# A Study on Sentiment Analysis of Product Reviews for Consumer Insights

**Li Wei Zhang*1, Chen J. Liu2, Wang H. Bo3, Xiang R. Huang4, Dr. Mei L. Zhao5**
1234 Department of Computer Engineering, Tsinghua University, Beijing, China
5 Assistant Professor, Department of Computer Engineering, Tsinghua University, Beijing, China

## ABSTRACT

Due to the increase in demand for e-commerce with people preferring online purchasing of goods and products, there is a vast amount information being shared. The e-commerce websites are loaded with large volume of data. Also, social media helps a great deal in sharing of this information. This has greatly influenced consumer habits all over the world. Due to the vivid reviews provided by the customers, there is a feedback environment being developed for helping customers buy the right product and guiding companies to enhance the features of product suiting consumer's demand. The only disadvantage of availability of this huge volume of data is its diversity and its structural non-uniformness. The customer finds it difficult to precisely find the review for a particular feature of a product that s/he intends to buy. Also, there is a mixture of positive and negative reviews thereby making it difficult for customer to find a cogent response. Also these reviews suffer from spammed reviews from unauthenticated users. So to avoid this confusion and make this review system more transparent and user friendly we propose a technique to extract feature based opinion from a diverse pool of reviews and processing it further to segregate it with respect to the aspects of the product and further classifying it into positive and negative reviews using machine learning based approach.

**KEYWORDS**: aspect; sentiment analysis; feature extraction; machine learning

## INTRODUCTION

In the recent years E-Commerce has exploded everywhere in the world, and majority of the population is preferring to buy products through these websites. Consequently large amount of data in the form of reviews is produced which helps prospective buyers to choose the right product. Furthermore these reviews contain opinionated contents which can be useful for the company to identify the areas which need to be enhanced.
However it is impractical for the user to read each and every review about the product. Moreover, reading only few reviews may present a biased idea about the product. It is quite possible that some of the reviews lack credible sources, which the users have no means to differentiate. Besides the reviews and ratings provided do little to assess the specific features of the product. Due to all the above constraints, the user is unable to make a fully informed decision about the product.

Opinion mining also known as sentiment analysis can be used to extract customer reviews from different sources on the internet. This technique implements various algorithms to analyze the corpus of data and make sense out of it. This technique helps to identify the orientation of a sentence thereby recognising the element of positivity or negativity in it. Automated opinion mining can be implemented through a machine learning based approach. Opinion mining uses natural language processing to extract the subjective information from the data (in this case it's customer reviews).

Opinion mining techniques can be applied to wide range of data. It can track the popular viewpoint or attitude of

the general public towards a particular thing, person or an event. There are three general levels for opinion mining tasks: document level, sentence level and phrase level in Liu[1]. Document level tasks mainly help in segregating the overall document into either subjective document or objective document. Further it can be distinguished into positive, negative or neutral. It can also help separate the spam from the non spam. The sentence level opinion mining is performed on the sentences which can help group certain sentences to summarise the opinion and also it can help identify comparative sentences to rank them accordingly. Phrase level deals with the aspects and is known as aspect based opinion mining. This helps to identify the reviewers sentiment about specific aspects of the product. This level does the finer-grained analysis of the opinions.

## RELATED WORKS

Product review sentiment analysis, also called as opinion mining, is a method of ascertaining the customers' sentiment about a product on the basis of their reviews. Liu [1**]** classifies the opinion mining tasks into three levels: document level, sentence level and phrase level.

Opinion can be represented as an entity consisting of five parameters: target tentity, aspect in opinion, opinion holder, time when opinion is expressed, and the sentiment orientation of the opinion holder of a feature entity at a particular time. [2] makes use of frequency itemset mining which by employing a certain minimum support count finds the itemsets. Further it makes use of Naive Bayesian algorithm for aspect and sentence orientations by using supervised term counting.

The opinionated reviews also contain other information that can be used to ascertain the sentiment about a product. Venkata Rajeev P et al [3] uses the reviews from flipkart.com and proposes the combination of four parameters: star ratings of the product, the polarity of the review, age of review and helpfulness score, for determining the opinion of a product.

