**Mining and Summarizing Customer Reviews of the Products using Data Mining and Natural Language Processing Techniques**

Semester project report for Data Mining, Dr. Xiaofei Zhang instructing Suraj Sharma

Department of Computer Science, University of Memphis, Memphis, TN, [sshrmspk@memphis.edu](mailto:sshrmspk@memphis.edu)

**ABSTRACT:**

Almost all of the products sold online has customer reviews. Sales of particular product depends on the reviews left by the previous buyers. So it is essential for both potential buyers and sellers to understand what those reviews are saying. But with the growing number of these reviews, thousands in some cases, it is difficult to make informed decision just by reading couple of the reviews and overall star rating. In this paper I present an application that mines and summarizes all the reviews of the product into key features and provide rating for each of these features. I have used Data Mining and Natural Language Processing techniques to create this application. I have also presented experimental results using reviews of different products sold on Amazon and demonstrated the effectiveness of our techniques by comparing it with the overall 5 star rating of Amazon for each product.

**INTRODUCTION:**

With increasing trend on online shopping and interaction, and with millions of products sold online, the number of customer reviews have grown drastically. Choosing the right product to buy from the pool of products has become difficult for the buyers as it is nearly impossible to read through and analyze hundreds of reviews on a popular product. It also makes difficult for the seller, who sells number of products online, to keep track and manage the customer reviews for each product.

In this paper I will be creating a feature based summarizing application using data mining and natural language processing models for mining and summarizing all the customers’ reviews. I am hoping this will allow new buyers to understand the key experiences other customers had with the given product, and also help the seller manage and understand the customer experiences for his/her items efficiently.

**RELATED WORK:**

Similar research has been done in summarizing the customer reviews but they have their own limitations. In [1] Minqing Hu and Bing Liu has considered the opinion sentences in each review to categorize the review as positive or negative, but their application doesn’t generate a summary of the reviews with significant features of the product. There are some other applications that provides the textual summary of each feature of the product but not an overall summary or the ratting.

In [2], the author has proposed reliable approach to build an algorithm for summarizing the customer reviews implementing a unique feature and opinion based analysis.

My research mainly use findings from Minqing Hu and Bing Liu [1], and Hanni and Patil [2].

**APPROACH:**

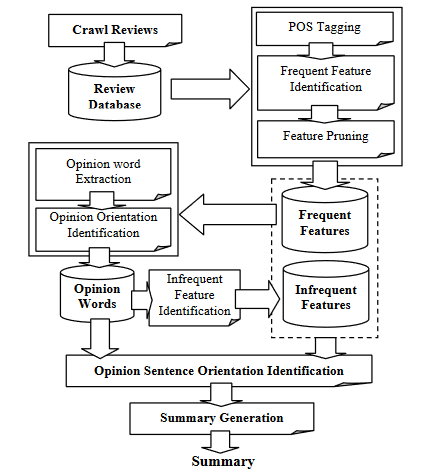
In [1] author has generated feature-based summaries using following 3 step;

1. Identifying feature of the product that customer has expressed their opinions on (called product features)
2. For each feature, identifying review sentences that give positive or negative opinions
3. Producing a summary using the discovered information.

This approach will first find all the *product features* from the pool of reviews for that item, then the number of reviews that expresses positive/negative opinion on each feature are counted. And based on the quantitative analysis of the features, a summary is generated to reflect the findings.

While the approach mentioned in [1] simplifies review summarization, it is not effective for all reviews. Some comments might be classified as negative, but that doesn’t mean the user doesn’t like any of the other features of the product. So, author in [2] uses some of the techniques explained in [1] but also uses unified opinion mining and sentiment analysis framework with Natural Language Processing approach to produce more realistic review summary.

I will be using following approach which both [1] and [2] uses with some modifications;



**Figure 1: Feature-based opinion summarization**

1. **Data Collection: Reviews Extraction:**

I have selected 4 different products sold on Amazon for this project; iPhone Case, Smart watch, Smart Garage door opener, and Kids safety glasses. Links to these products are given in Table 1. Reviews for these four products are extracted by web-crawling [3] [10] using python code. Code is provided in the deliverables.

1. **Tokenization:**

Reviews (review body/paragraph) collected are then parsed using NLProcessor linguistic parser [8] into individual sentences and then into meaningful units/words.

1. **POS:**

Product features are usually nouns or noun phrases, and adjective describes the noun in the review. Thus Parts of Speech tagging is crucial step. The words that are parsed from the reviews are tagged to see whether the word is noun, verb, adjective, etc. using Parts of Speech Tagging (POS)[8]. This helps in feature and opinion extraction.

