

Sentiment Analysis and Opinion Summarization of Product Feedback

Sindhu C and Vadivu G

Abstract: With the exponential growth of online shopping platforms, user interaction is made direct through their reviews and ratings. User's opinions and experiences are a significant source of valuable information in decision making process. In recent days, almost every website encourages users to express and exchange their views, suggestions and opinions related to product, services, policies, etc. publicly. Opinion mining is an extensive branch of Artificial Intelligence and a form of Natural Language Processing which illustrates the attitude of the customers, in specific services or products. Also known as Sentiment Analysis, it aims at determining the response and mood or attitude of the speaker or the overall contextual and emotional polarity or reaction. Existing algorithms determine sentiment by training on datasets, lexicon-based approach by calculating polarity and rule-based approach for classification. Opinion Summarization is the process of consolidating a large amount of sentiments and opinions into a clear and brief statement for an easier grasp on the underlying context. Major summarization methods include, Extractive method, Sentence Ranking, Abstractive method and Clustering of Textual Segments. Hence it is important to judge and classify these reviews and present a laconic opinion so it would be easier for users to obtain a gist and overall polarity on the various reviews instead of going through all of them.

Index Terms: Feature Extraction, Opinion Mining, Opinion Summarization, Review Based Opinion Mining, Sentiment Analysis.

I. INTRODUCTION

Online shopping websites have given users, a chance to share their views about a product they have purchased or used and users to have taken it as benefiting and are expressing openly. Users have varied views and opinions about a product and over that there are different ways to express those opinions which makes it a huge chunk of opinions that may be similar yet varying in words. Thus, it is a much necessary task to clean these opinion records and to publish a minimized, concise output while combining all the reviews. Sentiment analysis or opinion mining are the domains which deals with the extraction of exact sentiments of reviewers from their review to produce a brief straight forward review.

Sentiment analysis is the process of analysis of opinions and emotions from textual data [1]. It helps in understanding the attitudes, opinions and emotions through text. It also helps

Revised Manuscript Received on July 05, 2019.

Sindhu C, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India.

Vadivu G, Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, India.

predicting the latent information present in the text. Also, there are various important notions to be discussed under the topic of

Sentiment Analysis, such as Subjectivity/Objectivity, Polarity, Level of Sentiment and these can be defined as the righteous

steps to meet the endpoint of getting a perfect output over finding the exact sentiment of the user. The reviews can be classified as subjective or objective based on its content such as the one with a sentiment would be subjective while the one which is a fact would be classified as objective.

For example: I went to the Nike Store. There I purchased the Nike 2000X, which is the best amongst running shoes.

Here the first statement is objective and the second statement is subjective.

This example clears the idea of subjectivity, the ones consisting of opinions and objectivity, the factual statements. The Polarity of the review gives a classification on the sentiment to be specific as to positive negative or neutral. The level of Sentiment can be recognized as the value of how deep an assessment was made say document level, sentence level or phrase level but for the purpose of review classification we need the second (Sentence Level). Sentiment classification is a task of text classification with the objective of classifying a text according to the sentimental polarities of opinions it contains e.g., good or bad, positive or negative [1]. Various techniques have been introduced to extract the exact level of sentiments from the word blocks. Some of which are sentiment classification using Naive Bayes, Maximum Entropy and Support Vector Machine.

It is well known that as the days pass by, the data on the Internet increasing exponentially and so the data related to products such as reviews. Thus, it has become a necessity to provide a summary to them so that the user does not have to go through the complex task of processing all the facts, opinions and ratings and has to read all the reviews as given by the other people and also to interpret and to compare them. Thus, here opinion summarization comes into picture. Opinion Summarization is the process of analyzing the text and to produce a concise laconic output which is brief and depicts the exact meaning of the complete text analyzed. This could result in much shorter or maybe up to a word summary of all the text analyzed by it so as to provide the gist of the overall text or reviews providing the users with a brief easy to process without complexity valued word that can be trusted and believed upon. The main data sources for the reviews are websites such as YELP or micro-blogging or Social Networking or e-commerce

sites for example Amazon, Flipkart etc.

The paper is organized in a manner as follows. Section II contains research related to Sentiment Analysis and Opinion Summarization and their Techniques. Section III provides the workflow that can be used as a solution methodology. Section IV discusses about the topic in hand. Section V concludes the overall survey. The scope for future work is also highlighted.