The task of mining the features is of particular importance and many methods are suggested for it. Weishu Hu et al. [4] divides the opinion analysis tasks into three steps: identifying the opinion sentences and their polarity, mining the features that are commented upon by customers, and removing incorrect features.

The primary focus of product review system is identifying the adjective word in a sentence and identifying the sentiment behind it. Yan Luo et al. [5] suggests the final sentiment score of the review to be the cumulative sentiment score of all the adjectives in that review

D V Nagarjuna Devi et al. **[6]** proposes a system that uses a supervised classification approach called as support vector machine. This paper claims that the proposed classifier approach gives out the best result. It also identifies various challenges in sentiment analysis like sarcasm and conditional sentences, grammatical errors, spam detection and anaphora resolution. sentence level classification is done on input data which is further classified according to the subjectivity/objectivity.Further aspect extraction is done using SentiWordNet. This is then further fed to SVM classifier to find the overall opinion.

Shoiab Ahmed et al. [7] proposes that the count of scored opinion words be classified into seven possible categories i.e. strong-positive, positive, weak-positive, neutral, weak-negative, negative, strong-negative. Sentiment analysis is then done with the help of these score counts.

## PROBLEM DEFINITION

An application that collects reviews from the users about a certain product and analyzes them. It would

segregate the reviews into positive and negative reviews. The negative reviews will be helpful to the companies to further enhance their product based on the user's' feedback. The application further provides the pros and cons of the individual feature of the product. The application will further provide reports about the sentiment analysis performed on the products. We further aim to create a recommendation system that recommends products to users according to the feature requirement of user.

## CONSTRAINTS:

Every system has some shortcomings or in other words all the system round the world work under some predefined constraints. Our system basically has 4 constraint under which it works.

1. **Sarcasm:**
   It is an extremely difficult task for a machine to perceive a sarcastic review about any product and understand exact meaning of the review. It is even sometimes difficult for humans to interpret some of these sarcastic comments.
2. **Errors in Grammar:**
   Due to social media apps people often commit grammatical mistake, punctuation errors and spelling mistake. In most of the cases in order to express their feelings about the product people deliberately type wrong spellings.This makes it difficult for the machine to figure out the exact meaning behind the review of a customer.
3. **Detecting spam:**
   In this competitive world some users' usually try to post the negative reviews to spoil others' reputation.So it becomes extremely difficult rather impossible to segregate negative reviews from spam. Most of the time such reviews are considered as negative reviews by the machine.

4. **Anaphora Resolution**
   In our product review system we focus on nouns that occurs in a sentence and see whether that noun is a feature of that product or not using our feature database. But many times customer uses pronoun instead of the proper feature name to express his feeling about the product but in such case due to the absence of the feature name in that product the system fails to recognize it as an opinion sentence about a feature.

## TECHNOLOGIES

1. Python:
   Python is a widely used dynamic programming language with a clean syntax and an indentation structure easy to learn [8].
2. Stanford CoreNLP:
   Stanford CoreNLP is natural language processing tool. It has a comprehensive toolkit with a good range of grammar checking tools. It is fast and reliable. It identifies the part of speech of the words in a sentence. It is flexible and extensible [9].
3. BeautifulSoup:
   Beautiful Soup is a third party Python library from Crummy designed for scraping.
4. Scikit-learn:
   Scikit-learn is machine learning library in Python which provides implementations of various machine learning algorithms including classification, regression, clustering, etc.

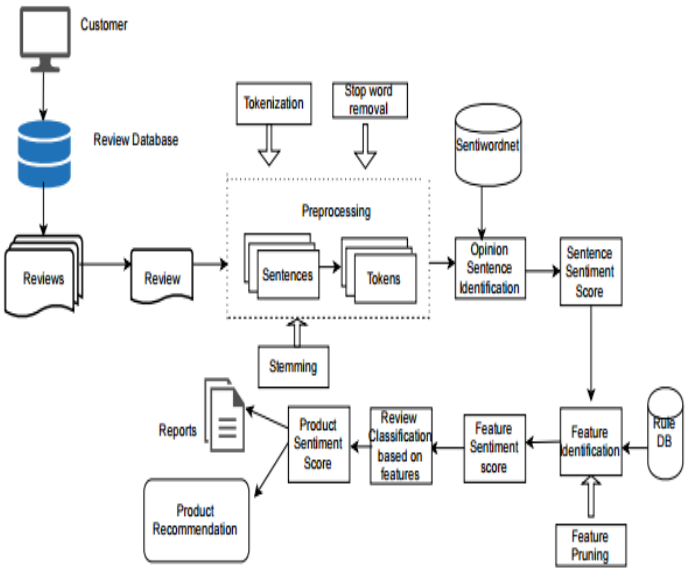
## PROPOSED METHODOLOGY

The dataset used for this project is the Amazon Reviews Database [10]. The reviews in the dataset are consists of the attributes such as: Reviewer ID, Product ID, Review Text, Rating and time of the review.