1. **Frequent Features:**

After tagging the words, features of the products which are represented by noun in a sentence are extracted. Some pre-processing of the words is performed to remove stopwords, stemming and fuzzy matching to deal with word variants and misspellings. [4]

I also use association rule mining to find the frequent itemsets. Itemset is simply words or phrases that occurs together in sentences. [1] Not all features collected are genuine features, so I perform 2 types of pruning;

1. **Compactness Pruning:**

This method checks features that contain at least 2 words, *feature phrases*, and remove those that are likely to be meaningless. Feature phrases are considered meaningless by identifying features phrases whose words do not appear together in a specific order.

1. **Redundancy Pruning:**

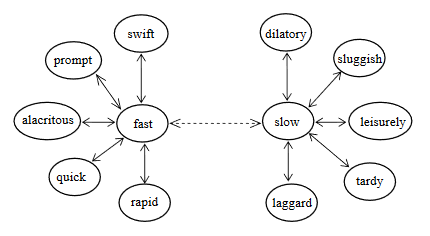
This step focuses on removing redundant features that contain single words. I use the concept of p-support. P-support of feature *ftr* is the number of sentences that *ftr* appears in as a noun or noun phrase, and these sentences must contain no feature phrase that is a superset of *ftr.*

Minimum p-support value is used to prune those redundant features. For instance, *life* by itself is not a useful feature while *battery life* is a meaningful feature phrase [4]

1. **Opinion Extraction:**

This is crucial part of the research where the opinion of the customers on the features collected is extracted. Opinion regarding these identified features are collected by analyzing the adjective in the sentence. [1] suggests using adjective synonym set and antonym set in WordNet [5] to predict the semantic orientations of adjectives.

In WordNet, adjectives are organized into bipolar clusters, as illustrated in Figure 2. The cluster for *fast/slow*, consists of two half clusters, one for senses of *fast* and one for senses of *slow*. Each half cluster is headed by a *head synset*, in this case *fast* and its antonym *slow*. Following the *head synset* is the *satellite synsets*, which represent senses that are similar to the sense of the head adjective. The other half is headed by the reverse antonymous pair *slow/fast*, followed by satellite synsets for senses of *slow* [1]



**Figure 2: Bipolar adjective structure, (-> similarity; ----> Antonymy)**

For those adjectives that WordNet cannot recognize, they are discarded as they may not be valid words. Similar opinions are grouped together.

1. **Analysis:**

After identifying actual features, and opinion related to the features, pair of feature and opinion are collected to identify the features that are frequent and most related to the product. The sentences are then analyzed and classified into positive, neutral, and negative sentiments using Sentiment analysis training set [2] with score(x) given to each statement between -1 <=x <= 1. Neutral statements are scored zero.

After analysis, feature based summary is generated according to the frequency of their appearances in the reviews. It is represented by a rating bar on a scale of 0-5; 0 being the worse and 5 being the best score. A paragraph summary which includes statements that speak of the features in their review is also generated using Sentiment analysis and the analysis of feature-opinion pairs along with the overall statement score. This provides detail overview of the opinions of the users and stresses on each feature of the product, making it easier for both customer and seller to know the response of the public.

**EXPERIMENTS AND RESULTS:**

Feature-Based Summarization has been implemented using the techniques proposed in Python. I have adopted following simple steps to achieve our goad;

1. Extract features (nouns) from each sentence in the reviews.
2. Extract close opinions (adjectives) about the features in each sentence.
3. Generate rating of each key-feature in the scale of 1-5, 5 being the best, based on the opinions collected.
4. Generate readable sentence, phrase to summarize the reviews