Online shopping websites have given users, a chance to share their views about a product they have purchased or used and users to have taken it as benefiting and are expressing openly. Users have varied views and opinions about a product and over that there are different ways to express those opinions which makes it a huge chunk of opinions that may be similar yet varying in words. Thus, it is a much necessary task to clean these opinion records and to publish a minimized, concise output while combining all the reviews. Sentiment analysis or opinion mining are the domains which deals with the extraction of exact sentiments of reviewers from their review to produce a brief straight forward review.

Sentiment analysis is the process of analysis of opinions and emotions from textual data [1]. It helps in understanding the attitudes, opinions and emotions through text. It also helps predicting the latent information present in the text. Also, there are various important notions to be discussed under the topic of Sentiment Analysis, such as Subjectivity/Objectivity, Polarity, Level of Sentiment and these can be defined as the righteous steps to meet the endpoint of getting a perfect output over finding the exact sentiment of the user. The reviews can be classified as subjective or objective based on its content such as the one with a sentiment would be subjective while the one which is a fact would be classified as objective.

For example: I went to the Nike Store. There I purchased the Nike 2000X, which is the best amongst running shoes.

Here the first statement is objective and the second statement is subjective.

This example clears the idea of subjectivity, the ones consisting of opinions and objectivity, the factual statements. The Polarity of the review gives a classification on the sentiment to be specific as to positive negative or neutral. The level of Sentiment can be recognized as the value of how deep an assessment was made say document level, sentence level or phrase level but for the purpose of review classification we need the second (Sentence Level). Sentiment classification is a task of text classification with the objective of classifying a text according to the sentimental polarities of opinions it contains e.g., good or bad, positive or negative [1]. Various techniques have been introduced to extract the exact level of sentiments from the word blocks. Some of which are sentiment classification using Naive Bayes, Maximum Entropy and Support Vector Machine.

It is well known that as the days pass by, the data on the Internet increasing exponentially and so the data related to products such as reviews. Thus, it has become a necessity to provide a summary to them so that the user does not have to go through the complex task

of processing all the facts, opinions and ratings and has to read all the reviews as given by the other people and also to interpret and to compare them. Thus, here opinion summarization comes into picture. Opinion Summarization is the process of analyzing the text and to produce a concise laconic output which is brief and depicts the exact meaning of the complete text analyzed. This could result in much shorter or maybe up to a word summary of all the text analyzed by it so as to provide the gist of the overall text or reviews providing the users with a brief easy to process without complexity valued word that can be trusted and believed upon. The main data sources for the reviews are websites such as YELP or micro-blogging or Social Networking or e-commerce sites for example Amazon, Flipkart etc.

The paper is organized in a manner as follows. Section II contains research related to Sentiment Analysis and Opinion Summarization and their Techniques. Section III provides the workflow that can be used as a solution methodology. Section IV discusses about the topic in hand. Section V concludes the overall survey. The scope for future work is also highlighted.

Paper	Topic	Related	Methodologies/Tools used	Outcome
		Papers		
[1]	Applications, Challenges,	[3] [5] [6]	Sentiment Polarity, Subjectivity	Exhaustive Survey
	Classification and	[10] [11]	Detection, Term presence vs	on all approaches
	Extraction methods		Frequency Single Document	and techniques and
	Approaches,		Opinion Oriented summarization,	even Challenges
	Summarization,		Negation, Domain Adaptation,	
	Availability of Datasets,		Syntax, Language Model and	
	Implications		Unsupervised Lexicon Adaption.	
[2]	Sentiment Classification	[1] [3]	Naive Bayes,	Survey
	using Machine Learning		Maximum Entropy,	
	Techniques		Support Vector Machines	
[3]	Machine Learning	[1] [5] [6]	Lexicon based,	Survey
	Approaches		Statistical,	
			Supervised and Unsupervised	

