priced products attract higher ratings illustrating higher satisfaction among buyers of expensive products.

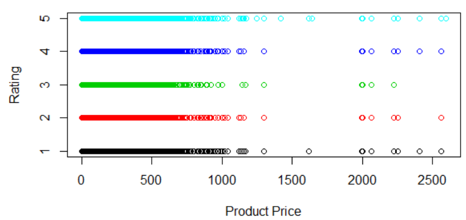

Fig. 7: Product Price and Rating

Fig. 9: Word Cloud for BLU Reviews

Fig. 10: Word Cloud for Apple Reviews

F. Word Cloud

The most frequent occurring words are found out in this analysis which can give both the consumer and the designer an idea of what the users are feeling about the product or what are the key aspects of the product. The words represented in the word cloud with a set frequency can aid in highlighting the most commonly cited words in the reviews. The height of the word represents its frequency. The word cloud of words having minimum frequency of 2000 is created by using 'SnowballC' [11] and 'WordCloud' [12] packages for each of the top three brands namely Samsung, BLU and Apple as shown below.

1) Word Cloud for Samsung Reviews: In the figure 8, highlighted words like 'great', 'good', 'quality', 'happy', 'love' provide positive feedback to the consumers.

2) Word Cloud for BLU Reviews: In the figure 9, highlighted words like 'great', 'good', 'excellent', 'easy' paint a positive picture before the consumer about the BLU brand.

3) Word Cloud for Apple Reviews: In the figure 10, highlighted words like 'great', 'good', 'works', 'perfect', 'recommend' are assuring the consumer that the products of apple are worth-buying.

V. Sentiment Analysis

Sentiment analysis has been conducted by employing an in built package named 'Syuzhet' [13]. which encompasses three sentiment dictionaries. NRC sentiment dictionary is used to extract eight different emotions and their corresponding valence in the text. Eight different emotions represented are: anger, anticipation, disgust, fear, joy, sadness, surprise and trust. Positive and negative valence is also calculated. In figure 11 it is seen that overall polarity of the text is positive with high valence of trust, joy and anticipation. Negative polarity is almost half the positive polarity as shown in figure 12 with average feelings of anger, disgust, fear and sadness.

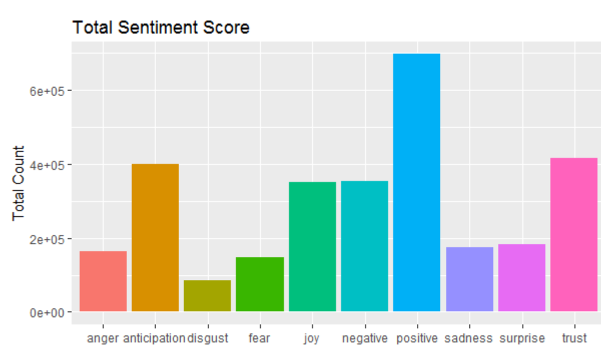

Fig. 11: Sentiment Analysis

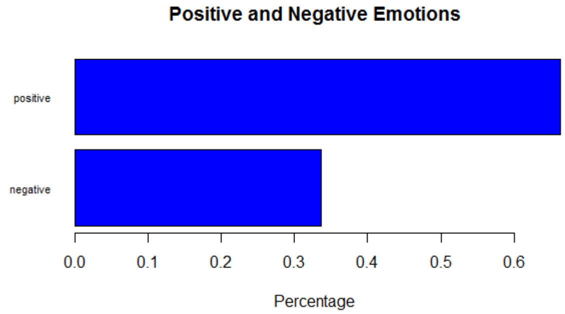

Fig. 12: Positive and Negative Emotions

Sentiment analysis of the top three brands, i.e. , Samsung, BLU and Apple is also conducted to understand and compare their respective review sets.

A. *Sentiment Analysis of Samsung Reviews*

According to Figure 13, feelings of anticipation, joy, trust and surprise are highly greater than the feelings of disgust, fear and sadness. As per figure 14 number of the positive reviews is far greater than the number of negative reviews. Thus, the overall sentiment of mobile phones manufactured by Samsung is positive.