The main source of data used is the product reviews from Amazon. The reviews for a few popular phones have been obtained by building a web crawler. The web crawler has been written in Python using a scraping library called BeautifulSoup. Along with the review text, some additional data related to the reviews such as reviewer name, review date, overall rating and comments were also obtained. The crawler is called periodically to get the

most up-to-date reviews. Each review is generally treated as a sentence or a group of sentences. They are cleaned and stored in a CSV file.

The first stage of analysis involves preprocessing of the reviews. Preprocessing involves the following operations: stemming, stopword removal and part-of-speech tagging.



Then, sentiment analysis is performed on the preprocessed reviews and overall sentiment score for each review is generated. Further for feature extraction, there are two cases:

1. Single Feature: If the review contains only a single feature, then the sentiment score of the review is assigned to the feature.
2. Multiple Features: Some reviews have multiple features contained in them. So the above procedure will not work in this case. Rules are defined to extract multiple features and assign the correct sentiment score to those features.

For reviews containing more than one sentence, we first check if the review contains a word from an adjective word set or not. If it does not contain one, then it is assumed to be objective. If it does contain an adjective, the feature that corresponds to that adjective is found by looking for a set of predetermined nouns near that adjective. We plan on using the implementation of the K-nearest-Neighbours (KNN) algorithm for this step, using CoreNLP or SentiWordNet libraries.

Sentiment scores range from 1 to 5, 1 being the most negative, 5 being the most positive and 3 being neutral. These scores are then averaged for each feature and phone and stored in a database.

The recommendation engine takes in a set of user preferences in terms of the features that the user would like in the product. It presents the most suitable product to the user based on the scores assigned to each phone in the previous step.

## CONCLUSION

The system proposed in this report aims to help a user select a phone based on his/her needs using review data from previous users. It pulls review data from the Amazon website periodically, processes it and assigns a score to each feature of each phone based on the review data. When the user inputs his/her preferences, the scores are used to determine the best match for the user. This match is guaranteed to be up-to-date. Additionally, depending on time constraints, we also plan on introducing a portal for phone manufacturers to analyze the public perception of their product over time and identify key areas where their product can be improved.

## REFERENCES

[1] Bing Liu (2012), 'Sentiment Analysis and Opinion Mining', Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers
[2] A.Jeyapriya, C.S.Kanimozhi Selvi, "Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm", 2nd International Conference On Electronics and Communication Systems(ICECS '2015)
[3] Venkata Rajeev P, Smrithi Rekha V, "Recommending Products to Customers using Opinion Mining of Online Product Reviews and Features", 2015 International Conference on Circuit, Power and Computing Technologies [ICCPCT]
[4] Weishu Hu, Zhiguo Gong, Jingzhi Guo, "Mining Product Features from Online Reviews", IEEE International Conference on E-Business Engineering
[5] Yan Luo,Wei Huang, "Product Review Information Extraction Based on Adjective Opinion Words", 2011 Fourth International Joint Conference on Computational Sciences and Optimization
[6] D V Nagarjuna Devi, Chinta Kishore Kumar,Siriki Prasad, "A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine", 2016 IEEE 6th International Conference on Advanced Computing
[7] Shoiab Ahmed, Ajit Danti, "A Novel Approach for Sentiment Analysis and Opinion Mining based on SentiWordNet using Web Data", 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15)
[8] Python (programming language) - https://en.wikipedia.org/wiki/Python_(programming_language)
[9] "Stanford Core NLP Toolkit" Available: http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf
[10] Inferring networks of substitutable and complementary products J. McAuley, R. Pandey, J. Leskovec Knowledge Discovery and Data Mining, 2015