Summarization and opinion detection of product reviews

Product review nowadays has become an important source of information, not only for customers to find opinions about products easily and share their reviews with peers, but also for product manufacturers to get feedback on their products. With the rapid expansion of e-commerce, more and more products are sold on the Web, and more and more people are also buying products online. In order to enhance customer satisfaction and shopping experience, it has become a common practice for online merchants to enable their customers to review or to express opinions on the products that they have purchased. With more and more common users becoming comfortable with the Web, an increasing number of people are writing reviews. As a result, the number of reviews that a product receives grows rapidly. Some popular products can get hundreds of reviews at some large merchant sites. Furthermore, many reviews are long and have only a few sentences containing opinions on the product. This makes it hard for a potential customer to read them to make an informed decision on whether to purchase the product. If he/she only reads a few reviews, he/she may get a biased view. The large number of reviews also makes it hard for product manufacturers to keep track of customer opinions of their products. Hence to solve this problem we aim to mine the features on which the customers have expressed their opinions and to know about the opinion whether it is positive or negative for the particular review sentence and then features based summarization along with the orientation of the review.

1. Introduction

With the rapid expansion of e-commerce, more and more products are sold on the Web, and more and more people are also buying products online.With more and more common users becoming comfortable with the Web, an increasing number of people are writing reviews. As a result, the number of reviews that a product receives grows rapidly. Some popular products can get hundreds of reviews at some large e-commerce sites.Furthermore, many reviews are long and have only a few sentences containing opinions on the product. This makes it hard for a potential customer to read them to make an informed decision on whether to purchase the product. If he/she only reads a few reviews, he/she may get a biased view. The large number of reviews also makes it hard for product manufacturers to keep track of customer opinions of their products.So in this project we are generating feature based summaries of customer reviews of products on those product features on which the customers have expressed their opinions.This summarisation with opinion dectection task involves three subtasks ie. (1) Identify the product features from the reviews of customers on which they have expressed their opinions, (2) For each extracted product feature identify the reviews that give the positive and negative opinions, (3) Then produce the summary from the above discovered information.
Let us use an example to illustrate a feature-based summary. Assume that we summarize the reviews of a particular digital camera. The summary contains graph and some useful reviews example:

Cannon Digital camera:

picture quality
Positive:9

Awesome picture quality.
It looks very crystal clear resolution

Negative:3

Picture lacks sharpness.
Nightmode is not at all good.

Our task is different from traditional text summarization in a number of ways. First of all, a summary in our case is structured rather than another (but shorter) free text document as produced by most text summarization systems. Second, we are only interested in features of the product that customers have opinions on and also whether the opinions are positive or negative. We do not summarize the reviews by selecting or rewriting a subset of the original sentences from the reviews to capture their main points as in traditional text summarization. As indicated above, our task is performed in three main steps:
(1) Mining product features that have been commented on by customers. We make use of both data mining and natural language processing techniques to perform this task.
(2) Identifying opinion sentence for each feature and deciding whether each opinion sentence is positive or negative. Note that these opinion sentence must contain one or more product features identified above. To decide the opinion orientation of each sentence (whether the opinion expressed in the sentence is positive or negative), we perform three subtasks. First, a set of adjective words (which are normally used to express opinions) is identified using a natural language processing method. These words are also called opinion words in this paper. Second, for each opinion word, we determine its semantic orientation, e.g., positive or negative. This task is completed using SentiWordNet. Finally, we decide the opinion orientation of each sentence.
(3) Summarizing the results. This step aggregates the results of previous steps and presents them in the format of above shown example.

2. Related Work

Our work is closely related to Hu and Liu's work on Mining and Summarizing Customer Reviews. We also followed the work of Ziheng Lin's work on Product Review Summarization. We have used the following tools also:

(1) Scraping: For scraping Jsoup is used to crawl the data of the flipkart. It provides a convenient interface to fetch content from the web, and parse them into Documents.

(2) Stanford Dependency Parser: A natural language parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences.

(3) SentiWordNet: It is used as machine learning tool to determine the polarity of the opinion words and the reviews.

3. Approach

3.1. Assumptions

We consider only product features that appear as nouns; our method has the limitation that it cannot handle implicit features that are not explicitly mentioned. To explain this crucial point, suppose the following two sentences from camera reviews:
(1) The pictures of this camera are very clear.
(2) The camera fits nicely into my palm.
In the sentence (1), the user expresses his/her satisfaction about the quality of the picture taken by the camera, and we can infer that the noun picture is a feature of the camera. On the other hand, the sentence (2) discusses the size of the camera. However, the word size does not appear explicitly in the sentence.