[4]	Sentiment Classification	[2] [5]	Support Vector Machine,	SVM is most
			Maximum Entropy,	viable for
			Naïve Bayes Classification	classification
[5]	Feature Extraction,	[1] [6] [7]	Latent Semantic Analysis (LSA),	LSA is easier and
	Sentiment Classification		SVM	faster than most.
[6]	Opinion Classification	[1] [2] [3]	Naïve Bayes Text Classification,	Naive Bayes
			Logistic Regression,	Classification was
			SentiWordNet	most efficient
[7]	Aspect Term Extraction and Analysis	[4] [7]	Natural Language Processor,	Various tools are
			SentiWordNet,	justified for the
			Amazon web services,	different steps in
				Term Extraction
[8]	Product Aspect Ranking	[4] [6]	Probabilistic Aspect Ranking -	Important Aspects
			Frequency-based,	of products are
			Correlation-based,	classified and
			Hybrid method	ranked
[9]	Semantic analysis based on Adjectives and Conjunctions	[4] [7]	Sentiment Classification, Cluster	Sets of adjectives
			labeling, Sentence Polarity.	along with
				conjunctions with
				polarity.
[10]	Challenges and Applications	[1] [4]	Sentiment Classification, Spam	Survey
			Detection, Natural Language	
			Processing, Opinion Mining	
			Software	
[11]	Product Recommendation and Performing sentiment analysis on product reviews	[1] [7]	OpenNLP tool, HTML, CSS,	Almost accurately
			JQuery, JSON, Feature Polarity	recommended
				(88.33%

				accuracy).
		Table 1: Reference Paper Summary		

II. RELATED WORK

[1] is an exhaustive research on the opinion mining domain and deals with the applications, challenges, classification and extraction methods, approaches, summarization, availability of datasets, implications and even how and in what fields opinion mining could be beneficial. [2], is basically a survey paper for [1] and is full of details about sentiment classification and polarity check and on the use of SVM, Naive Bayes and Maximum Entropy and throws light on machine learning methods to accomplish Sentiment classification. In [3], points at the various approaches and methodologies for Sentiment Classification. These methods include 1. Dictionary based, searching for synonyms and antonyms and adding them to a seed list. 2. Subjective Lexicon, which classifies each word into positive, negative and neutral categories. 3. Corpus Based, where we deal with the complete context of the text and classify it wholly as positive or negative.

At [4], the paper has compared the well-known machine learning approaches including Naïve Bayes Classification, Maximum Entropy and Support Vector Machines (SVM), with the conclusion that SVM outperforms the other two at categorization and that it can work well even with smaller training data. At [5], proposes an approach namely Latent Semantic Analysis (LSA) for product feature identification, which is a method to inspect links between the terms contained in a document, thus extracts the feature terms from a then formed concurrency matrix. At [6], the paper compares approaches for opinion classification using Naive Bayes Classification, Logistic Regression and making use of SentiWordNet, coming to the conclusion that Naive Bayes was more efficient among the three at classification.

The [7], paper outlines a step-by-step procedure for sentiment analysis and summarization of a customer review with the steps being, 1. Review Collection 2. Sentence segmentation and POS tagging 3. Frequent features identification 4. Feature pruning 5. Opinion words extraction 6. Opinion orientation identification 7. Opinion sentence orientation identification 8. Opinion summarization. [8] showcases a Probabilistic Aspect Ranking Algorithm for identifying the various product aspects from consumer generated reviews. The algorithm works on the following characteristics - A. consumers' opinions on these aspects greatly influence their overall opinions on the product. Then it aims at ranking the various aspects of the product to estimate the Pros and Cons of the product. [9] has utilized adjectives and conjunctions such as "fair and legitimate" or "simplistic but well-received" to refine and separate similarly connoted

and oppositely connoted words. A summary of all the related works has been added in Table 1.

III. WORKFLOW OF OPINION SUMMARIZATION IN SENTIMENT ANALYSIS

The complete workflow of a review sentiment analysis and summarization has been provided as a flowchart in figure 1.

A. Crawl Review Data

Reviews can be extracted using a simple regex and a text crawler or DOM manipulator. Otherwise review datasets from YELP and Amazon can be used for testing. Also, Handy extractor, import.io, Helium Scraper, parse hub, etc. [11] are some of the tools for review extraction. Most researchers scrapped the datasets from the public domain websites and e-commerce sites such as Flipkart, Amazon, Baidu by using web technologies like HTML, JavaScript, JQuery, Selenium etc. through DOM manipulations [11] and accessed all the required dataset.

B. Review Pre-Processing

Review preprocessing includes *Tokenization*, *Stop Word Removal* (Stop words are the short function words such as ‘where’, ‘why’, ‘is’, ‘the’, ‘that’, etc., although there is a universal dictionary for these words. Removal of stop words helps increase performance. Natural Language Toolkit in python has a dictionary for stop words in 16 languages), *Stemming* (it is the process of reducing the words to their root form so that the various form of the same word used in the text, traces back to the same word [7], For example- words like industrial, industrialized, industrialization, etc., will all be reduced to “industry” thereby reducing the size of the dictionary dataset, the major algorithms include Suffix-Stripping, Production Technique. Lemmatization is a more complex process for stemming which groups similar occurring words together so they be analyzed as a single unit but it also takes into account the context and meaning of all the words into account). Preprocessing is an important step that helps normalize the review dataset hence facilitates easier comparison, categorization and removing ambiguity.