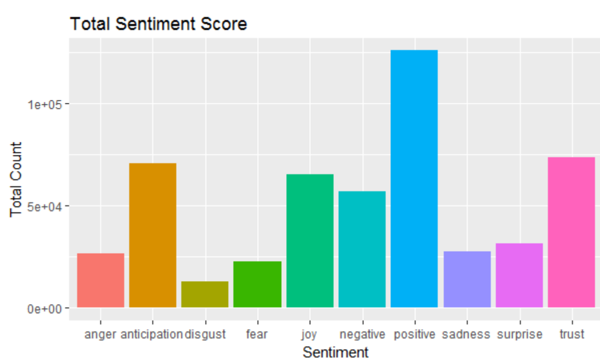

Fig. 13: Sentiment Analysis of Samsung Reviews

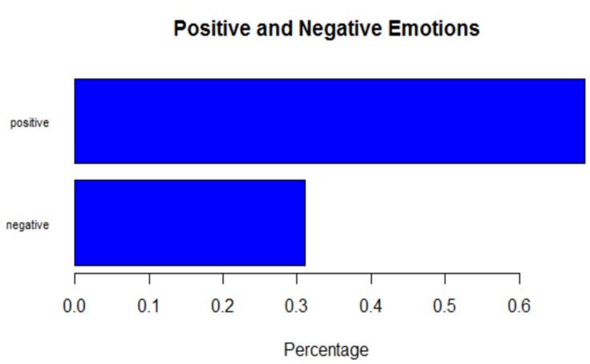

Fig. 14: Positive and Negative Emotions of Samsung Reviews

B. *Sentiment Analysis of BLU reviews*

The positive sentiments of joy, trust, anticipation and surprise associated with mobile phones manufactured by BLU are high than the negative sentiments of disgust, fear and sadness as shown in figure 15. Likewise, the overall polarity of the brand is positive as seen in figure 16.

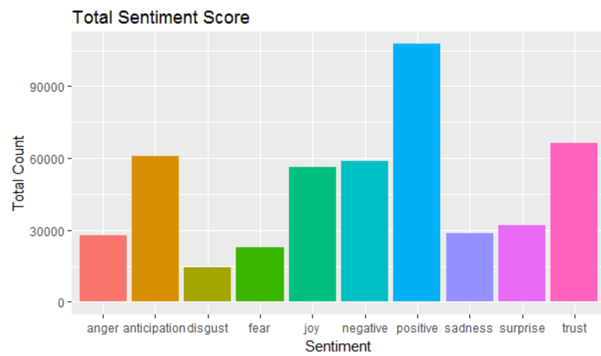

Fig. 15: Sentiment Analysis of BLU Reviews

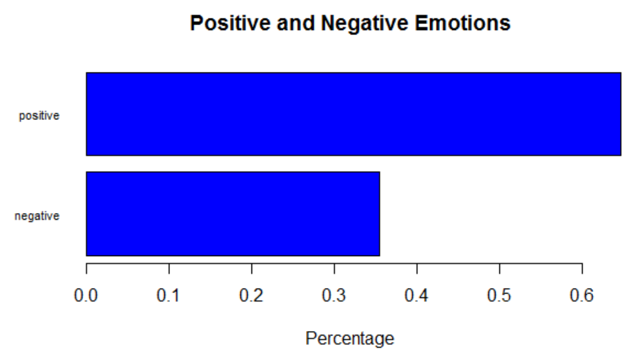

Fig. 16: Positive and Negative Emotions of BLU Reviews

C. *Sentiment Analysis of Apple Reviews*

In Figure 17, feelings of anticipation, joy, trust and surprise are far greater than the feelings of disgust, fear and sadness. In figure 18 the number of positive reviews is almost double the number of negative reviews. Thus, it can be said that the overall sentiment of mobile phones manufactured by Apple is positive.

D. *Cross Validation*

A sample of the labelled data is then trained and tested on Support Vector Machine(SVM) classifier, whose performance is validated using 10-fold cross validation. The performance of SVM is depicted in table II. The predictive accuracy of SVM comes out to be 84.87% which is quite good.