3.2. Architecture

Figure 1 gives the architectural overview of our feature based opinion summarization system.

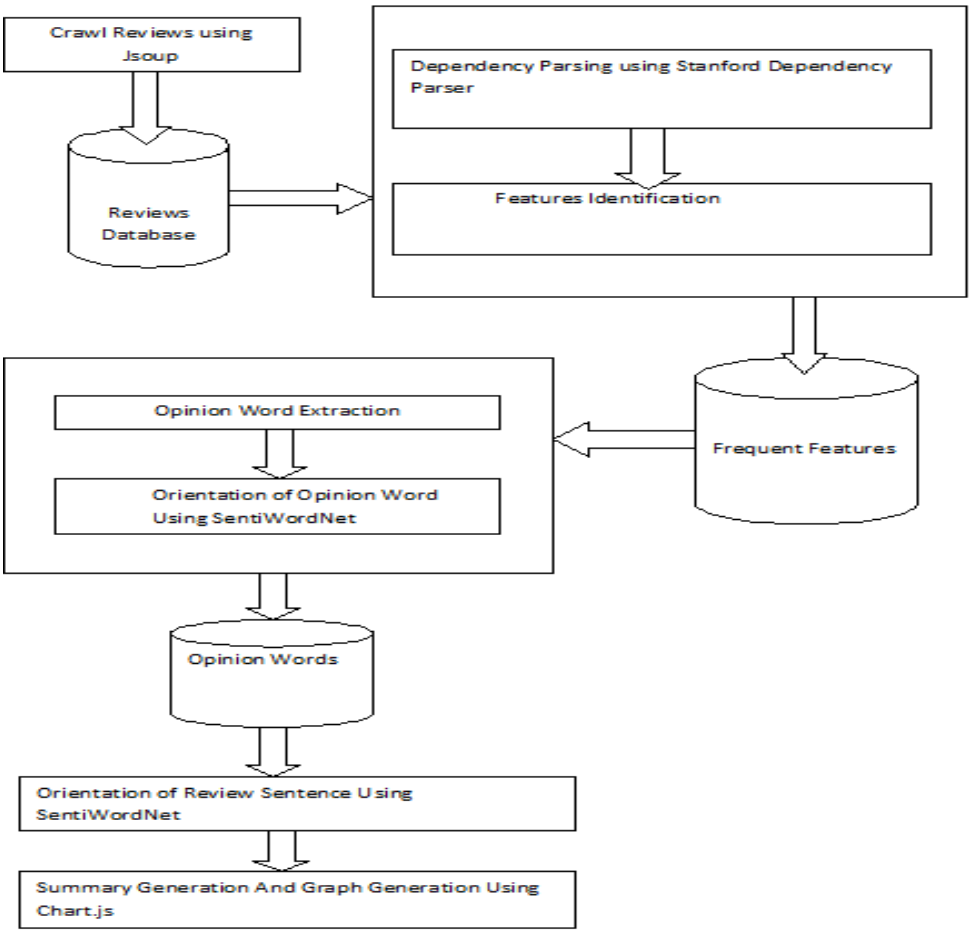

Figure 1 : Summarization and Opinion Detection of Product Reviews

3.3. System Overview

The inputs to the system are a product name and all the reviews of that product. The output is the summary of the reviews as the one shown in the introduction section with the graph.
Given the inputs, the system first crawls all the reviews, and put them in the review database. It then finds those frequent features that many people have expressed their opinions on. After that, the opinion words are extracted using the resulting frequent features, and semantic orientations of the opinion words are identified with the help of SentiWordNet. Then the orientation of each opinion sentence is identified, and a final summary is produced. The Dependency parsing from natural language processing helps us to find features and their opinion words from the opinion sentences. Below, we discuss each of the sub-steps in turn.

3.4. Scraping

Using Jsoup we crawl the data of a ecommerce site for the given link of the product. We extract all the reviews and title of that particular product. Output of this procedure will the text file containing title and all the reviews of the input product.

3.5. Product Feature Identification

We used the Stanford Dependency Parser to parse the reviews. Consider the following sentences parsed by Stanford Dependency Parser:

(1) The larger lens of the g3 gives better picture quality in low light.
nsubj(gives-7, lens-3), dobj(gives-7, quality-10), etc...

(2) When I took outdoor photos with plenty of light, the photos were awesome.
dobj(took-3, photos-5),nsubj(awesome-14, photos-12),etc...

