



e-ISSN: 2319-8753 | p-ISSN: 2320-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 1, January 2021

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.512

9940 572 462

6381 907 438

ijirset@gmail.com

www.ijirset.com



Sentiment Classification for Product Reviews using NLP and Machine Learning

Ashmaki Suresh Waydande¹, Amol S. Dange²

PG Student, Dept. of Computer Engineering, Annasaheb Dange College of Engineering and Technology,
Astha MH, India¹

Assistant Professor, Dept. of Computer Engineering, Annasaheb Dange College of Engineering and Technology,
Astha MH, India²

ABSTRACT: With the growth of the Internet the number of online reviews and recommendations is increasing. Both users and organizations use this data for their needs. Users check the reviews before purchasing any item so that they can compare between two or more items. Organizations use these reviews to understand the issues and positive points about their product and hence can make decision accordingly. Consumer reviews contain rich and valuable knowledge for both firms and users. However, the reviews are often disorganized, leading to difficulties in information navigation and knowledge acquisition. In this report we have proposed a product aspect ranking framework, which identifies the important aspects of products, aiming at improving the usability of the numerous reviews. In particular, given the consumer reviews of a product, we first identify product aspects and determine consumer opinions on these aspects via a sentiment classifier. We then develop a aspect ranking algorithm to infer the importance of aspects by simultaneously considering aspect frequency and the influence of consumer opinions given to each aspect over their overall opinions. We then weight these aspects and then decide the overall rating of the product. We propose a framework for sentiment sentence compression. Different from the previous sentence compression model for common news sentences, it seeks to remove the sentiment unnecessary information for sentiment analysis, thereby compressing a complicated sentiment sentence into one that is shorter and easier to parse. We apply a preprocessing technique, with certain special features, to automatically compress sentiment sentences. We are also handling the Negation sentiments and neutral counts for better accuracy. We will not only obtain the ranking of products but also the ranking of the individual aspects of the products.

KEYWORDS: Product review mining, sentiment classification, aspect mining, natural language processing, machine learning, Bag of Words.

I. INTRODUCTION

Recent years have witnessed the rapidly expanding e-commerce. Millions of products from various merchants have been offered online. For example, Bing Shopping has indexed more than five million products. Amazon.com archives a total of more than 36 million products. Shopper.com records more than five million products from over 3,000 merchants. Most retail Websites encourages consumers to write reviews to express their opinions on various aspects of the products. Here, an aspect, also called feature, refers to a component or an attribute of a certain product. A sample review the battery of Nokia N95 is amazing.” reveals positive opinion on the aspect battery” of product Nokia N95. Besides the retail Websites, many forum Websites also provide a platform for consumers to post reviews on millions of products.

Textual information can be broadly classified into two main types: facts and opinions. Facts are objective expressions about entities, events and their properties. They are something which already happened or actually the case itself. (e.g., iPhone is an Apple product). Opinions are subjective expressions that describe peoples judgment, viewpoint, feelings towards entities, events and their properties. (e.g., I like this Apple iPhone 6). A decade ago, when an individual needed to make a decision, Consumer typically asked for opinions from friends, neighbors and families. Similarly, when an organization wanted to find the opinions about its products and services, it conducted opinion polls, surveys, and focus groups. In the last few years, volumes of opinionated text have grown rapidly and are also publically available. Social media plays a major role by allowing people to share and express their opinion on products, events, topics, individuals, and organizations in the form of comments, reviews, blogs, tweets, status updates, etc. instantly. Therefore, its quite obvious that people always prefer to hear others opinion before making a decision. Some people express their opinions

in binary scale (i.e. Positive or Negative) and some other expresses their opinions explicitly in terms of ratings (i.e. one to three or five stars).

Motivated by the above observations, we in this paper propose a product aspect ranking framework to first identify the important aspects of products from online consumer reviews. Synonym clustering is done to remove duplicate aspects. We have to develop a system with machine learning as well NLP based approach which provides better accuracy. The reviews will be classified as a positive or negative sentiment for that aspect via a sentiment classifier. After all the reviews have been classified then we will find the weight for each of these aspects. After this we calculate the overall weight of the product. We have to also reduce the neutral count of users view, so it will reduce the system false negative ratio. We also focus on negation handling, which is to improve the correct review from end users.