C. Word Tokenization

Tokenization [7] is the process of breaking up a large text into individual words called tokens, also referred to as Segmentation. Token normalization is the process of checking tokens to match even though superficial differences in the character sequence of the tokens may occur. Consider the following example where there are three-word tokens.

I have friends.

D. Aspect Extraction

It is the process of extracting words that define the aspects or specifications of an object in the opinionated text [7]. For example - “*I love the camera but the battery life is short*” - include two aspects, namely camera and the life of battery which are polarized by the words, “love” and “short” respectively.

Extracting aspects from the text helps in identifying key features that need addressing so that the users get sentimental analysis based on them and also, they are of utmost importance in text summarization.

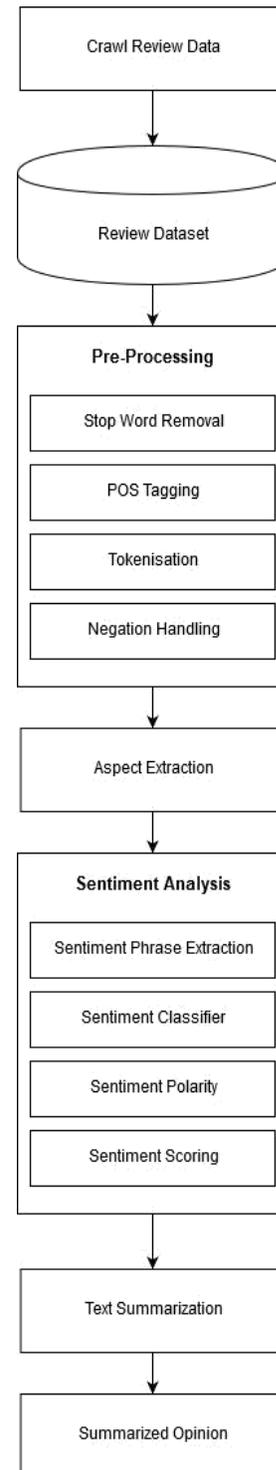


Figure. 1. General Workflow for Review based Opinion Summarization.

E. Sentiment Analysis Phase

Analysis on text includes the various processes of Sentiment Phrase Extraction, Sentiment Classifier and Scoring. The input text is removed of stop words, then using any of the widely used classifiers like Naive Bayes Classification, Maximum Entropy, Support Vector Machines the words are classified according to their polarities, i.e., positive, negative and neutral. These words are then scored between 0 to 1, according to their correlation, impact and context.

Sentiment Phrase Extraction: Every text written isn't necessarily be providing any real sentiment. Thus, to reduce workload and time consumption effective texts are extracted directly from the reviews by the means of noun phrase extraction and related word matching.

Sentiment Classifier: [4] classifies the polarity of the words and clauses as positive or negative or neutral and is accomplished by SVM or Naive Bayes or Decision Tree or Maximum Entropy Algorithm.

Sentiment Scoring: the final sentiments are scored based on the emphasis and the specifications relativity with the product and is marked from 0 to 1. [2] has discussed the use of SVM light package for training and testing which proved to be higher performance yielding.

Classification can also be performed using based on relationship information as used by the researchers of Cornell University [1][2]. Adding up language models can be beneficial too for language-based approach to review mining.

On application of the machine learning techniques for sentiment analysis researchers found out that SVM proved to be the best out of the three with a higher score [1][2].

F. Text Summarization

It is the process of extraction and crunching of the review text by some the following techniques such as Word frequency check along with heading wise summary generation and ATS (Automatic Text Summarization) [13]. This helps in reducing users time consumption and faster checkout based on the crisp summary of all the reviews.

G. Summarized Opinion

The final output comes as a brief summary of the text taken as input. This along with the polarity and sentiment scores gives a proper and a deep level of insight of the product reviewed in just a few words. This produces a fine laconic output which is crisp and clear and helpful in saving user's time.