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.N** | **Summarization Results** | | |  |  |
| **Product link** | **Features and rating** | **Review Summary** | **Amazon Review** | **Number of Reviews** |
| 1 | **I-Phone Case:** <https://www.amazon.com/TORRAS-iPhone-Plastic-Coating-Compatible/dp/B0759JNCJ3/ref=cm_cr_arp_d_product_top?ie=UTF8> | **Case**: 4.42  **Fit**: 3.7  **Screen**:3.52  **Feel**: 3.45  **Service**: 4.9  **Cut**: 4.9  **Color**: 4.2 | Great, slim and thin case. Perfect fit.  Slim and great feel.  Protects screen. Great service. Perfect cut. | 4.6/5 | 2059 |
| 2. | **Smart Watch:**  <https://www.amazon.com/Padgene-Bluetooth-Smart-Watch-Camera/dp/B00X9L90RG/ref=cm_cr_arp_d_product_top?ie=UTF8> | **Watch**: 4.08  **Work**: 2.98  **Card**: 3.53  **Charge**: 2.1  **Price**: 3.67  **Band**: 3.70 | Great smart watch. Works with iphone. Has sim card and sd-card slot. Poor charge. Over price. Cheap band. | 3.1/5 | 1943 |
| 3 | **Smart Garage Door Opener**  <https://www.amazon.com/dp/B07H57NPRV/ref=cm_sw_em_r_mt_dp_U_ZoZWCbJNXYKVQ> | **Work**: 4.06  **Door**: 4.41  **Service**: 4.72  **Product**: 4.48  **Garage**: 4.74 | Works great, easy app. Great smart garage door opener. Excellent service. Awesome product. | 4.0/5 | 2634 |
| 4 | **Kids safety glasses.**  <https://www.amazon.com/dp/B07FXPDK9M/ref=cm_sw_em_r_mt_dp_U_17ZWCb5C8GXH5> | **Item**: 0  **Gun**: 4.8  **Party**: 0  **Price**: 4.8 | Not item for outdoor and Gun. Great Price.  Horrible product. | 2.3 | 9 |

**Table 1: Result of the summarization**

**CONCLUSION:**

Upon running the proposed technique, I were able to get Table 1 result. Looking closely with the result I are able to clearly say that our finding is almost close to the amazon rating. Besides providing just one overall rating like Amazon does, our research also provides key features and dissect the rating into deeper level providing in depth understanding of the product. Our research also provides human readable phrases that gives enough information about the product.

While implementing the proposed technique, I saw several challenges with language processing and sentiment analysis. Here are some of the challenges and findings;

1. Higher the number of reviews, the better the rating process would be. Since I didn’t have enough reviews for product 4 (Kids safety glasses), I were unable to extract accurate features and their ratings, while other products have enough reviews to accurately rate them.
2. Not everyone write reviews in grammatically correct English. Since I were looking for certain key features, it was challenging to find an exact feature described in the review. As an example, “*dimension”* was one of the feature of the case I were analyzing, I found some reviews had “*size”* in the review. I used WorldNet synsets to catch similarity between features collected. But even then due to ambiguous nature of writing and unstructured data, this is still not perfect.
3. Some of the reviews didn’t have any content but just title, so I had to manually process the data.

**FUTURE WORK:**

Even though I were able to parse the reviews and collect features and the rating for them, I still have some opportunity to improve our performance. Just collecting adjectives gave us fair idea about the feature, but I need to work deeper to look into phrases and the structure of the sentence to better understand the opinion of the customer. Thus the future work would be to improve the algorithm to look wider than just the adjectives to get the true opinion of the reviewer.

**REFERENCES:**

[1] Minging Hu and Bing Liu. “Mining and Summarizing Customer Reviews.” KDD. August 2004. <https://dl.acm.org/doi/abs/10.1145/1014052.1014073>

[2] Akkamahadevi R Hanni and Mayur M Patil. “Summarization of Customer Reviews for a Product on a website using Natural Language Processing.” ResearchGate. September 2016. <https://www.researchgate.net/publication/306285022_Summarization_of_Customer_Reviews_for_a_Product_on_a_website_using_Natural_Language_Processing>

[3] <https://www.webscraper.io/>

[4] Minging Hu and Bing Liu. “Mining Opinion Features in Customer Reviews.” To appear in AAAI’04, 2004. <https://dl.acm.org/doi/10.5555/1597148.1597269>

[5] Miller, G., Beckwith, R, Fellbaum, C., Gross, D., and Miller, K. “Introduction to WordNet: An on-line lexical database”. International Journal of Lexicography (special issue), 3(4):235-312.

[6] <http://www.nltk.org/_modules/nltk/probability.html>

[7] <https://gawron.sdsu.edu/python_for_ss/course_core/book_draft/web/web_intro.html>

[8] <https://textminingonline.com/dive-into-nltk-part-iii-part-of-speech-tagging-and-pos-tagger>

[9] <https://glowingpython.blogspot.com/2014/09/text-summarization-with-nltk.html>

[10] <https://www.scrapehero.com/amazon-review-scraper/>

[11] <https://gist.github.com/scrapehero/cefaf014076b953f865a63ad453d507b>