Mining and Summarizing Customer Reviews

By Team 36

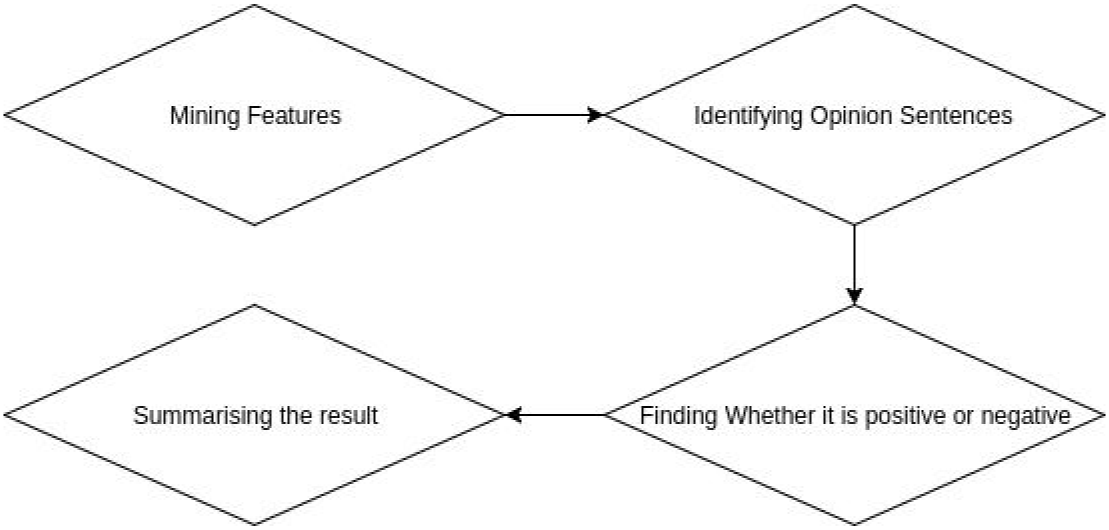
Abstract

Through this project our aim is to mine and summarize all the customer review of a product. We will mine the features of the product on which customer have expressed their opinions and whether they are positive and negative. Our task is done in three steps-:

- Mining product features that have been commented on by customers
- Identifying features of the product that customers have expressed their opinions on (called product features)
- For each feature, identifying review sentences that give positive or negative opinions
- Producing a summary using the discovered information

Approach

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. In this we used various libraries like **NLTK** to mine the features.


Mining Features
Identifying Opinion Sentences
Finding Whether it is positive or negative
Summarising the result

POS Tagging:-

We used the NLProcessor linguistic parser to parse each review to split text into sentences and to produce the part-of-speech tag for each word (whether the word is a noun, verb, adjective, etc).Each sentence is saved in the review database along with the POS tag information of each word in the sentence. A transaction file is created for generation of frequent features. In this file , each line has words identified as noun or noun phrases from a single file. Some pre-processing of words is also performed, which includes removal of stopwords, stemming and fuzzy matching.

Stop words are words which are filtered out before or after data.These a words such as *the*, *is*, *at*, *which*, and *on.*

Lemmatization is the process of converting the words of a sentence to its dictionary form. For example, given the words amusement, amusing, and amused, the lemma for each and all would be amuse**.**

Fuzzy matching is used to deal with word variants and misspellings.

Frequent Features Identification:-

This step identifies product features from the nouns on which the users have expressed their opinions. Here, our focus is on finding the frequent(common) features.We use apriori algorithm to find the frequent itemsets.It is common that a customer review contains many things that are not directly related to product features. Different customers usually have different stories. However, when they comment on product features, the words that they use converge.

Feature Pruning

Compactness pruning: This method checks features that contain at least two words, which we call feature phrases, and remove those that are likely to be meaningless.