IV. DISCUSSION

Various tools and techniques have been tried and made use of by many researchers, out of which some really stand out against all the others. As for the case process of review extraction from websites no such proper methodology was used but could be downloaded directly from the websites such as YELP or other which provide it in public domain and also, could be extracted using JavaScript by making some DOM manipulations. As for Sentiment analysis SVM proved to be the best choice of most researches and with higher specific results along with feature identification [2] [12]. Also, other approaches such as the Maximum Entropy Classifier and Naive

Bayes were the second choice though the best results came out with SVM these approaches proved to be good for other than review classification. Text Summarization is another bigger task at hand which could be accomplished using NLTK (Natural Language Tool Kit), this toolkit was a choice of the most. So, all these domains can be combined together to produce an effective review sentiment analysis and scoring system which could be beneficial for users as well as a benchmark for all kinds of products. Thus, it can be said that since SVM, NLTK, review datasets and Machine Learning Concepts which proved to provide better outputs at their jobs can be used to produce a better review scoring system.

V. CONCLUSION

As it has been well established that customer reviews play a crucial role in swaying the mood of public be it in the matter of choosing a product, opting to watch a movie and similar matters it is important to use the resources at hand to analyze and summarize these reviews to provide a concise output and recommendation to the users so as to help them make better decision. After surveying for the various methodologies for performing Sentiment Analysis and Opinion Summarization of a customer review, it can be concluded that so far Support Vector Machine would be the appropriate choice for the categorization [3], while any of the approaches mentioned above for Lexicon classification are equally viable [2]. Also, Latent Semantic Approach [4] seems most reasonable for feature identification.

REFERENCES

- [1] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis" *Foundations and Trends in Information Retrieval*, Vol. 2, No 1-2, pp. 1-135, 2008.
- [2] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan "Thumbs up? Sentiment Classification using Machine Learning Techniques", 2008.
- [3] Harpreet Kaur, Venu Mangat and Nidhi, "A Survey of Sentiment Analysis techniques", *International Conference on IoT in Social, Mobile, Analytics and Cloud (I-SMAC-2017)*, 2017.
- [4] Jayashri Khairnar and Mayura Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification", *International Journal of Scientific and Research Publications*, Volume 3, Issue 6, June 2013.
- [5] Chien-Liang Liu, Wen-Hoar Hsiao, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou, "Movie Rating and Review Summarization in Mobile Environment", *International Journal of Scientific and Research Publications*, Volume 3, Issue 6, June 2013.
- [6] Santhosh Kumar K L, Jayanti Desai and Jharna Majumdar, "Opinion Mining and Sentiment Analysis on Online Customer Review", *IEEE International Conference on Computational Intelligence and Computing Research*, 2016.
- [7] Kajal Sarawgi and Vandana Pathak, "Opinion Mining: Aspect Level Sentiment Analysis using SentiWordNet and Amazon Web Services", *International Journal of Computer Applications (0975 - 8887)*, Volume 158, No 6, January 2017.

[8] Zheng-Jun Zha, Jianxing Yu, Jinhui Tang, Meng Wang and Tat-Seng Chua, "Product Aspect Ranking and Its Applications", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 5, 2014.

[9] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proc. 8th Conf. Eur. Chap. Assoc. Computer Linguist., Morristown, NJ: Assoc. Computer Linguist.*, pp. 174–181, 1997.

[10] Haseena Rahmath P., "Opinion Mining and Sentiment Analysis- Challenges and Applications", *International Journal of Application or Innovation in Engineering & Management*, 2014

[11] Vamsee Krishna Kiran M, Vinodhini R E, Archanaa R and Vimalkumar K., "User specific product recommendation and rating system by performing sentiment analysis on product reviews", 2017.

[12] Moraes Rodrigo, Valiati Joã o Francisco, Gaviã o Neto Wilson P, "Document-level Sentiment Classification: An Empirical Comparison between SVM and ANN", *Expert Systems with Applications*, 40(2): 621–633, 2013.

[13] P. Krishnaveni and Dr. S. R. Balasundaram, "Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence", *Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC)*, 2017.

AUTHORS PROFILE



Sindhu C is working as Assistant Professor in the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur for about 6 years. She holds Master of Engineering degree from Velammal Engineering College, Chennai and Master of Business Management from SRM University. She is pursuing PhD on Sentiment Analysis in SRM Institute of Science and

Technology. She has published 9 journal papers and 18 conference papers. She is a member of IET and ISCA.



Vadivu G is working in SRM Institute of Science and Technology as Professor and Head in the Department of Information Technology. She completed her PhD and M.Tech on Semantic Web and Computer Science and Engineering respectively from SRM Institute of Science and Technology. She has published 27 journal

papers and few conference papers.