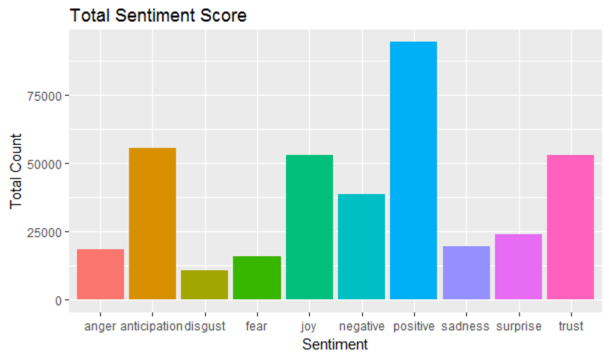

Fig. 17: Sentiment Analysis of Apple Reviews

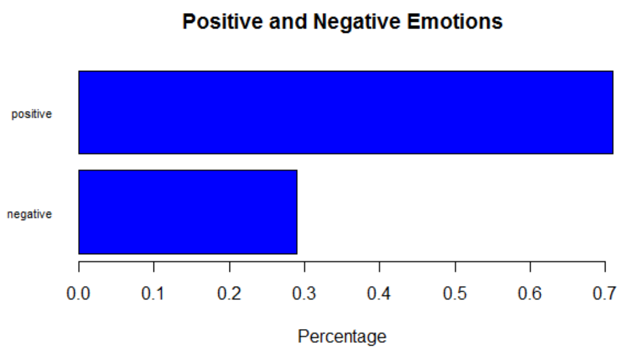

Fig. 18: Positive and Negative Emotions of Apple Reviews

TABLE II: Cross Validation for SVM

Runs	1	2	3	4	5
Accuracy	85.95	84.44	83.69	86.11	85.09
Runs	6	7	8	9	10
Accuracy	84.58	82.11	85.85	85.66	85.22

VI. Conclusion and Future Scope

Rapid transformation from the offline markets and ''word-of-mouth'' tradition to tell the consumers about the durability of a product for growth of e-commerce online markets has given birth to online reviews as the source of trust building community. Through this research, we intend to focus on understanding the relationship between different features and conduct sentiment analysis of the mobile reviews data set which is useful both from the consumer's point of view and the designer's end. Through statistical analysis, it can be concluded that Samsung, BLU and Apple are the three top ranked brands currently in the market. The most positive feedback is received for the brand Samsung. It was also observed that detailed reviews not necessarily attracted better rating and that higher priced products didn't fetch delineated reviews. However, better rating was observed in case of high priced products depicting high levels of customer satisfaction and better quality of the products than the low-priced products. The sentiment orientation of the top three brands Samsung, BLU and Apple was found out to be positive coupled with high positive sentiments of joy, trust, anticipation and surprise. The classification of data is very efficient as the accuracy of SVM after cross validation is equal to 84.87%.

This research can be extended to mine customer requirements keeping in mind the designer's concerns. So, helpfulness of reviews can be calculated to provide the designer with maximum useful information enabling him to improvise the product or roll a new product in the market by meeting maximum customer requirements. This extensibility in the research will be of great benefit to the industry and will make requirement gathering much less time-consuming, shrinking down the expenses incurred using surveys, questionnaires, interviews, market research and trends.

References

[1] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," *Journal of Business Research*, vol. 69, no. 2, pp. 897–904, 2016.
[2] P. Russom *et al.*, "Big data analytics," *TDWI best practices report, fourth quarter*, pp. 1–35, 2011.
[3] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.
[4] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997, pp. 174–181.
[5] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 347–354.
[6] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 271.
[7] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREc*, vol. 10, no. 2010, 2010.
[8] M. WAHYUDI and D. A. KRISTIYANTI, "Sentiment analysis of smartphone product review using support vector machine algorithm-based particle swarm optimization." *Journal of Theoretical & Applied Information Technology*, vol. 91, no. 1, 2016.
[9] D. N. Devi, C. K. Kumar, and S. Prasad, "A feature based approach for sentiment analysis by using support vector machine," in *Advanced Computing (IACC), 2016 IEEE 6th International Conference on*. IEEE, 2016, pp. 3–8.
[10] I. Feinerer, "Introduction to the tm package text mining in r," 2015.
[11] M. Bouchet-Valat and M. M. Bouchet-Valat, "Package âĂŸsnowballcâĂŹ," 2015.
[12] I. Fellows, M. I. Fellows, and L. Rcpp, "Package âĂŸwordcloudâĂŹ," *Retrieved*, vol. 4, p. 2013, 2012.
[13] M. Jockers, "Package âĂŸsyuzhetâĂŹ," 2016.