According to the examples above, we observe that genuine features tend to appear as either subjects or objects within the sentences. In fact, our analysis on a subset of camera reviews (more than 150 sentences) shows that more than 90% of the instances correspond to the above observation. This is not too surprising as subjects and objects in the sentences are usually the targets at which the users express their opinions. These findings suggest that we can filter non-subject and nonobject nouns from the set of identified features. We are removing also those nouns that does not appear above the certain number of times in the all the reviews of the product.

3.6. Opinion Identification of opinion sentence

There are three steps in this process:
(a) Opinion Word Extraction: Consider the above review of a camera ie.

(1) The larger lens of the g3 gives better picture quality in low light.

Stanford parser parse the above review and give the following relations also :
amod(picture-9, better-8), etc...
So in the above parsed result the opinion word related to "picture" is "better" . In this way the associated adjectives(opinion words) to a feature are extracted.

(b) Orientation Identification for Opinion Words: For each opinion word, we need to identify its semantic orientation, which will be used to predict the polarity of each opinion sentence. The polarity of a word indicates the direction that the word deviates from the norm for its semantic group. Words that encode a desirable state (e.g., beautiful, awesome) have a positive orientation, while words that represent undesirable states have a negative orientation (e.g., disappointing). For this we used the SentiWordNet to know the orientation of the opinion words.

(c) Predicting the Orientations of Opinion Sentences: Now after getting the orientation of the individual opinion words we predict the orientation of the whole review sentence for the particular feature whether the review is positive or negative for that feature. In general, we use the dominant orientation of the opinion words in the sentence to determine the orientation of the sentence. That is, if positive/negative opinion prevails, the opinion sentence is regarded as a positive/negative one. In the case where there is the same number of positive and negative opinion words in the sentence, we predict the orientation using the average orientation of effective opinions or the orientation of the previous opinion sentence (effective opinion is the closest opinion word for a feature in an opinion sentence).We also take care of negative words to determine the orientation of the whole review. Consider a review i.e. "the camera is not easy to use".
Here the opinion word if "easy" which is positive but if we see the whole review then we observe that the whole sentence is negative because of "not" related to "easy". So we extract these related negative words for the opinion words using the Dependency Parser relations.

3.7. Summary Generation:

After all the previous steps, we are ready to generate the final feature-based review summary, which is straightforward and consists of the following steps:
• For each discovered feature, related opinion sentences are put into positive and negative categories according to the opinion sentences' orientations. A count is computed to show how many reviews give positive/negative opinions to the feature.
• All features are ranked according to the frequency of their appearances in the reviews. Feature phrases appear before single word features as phrases normally are more interesting to users. Other types of rankings are also possible. For example, we can also rank features according the number of reviews that express positive or negative opinions.
Here we are giving summary on one product ie. "Apple 16GB iPad Mini with Wi-Fi"