II. LITERATURE SURVEY

There are various machine learning approaches in which mostly used are the supervised classification methods. Machine learning techniques like Naïve Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM) are widely used. In machine learning technique the training dataset is collected first and classifier is trained on the training data and the feature selection is done on the basis of the term present and the term frequency, once the classification technique is used. Zhaet. al. [1] categorized the positive and negative opinions on the aspects from the pros and cons of the reviews. Here the classifier is trained by the Pros and cons reviews, which is used to determine customer opinion on the aspects in free text reviews. They compared the three supervised classifiers SVM, NB, ME in terms of F1- Measure, which is a combination of precision and recall. Here the SVM was implemented using the libSVM, NB with Laplace smoothing and ME was implemented with LBFGS parameter estimation. The experiment are conducted on the 21-products in eight domains found that the SVM performs better than NB and ME. D.K.Kirangeet. al. [2] conducted experiments on the laptop and the restaurant reviews dataset and compared the performance of the system with the KNN(K- Nearest Neighbor) and SVM, and compared their accuracy for sentiment classification, and concluded that the SVM performs better than KNN. M.GOVINDARAJAN [3] found that there are the disadvantages for the SVM classifier as its performance reduces for the small dataset. So proposed a new approach which is the assemble of classifier NB and SVM known as Hybrid approach, which is applied for the sentiment classification tasks, with the aim of efficiently integrating the advantages of the NB and SVM. The proposed approach is based on 5 phases and the experiments were conducted on the 10*10 fold cross validation for evaluating the accuracy on the dataset of 2000 movie reviews. And compared the results with the NB and SVM and concluded that the proposed approach NBSVM gives higher accuracy than the base classifiers (NB and SVM). Nguyen et. al. [4] introduced a study combining the advantages of the both NB and SVM classifier into a two stage system by applying the reject option for the document-level sentiment classification. Here the NB classifier was used in the first stage and the documents rejected by the NB where given as input to the second stage classifier SVM and the classification is done. The experiments were conduct on publicly available standard polarity dataset of movie reviews and evaluate the classifier based on 10 fold cross validation. Results were obtained without reject option and with reject option. SVM gains higher accuracy without reject option and NBSVM gains higher accuracy with reject option. Wang and manning [5] showed that the inclusion of the word bigram features gives consistent gain on sentiment analysis. They presented a simple model where SVM is built over NB log-count ratios as feature values. The experiments were carried on the different publicly available dataset on different methods and found that the NBSVM is the robust performer than other classifiers and the benefits of the bigrams depend on the dataset used. Kim et. al. [7] proposed a simple but novel approach to adjust the term frequency portion in tf-idf, which is an unsupervised weighting scheme, by assigning credibility adjusted score to each tokens. The comparison is done against traditional tf-idf weighting scheme on multiple benchmarks and the method proposed gives better results on multiple benchmarks, which included both the snippets and longer documents.

D. Tang et. al. [8] proposed a method, in which the problem of polarity inconsistency in the pipelined methods was overcome. The method proposed is joint segmentation and classification framework for sentiment analysis, which simultaneously conducts the sentence segmentation and sentence level sentiment classification. The experiments results were conducted on the Twitter dataset in SemEval 2013 and shows that the joint model outperforms the pipelined methods. WouterBanckenet. al. [9] proposed and domain independent approach for aspect based sentiment analysis to obtain most positive and negative aspects of a specific product from the free text customer reviews. The aspects are identified by the handcraft dependency paths in individual sentence. This approach does not require the seed word or domain specific knowledge as it employs offthe-shelf sentiment lexicon. The experiments were conducted on the on-line movie reviews and the MP3 player reviews and the accuracy of the top n aspects were calculated

III. RESEARCH METHODOLOGY

We propose a product aspect ranking framework to identify important aspects of products from numerous consumer reviews. We develop a probabilistic aspect ranking algorithm to infer the importance of various aspects by simultaneously exploiting aspect frequency and the influence of consumers opinions given to each aspect over their overall opinions on the product.