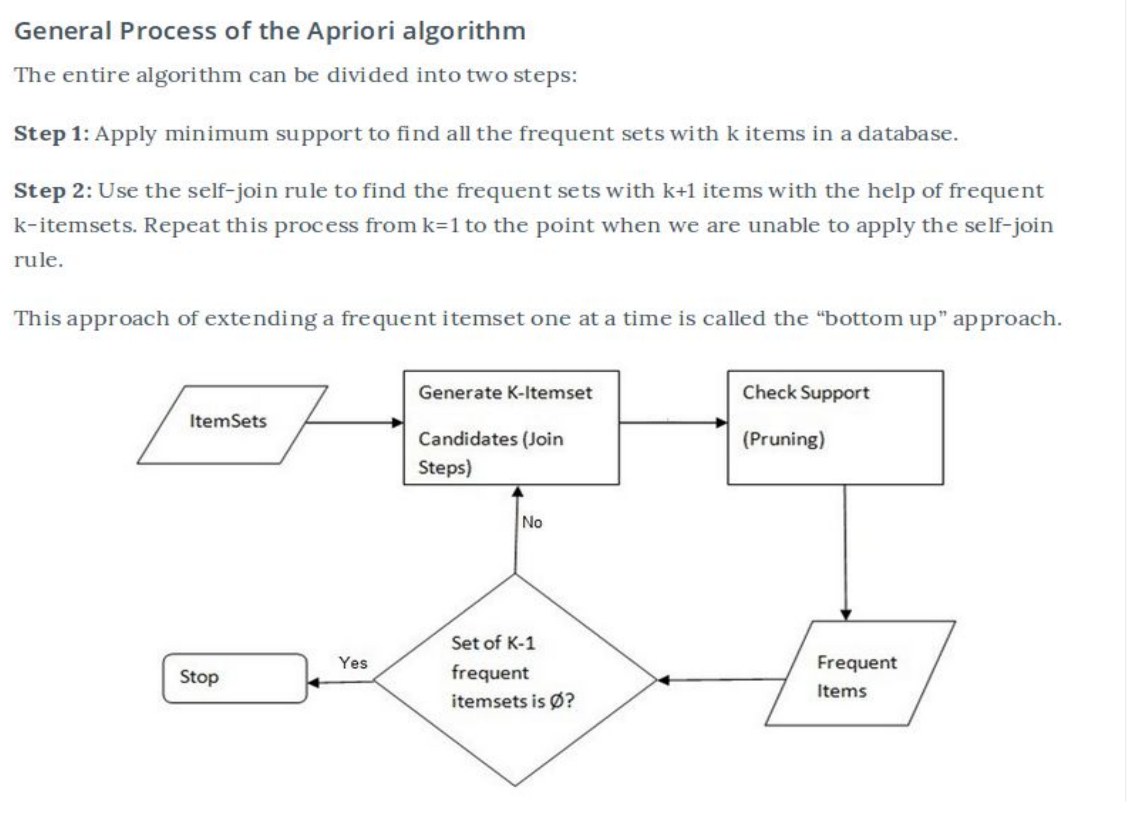
Redundancy pruning: In this step, we focus on removing redundant features that contain single words. To describe redundant features, we have the following definition.

APRIORI ALGORITHM

Short stories or tales always help us in understanding a concept better but this is a true story, Wal-Mart's beer diaper parable. A sales person from Wal-Mart tried to increase the sales of the store by bundling the products together and giving discounts on them. He bundled bread and jam which made it easy for a customer to find them together. Furthermore, customers could buy them together because of the discount.

A key concept in Apriori algorithm is the anti-monotonicity of the support measure. It assumes that ---

1. All subsets of a frequent itemset must be frequent.
2. Similarly, for any infrequent itemset, all its supersets must be infrequent too.



General Process of the Apriori algorithm

The entire algorithm can be divided into two steps:

Step 1: Apply minimum support to find all the frequent sets with k items in a database.

Step 2: Use the self-join rule to find the frequent sets with k+1 items with the help of frequent k-itemsets. Repeat this process from k=1 to the point when we are unable to apply the self-join rule.

This approach of extending a frequent itemset one at a time is called the "bottom up" approach.

ItemSets
Generate K-Itemset Candidates (Join Steps)
Check Support (Pruning)
Frequent Items
Set of K-1 frequent itemsets is Ø?
No
Yes
Stop

Opinion Prediction and Extraction :-

We utilize the adjective synonym set and antonym set in WordNet to predict the semantic orientations of adjectives. In WordNet, adjectives are organized into bipolar clusters.our strategy is to use a set of seed adjectives, which we know their orientations and then grow this set by searching in the WordNet. To have a reasonably broad range of adjectives, we first manually come up a set of very common adjectives. Procedure Orientation Prediction takes the adjective seed list and a set of opinion words whose orientations need to be determined. It calls procedure Orientation Search iteratively until no new opinion word is added to the seed list. Every time an adjective with its orientation is added to the seed list, the seed list is updated; therefore calling Orientation Search repeatedly is necessary in order to exploit the newly added information.