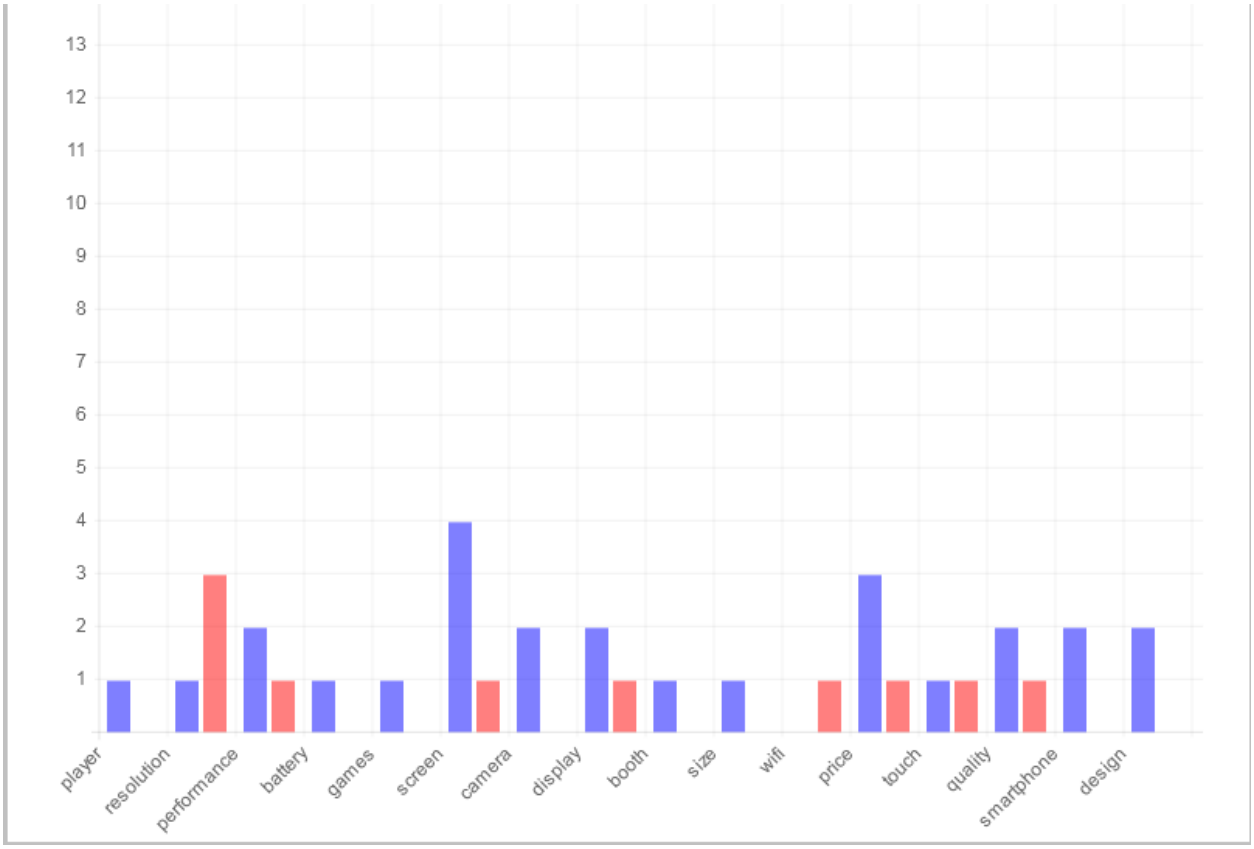

Figure 2 : Graph Summary for Apple 16GB iPad Mini with Wi-Fi

player

Positive

1) It will not let you transfer files except iTunes which needs to be installed in your PC. 2) It does not support video formats except mp4 and m4v niether it has a free player like MxPlayer in android. 3) Battery problems after an year. This is from my personal experience. If you still want to go for Apple products then its your take.

resolution

Positive

nice apple product..wish it would have larger screen like ipad 2.good resolution..nice for reading ebooks..playing games..but some apps are useless without presence of wifi..

Negative

Cons - No retina display (very low resolution) Still On A5 chip No Siri (Even iPod touch 5G Has got Siri Cost (with 3k more you can buy iPad 2) Not the best 7 inch tab yet,wait for iPad mini 2 which will come with retina display.
I have purchased last oct 13 and using it rough and tough. Battery life is amazing. Never hang till date. Running more than 20 application at a time but no problem observed. Very happy to use. Only resolution is slightly disappointed.

performance

Positive

Do not buy any other tablet wait and buy this. This mini is too good handy and awesome performance :) Apple tusi great ho !!!!
The timing and the delivery of the product was as per schedule and as promised. The performance of the product is also very good considering that this is my first experience

Figure 3 : Text Summary for Apple 16GB iPad Mini with Wi-Fi

4. Evaluation and Results

The type for the evaluation of our tool will be manual. We will evaluate this summarization in the following three perspectives:
1.The accuracy of product feature extracted.
2.The accuracy of the opinion sentence extraction
3.The accuracy of the opinion prediction for that sentence.
We compared our results manually with flipkart results. We are getting approximately 60% product features.

5. Conclusion

In this paper, we proposed a set of techniques for mining and summarizing product reviews based on data mining and natural language processing methods. The objective is to provide a feature-based summary of a large number of customer reviews of a product sold online. Our experimental results indicate that the proposed techniques are very promising in performing their tasks. We believe that this problem will become increasingly important as more people are buying and expressing their opinions on the Web. Summarizing the reviews is not only useful to common shoppers, but also crucial to product manufacturers. Through our project we are providing a better and easy way for online customer to decide whether to purchase the product or not.
In our future work, we plan to further improve and refine our techniques, and to deal with the outstanding problems identified above, i.e., pronoun resolution, determining the strength of opinions, and investigating opinions expressed with adverbs, verbs and nouns. We also plan to take care of implicit features of the product.

References

- http://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf
- http://www.comp.nus.edu.sg/~kanmy/papers/Khang-TP-final.pdf
- https://www.ideals.illinois.edu/bitstream/handle/2142/18702/survey_opinionSummarization.pdf?sequence=2
- http://gate.ac.uk/sale/eswc11/opinionmining.pdf
- http://www.seas.upenn.edu/~cse400/CSE400_2009_2010/final_report/Schaye_Feczko.pdf