Modules

1. Preprocessing Module

2. Product Aspect Identification

3. Sentiment Classification

4. Aspect Ranking

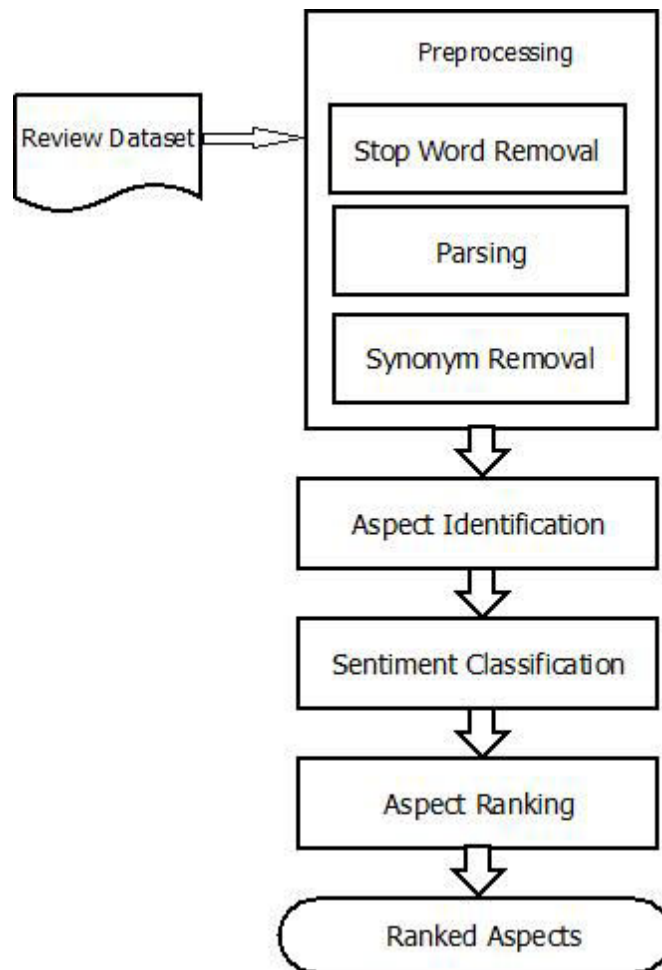


Figure 1: Flow of proposed research methodology

A. The preprocessing module consists of following sections,

Tokenization and StopwordRemoval :Tokenizing (splitting a string into its desired constituent parts) is fundamental to all NLP tasks. In lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases,

symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Stop words are words which are filtered out before or after processing of natural language data.

Stemming:

Stemming is the term used in information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form a generally written word form. Stemming programs are commonly referred to as stemming algorithms or stemmers. A simple stemmer looks up the inflected form in a lookup table. The advantages of this approach is that it is simple and fast.

B. Product aspects Identification:

Generally, a product may have several aspects. For example, iPhone has aspects, such as usability, design, application, 3G network. Identifying important product aspects will improve the usability of numerous reviews and is beneficial to both consumers and firms. Consumers can conveniently make wise purchasing decision by paying more attention to the important aspects, while firms can focus on improving the quality of these aspects and thus enhance product reputation effectively. For the Pros and Cons reviews, we first identify the aspects. Because consumer uses different words for same aspect. So this will reduce the accuracy of ranking algorithm. So here we use synonym clustering to obtain unique aspect. We collect synonym terms of aspect as a feature. The ISODATA (Iterative Self-Organizing Data Analysis technique) clustering algorithm is employed for synonym clustering.

C. Sentiment Classification

After the identification of important aspects the next step is sentiment classification. In this step the sentiments expressed on each aspect is identified. The sentiment is classified as a positive or a negative sentiment for that particular aspect.

D. Aspect Ranking

After the identification of important aspects the next step is sentiment classification. In this step the sentiments expressed on each aspect is identified. TFIDF, short for term frequency inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents.

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization.

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones.

Supervised binary Classification

Step 1: Divide items into k nonempty subgroups

Step 2: Compute seed point to the centroid of the cluster in current division. The centroid is the center which is middle point of cluster group.

Step 3: each object assign to the cluster nearest seed point.

Step 4: Repeat step 2 until data != Null.



TF/IDF

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones.

Step1: $TF(t) = (\text{Number of times term } t \text{ appears in a users comment}) / (\text{Total number of terms in the comments}).$

Step 2: $IDF(t) = \log_e(\text{Total number of comments} / \text{Number of comments with term } t \text{ in it}).$

Step 3: $tf-idf_{t,d} = tf_{t,d} \times idf_t.$

Step 4: Calculate Score

$$\text{Score}(q, d) = \sum_{t \in q} tf-idf_{t,d}.$$

Step 5: Finish

Negation Handling

For this approach we use **Keyword spotting**, it is classifies text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored.

Step 1: Read each C comment from D

Step 2: Tokenize each keyword using word splitting approach

Step 3: Call to train dataset which having positive and negative and synonyms.

Step 4: verify each word to each train DB word.

Step 5: Classify data separation as positive, negative as well as separation.

Step 6: store all classify data in new dataset D1.

Preprocessing Algorithm

Step 1: Read each C comment from dataset D1.

Step 2: Apply Sentence detection algorithm

Step 3: Apply stopword removal approach on C

Read all words from stop word Db

Step 4: Apply porter stemmer algorithm

Step 5: Return Text []



Sum Rule Algorithm

It was existing approach for analyze a user reviews for different kind of products. Some system support like positive and comments or some rating base. If the admin wants to identify the comment then he used sum rule approach with given procedure.

Step 1: Set the overall rating scenario for product like *R*.

Step 2: apply sum $f(x)$

Step 3: Calculate rating with all summation.

