Product Review Summarization With Feature Extraction and Opinion Mining

Prof. Alka Londhe, Yash Deshpande, Gaurav Ghongde, Pratik Deshmukh, Hrushikesh Gate
(Department of Computer Engineering, PCCOE, Pune)
Email: alka.londhe.pccoe@gmail.com, yashd299792.458@gmail.com, 7gaurav.ghongde@gmail.com, pratik.desh.2015@gmail.com, hrushigate7@gmail.com)

Abstract
The boom of e-commerce sites has resulted in the shift of shopping from traditional retailers to online platforms. This has resulted in a host of products being sold on e-commerce platforms and with them, there is an increasing number of reviews, feedbacks and suggestion being posted on those sites. These reviews contain customer opinions which are of high value in an e-commerce environment.

We propose a review summarization tool which enhances existing text summarization techniques notably feature extraction and opinion mining. We try to accomplish an accurate model for opinion mining by taking in account domain-specific vocabulary and sentiment analysis of opinion to exact an unambiguous opinion of the feature. After performing opinion mining we use TextRank to extract sentences relevant to the features to extract sentences for a review summarization.

Such a review summarizing tool would obviate the need to manually go through each and every review and instead provide a pithy aggregation which would provide comprehensive summary covering varied facets of the product. Such an aggregation would help in the understanding of traits of listed products which would, in turn, help the seller to identify shortcomings or strengths of the product as well gauge the opinion of the customers on various features of the product. Whereas, the customers can use this aggregation to make an informed decision while making a purchase.

Keywords: Feature Extraction; Natural Language Processing; Opinion Mining; TextRank; Text Summarization

1. Introduction

The exponential growth in the e-commerce sector has led to the increase in the number of products that are being sold; as a foreseeable consequence, the number of reviews of the product has also increased. These reviews help other potential customers to make a more informed choice and also allow customers to make their appreciation or grievances known. Although reading through the customer reviews gives a comprehensible picture, it might be very time-consuming in some cases where the product has thousands of reviews listed. Since these reviews may run into thousands it is not humanly possible to apprehend each and every review. To solve this problem, there is a need for summarization of all these reviews into one to provide a holistic gist of all reviews. In this project, the design of a unified opinion mining and sentiment analysis framework is presented with the natural language processing approach. Text summarization is the process that is used to shorten a body of text in an intelligent and efficient manner to provide a brief gist of the text which encompasses all the important points in the text. In review summarization for products, instead of simply shortening a vast amount of text it identifies the relevant features in all the reviews which have the highest correlation and relevance to the product. The aim of review summarization is to allow the customer and the seller to identify the feedback to various features and aspects of the product. In accordance with the features and their feedback a summary is generated which reflects the entire review corpus.

Review summarization follows the steps of natural language processing (NLP) for the processing of the text. These steps are as follows:

1.1. Gathering of data:
Gathering of data is the initial stage which involves mining the text from sources such as datasets or through web crawling sites like Yelp, Amazon, Tripadvisor.

1.2. Tokenization:
The excavation of text results in assemblage of sentences, but this not useful to us since we apply cleaning and extraction algorithms to words not to entire paragraphs. Tokenization splits the sentences into individuals words. Taking a sample review for example, “Perfect packaging and laptop. It runs fine and exactly how described.” After tokenization this would turn into: ['Perfect', 'packaging', 'and', 'laptop.', 'It', 'runs', 'fine', 'and', 'exactly', 'how', 'described']

1.3. Removal of stop words:
Removing stop words is one of the essential preprocessing stages where frequently occurring words such as pronouns, articles and conjunctions are removed. ['Perfect', 'packaging', 'laptop', 'runs', 'fine', 'exactly', 'described']

1.4. POS tagging:
POS tagging is used to classify words into parts of speech such as noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection.

1.5. Feature Extraction:
Feature extraction is used to retrieve features from the word corpus. There are various algorithms to achieve feature extraction. One of them is TF-IDF. Term Frequency Inverse Document frequency uses all tokens in the word corpus to extract frequent vocabulary. Frequency of tokens from word corpus in each document consists of the term...
frequency and number of documents in which token occurs determines the Inverse document frequency.[7] If the token occurs frequently in the word corpus the TF would be high on the other hand, if the token occurs frequently in majority of documents then it reduces the IDF. [6]

\[ w_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right) \]  

In addition to the general preprocessing steps we also prioritize reviews based on their relevance to the product since there are go astray from the topic and thus making them less relevant to the product. TextRank can be used for chronological ranking of text, TextRank is graph based ranking model which is used in natural processing application and uses unsupervised learning for keyword and sentence extraction. Using TextRank and feature extraction we aim to generate summaries which contain the features as well as their attributes to give a holistic summarization of all reviews.