Procedure Orientation Search searches WordNet and the seed list for each target adjective word to predict its orientation. It searches synset of the target adjective in WordNet and checks if any synonym has known orientation. If so, the target orientation is set to the same orientation as the synonym and the target adjective along with the orientation is inserted into the seed list . Otherwise, the function continues to search antonym set of the target word in WordNet and checks if any antonym has known orientation. If so, the target orientation is set to the opposite of the antonym and the target adjective with its orientation is inserted into the seed list. If neither synonyms nor antonyms of the target word have known orientation, the function just continues the same process for the next adjective since the word's orientation may be found in a later call of the procedure with an updated seed list. For those adjectives that WordNet cannot recognize, they are discarded as they may not be valid words. For those that we cannot find orientations, they will also be removed from the opinion words list.

Predicting the Orientations of Opinion Sentences:-

In this we are predicting the orientation of an opinion sentence, i.e., positive or negative. Orientation of the opinion words in the sentence to determine the orientation of the sentence.

For this first we have identified the orientation of each feature from the orientation of nearby adjective. The Algorithm as follows :

- In each sentence, for each feature in the sentence , we identify the closest adjective to the left of feature with a range of 2 words.

  For eg : *good battery, awesome image etc.*
- Then we identify the closest adjective to the right of feature with a range of 6 words.

  For eg : *camera quality is nice, body is too fragile*
- If we encounter any stop word (‘ . ’ ‘ , ’ ‘!’ ‘?’ ‘and’ ‘or’ ‘but’) while searching for an adjective we stop right there.

  For eg : *material is pure cotton but texture is rough*
- The orientation of the feature will be same as that of closest adjective.

After we get feature orientation, sentence orientation is predicted, for that we follow the following Algorithm :

- If the net sum of adjectives orientation is positive in the sentence then the sentence is predicted as positive
- If the sum is negative then, sentence is predicted as negative.
- If the sum is 0 i.e positive adjectives equals the negative adjectives, then sentence orientation is predicted by the feature effective opinion in that sentence
- The feature orientation is calculated in the above part, and we use that to get the sentence orientation

After the sentence orientation , it is quite easy to get the review orientation.

Summary Generation:-

To generate final feature based review we have done the following steps:-

- For any product feature we put opinion sentences into positive and negative category depending on the opinion sentence orientation.
- A count is generated to show how many reviews gives positive/negative opinions to the feature.
- All features are ranked according to the frequency of appearance in the reviews.
- Feature phrases appear before single word features as phrases normally are more interesting to users.
- We can also rank features based on the number of reviews that express positive/negative opinions.

Experimental Evaluation:-

A system, called FBS (Feature-Based Summarization), based on
the proposed techniques has been implemented. We have to evaluate FBS from three perspective:-

1. The effectiveness of feature extraction
2. The effectiveness of opinion sentence extraction.
3. The accuracy of orientation prediction of opinion sentences.

We conducted our experiments using customer review of 4 products:- Mobile Phone, DSLR camera, mp3 player and Laptop. We first collected some reviews from amazon.com. From each of the review we extract a text review.

Manual Evaluation:-

We manually read all the reviews. If it shows user's opinion all the feature on which user has expressed his view are tagged. If the user gives no opinion then that is not tagged. For each product we produced a manual feature list having "No of Manual features" as a column and all the results generated by our systems are compared with the manually tagged results.

Issues:-

- A minor complication regarding feature tagging is that feature can be explicit or implicit. Both explicit and implicit features are easy to identify by the human tagger.
- Judging opinions in reviews can be somewhat subjective. It is usually easy to judge whether an opinion is positive or negative if a sentence clearly expresses an opinion. However, deciding whether a sentence offers an opinion or not can be debatable. For those difficult cases, a **consensus** was reached between the **primary human tagger** (the first author of the paper) and the **secondary tagger** (the second author of the paper).

Results:-

On applying the algorithm to some of the examples we found the average accuracy to be 71.02% precision to be 66% and recall to be 78%.

Conclusion:-

Our main objective was to provide a feature based summary of large number of customer reviews of a product sold. Our experimental result suggest that our technique is promising but it can be improved by further implementing pruning. Our results can also be improved by further adding pronoun resolution, determination of strength of opinion and investigating opinions expressed with adverbs, verbs and noun.