Step 4: Assign and display overall eating

IV. RESULTS AND DISCUSSIONS

The execution of various systems utilized for sentiment calculating so as to mine is assessed different measurements like accuracy, review and F-measure. Accuracy is the division of recovered cases that are applicable, while review is the division of pertinent occurrences that are recovered. The two measures are once in a while utilized together in the F1 score (likewise F-score or F-measure) is a measure of a test's exactness. A diagram of the work done in the undertaking of Sentiment Analysis is appeared in Table 1. This table speaks to an example of work done and a few works distributed on the theme of Sentiment Analysis.

Approach	Feature Selection	Data Source	Performance Accuracy
Lexical Resource	POS Apriori	Amazon Reviews	87.06%
Lexical Approach	Graph Distance Measurement	Blog Posts	82.95%
Hybrid	n-gram	Movie Review	91.00%
Naïve Bayes	Information Gain	Canteen Review	92.50%
Naïve Bayes and SVM	Base on minimum cuts	Movie Review	86.00
Proposed Approach	NLP and ML	Product base Review	95.00% (Estimated)

Table 1: Performance Analysis of Proposed System

The difficulties of users audit mining lie in that verifiable data is constantly blended with genuine audit information also, humorous words are utilized as a part of composing motion picture surveys. Item survey area extensively varies from motion picture survey area due to two reasons. Firstly, there are highlight particular remarks in item surveys. Individuals might like a few elements and despise a few others. Therefore audits comprise of both positive and negative suppositions, which make the errand of grouping the audit as positive alternately negative harder. Such element particular remarks happen less often in film audits. Besides, there are a ton of relative sentences in item audits and individuals discuss different items in surveys.

V. CONCLUSIONS

A product aspect ranking System will be used to identify the important aspects of products from numerous consumer reviews. We have seen the modules used for getting the ranked specs as output. They include Preprocessing Module, Product Aspect Identification, Sentiment Classification, Aspect Ranking. The preprocessing module contains the submodules for tokenisation, Stop word removal, Stemming. Synonym handling is done for aspects with same meaning but different name to avoid ambiguity. We have also considered the Negation handling for obtaining better accuracy. We will also consider the polarity and intensity of the reviews to minimise the errors in classification. This framework will allow the user not only to judge between the different products but also the user can judge the individual aspects of the products.



REFERENCES

- [1] Z-J-Zha, J-Yu, J-tang, M-wang, T-S Chua Product aspect ranking and its applications “, IEEE Transaction of Knowledge and Data Engineering, Vol.26, no.5, may-2014.
- [2] Dai Quoc Nguyen and DatQuoc Nguyen and Son Bao Pham,” A Two-Stage Classifier for Sentiment Analysis, International Joint Conference on Natural Language Processing”, pages 897–901, Nagoya, Japan, 14-18 October 2013.
- [3] Yoon Kim, Owen Zhang,”Credibility Adjusted Term Frequency: A Supervised Term Weighting Scheme for Sentiment Analysis and Text Classification”, June 2014.
- [4] D. K. Kirange, Ratnadeep R. Deshmukh,” Emotion Classification of Restaurant and Laptop Review Dataset: SemEval 2014 Task 4 “, International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 6, March 2015.
- [5] M.GOVINDARAJAN,“SENTIMENT CLASSIFICATION OF MOVIE REVIEWS USING HYBRID METHOD”, International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume- 1, Issue-3, Jan.-2014.
- [6] G.Vinodhini et al. “Performance of Hybrid Sentiment Classification Model on Online Product Reviews”. International Journal of Computer Science
- [7] Andr esCassinelli, Chih-Wei Chen,” CS 224N Final Project Boost up! Sentiment Categorization with Machine Learning Techniques”, June 5, 2009.
- [8] Duyu Tang, Furu Wei, Bing Qin , Li Dong , Ting Liu , Ming Zhou , “A Joint Segmentation and Classification Framework for Sentiment Analysis”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 477–487, October 25-29, 2014, Doha, Qatar. c 2014 Association for Computational Linguistics.
- [9] Sida Wang and Christopher D.Manning, “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification”, Department of Computer Science, Stanford University, Stanford, CA 94305
- [10] WouterBancken, Daniele Alfarone and Jesse Davis, “Automatically Detecting and Rating Product Aspects from Textual Customer Reviews”, Proceedings of DMNLP, Workshop at ECML/PKDD, Nancy, France, 2014.



INNO  **SPACE**
SJIF Scientific Journal Impact Factor

**Impact Factor:
7.512**

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 **9940 572 462**  **6381 907 438**  **ijirset@gmail.com**



www.ijirset.com

Scan to save the contact details