2. Related work

Our work is similar to Minqing Hu and Bing Liu propose a feature and opinion mining system for customer review summarization. They have adopted a 3 step process:

1. Mining of features from customer reviews using data mining and natural processing techniques.

2. Extraction of opinion sentence and determination of the opinion on the features, and calculation of the sentiment of those features whether they are positive or negative. Opinion orientation is done in two steps. First, the adjectives describing features are identified using natural language processing (NLP). Then, for each such word, the polarity is calculated to determine whether the word has a positive or negative connotation. Ultimately, opinion orientation of the sentence is determined.

3. The results of the above two stages are summarized. We use this work as the foundation and eliminated shortcoming in their procedure by providing our own approaches.

Hence, we extrapolate their findings and build upon them to generate our own model of review summarization.

3. Proposed system

Our proposed system is an enhancement of earlier works to make a text summarization tool which is highly accurate as well as more helpful for the end users. Objectives of the system:

- To build an algorithm for summarization of customer reviews.
- To extract reviews, clean them, perform analysis on them, extract important features pertinent to the product.
- To identify quality of features and aggregate them.
- To provide a feature based summary of the product.

4. Implementation

We propose a system to summarize text for products by extracting their features in accordance with the category that they belong to. Firstly, we extract the reviews through either crawling through the web or via a dataset of reviews. These reviews consist of sentences, and these sentences may be either simple sentences or compound sentences. We break these sentences into words in order to perform analyses on them with the help of word-tokenization. After the sentences are split into words, the words are then classified into various parts of speech such as nouns, adjectives, prepositions, verbs and adverbs which would help in extracting review features as well as in understanding their role in the sentence. [9]

![Fig. 1: Figure showing an example of WordNet.](image-url)
To achieve maximal accuracy we train a machine learning model which would go through various product listings and classify words in accordance with the product category. After we have trained our machine to recognize domain-specific vocabulary, we extract features which are nouns by using TF-IDF and other logical regression techniques. Next, we determine whether users have favorable or unfavorable opinions regarding these features. For calculating favorability or unfavorability, we search back those features in sentences they belong to and determine it using sentiment analysis. [5]

We extract the feature as well as the adjective associated with it using apriori algorithm. Apriori algorithm would mine frequent subsets and these subsets would help to extract features as well the adjectives along with it. There is a drawback to this approach when we extract adjectives related to a feature, it is possible that the adjective expresses similar opinion to varied magnitude. For example, “the phone works great” and “the phone is splendid”. Here both “great” and “splendid” are used to give positive connotation to the feature, but our earlier approach would fail since it considers “splendid” and “great” as different and unrelated entities. [3]

To solve this problem we calculate orientation of the adjective for which we use wordnet to find synonyms and hyponyms of the adjective and by the sentiment analysis we calculate the degree of correlation of these synonyms and replace all of the occurrences of the adjectives and its synonyms with the most frequent one that is the mode of such adjectives. [4]

Furthermore, it is not necessary that the feature important to us is frequent in the word corpus, in order to solve this problem we have to identify infrequent features. To identify infrequent features we use the method suggested by Minqing Hu and Bing Liu where for every sentence in the corpus, if a sentence has no frequent feature it has opinion words, the nearest noun of the opinion word is subsequently stored in the feature set as an infrequent feature. The nearest noun that opinion word modifies is used because it is the most common occurrence in natural language. [8]

We use the nearest noun/noun phrase as the noun/noun phrase that the opinion word modifies because that is what happens most of the time. We use this simple heuristic method to find the nearest noun/noun phrase.

After the completion of feature extraction and opinion mining on the corpus, we proceed to the last step, which is, summary generation. The set of features which we extracted in the last step are searched in the corpus to extract phrases containing them. There might be multiple sentences having the same feature which poses the problem of which sentences to use for summary generation. To solve this problem, we use TextRank. TextRank enables us to choose the most representative sentence containing that feature. [1]

After the extraction of the required sentences, we shorten them into phrases for summary generation purposes. We aggregate all the phrases generated to render a terse summary of product reviews.

5. Abbreviations

- TF-IDF - Term Frequency Inverse document Frequency
- NLP- Natural Language Processing
- NLTK - Natural Language ToolKit
- POS - Part Of Speech
- TF - Term Frequency
6. Conclusion

In this paper, we tried to find an optimal solution to all the difficulties that can occur during summarization of product reviews. The aim is to provide a utilitarian summary which would help product sellers and customers to get relevant knowledge from the summary.

We have used opinion mining and feature extraction approaches which are best suited for review summarization and solves existing problems with review summarization tools. Through our experiments we have found these approaches to be successful in increasing the efficacy and accuracy of product summary.

In our future work, our objective is to design efficient summarization generation technique which uses the phrases generated to render a human-like intelligently written summary.

Acknowledgement

We would like to thank the Computer Engineering department of Pimpri-Chinchwad College of Engineering for their support and guidance which enabled us in conducting our research.

References


