Data Mining and Customer Review

By
Kanak Garg
Mohit Gupta
Srijan Singh
Anam Raihan

Procedure

- 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
- Identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative.
- 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.
    - A bootstrapping technique is proposed to perform this task using WordNet.
- Summarizing the results

POS TAGGING

We used the NLTk library in python 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).

```
The camera is good but the body is plastic so it is fragile

[[('The', 'DT'), ('camera', 'NN'), ('is', 'VBZ'), ('good', 'JJ'), ('but', 'CC'), ('the', 'DT'), ('body', 'NN'), ('is', 'VBZ'), ('plastic', 'JJ'
), ('so', 'IN'), ('it', 'PRP'), ('is', 'VBZ'), ('fragile', 'JJ')]]
```

VBZ	**Verb (3rd person singular present)**
JJ	**Adjective**
NN	**Noun**
PRP	**personal pronoun**

POS TAGGING

Each sentence is saved in the review database along with the POS tag information of each word in the sentence.

A transaction file is then created for the generation of frequent features in the next step.

In this file, each line contains words from one sentence, which includes only the identified nouns and noun phrases of the sentence.

Some pre-processing of words is also performed, which includes removal of stopwords, lemmatization.

POS TAGGING

STOPWORD- Stopwords are words which are filtered out before or after processing of natural language data.(such as the, is, at, which and on)

LEMMATIZATION- Lemmatization is the process of converting the words of a sentence to its dictionary form. (given the words amusement, amusing, and amused, the lemma for each and all would be amuse).

Apriori Algorithm - step 1

Minimum support count - 2

Transaction ID	Items
1	1 3 4
2	2 3 5
3	1 2 3 5
4	2 5
5	1 3 5

Candidate Itemset	Support
1	3
2	3
3	4
4	1
5	4

Frequent Itemset	Support
1	3
2	3
3	4
5	4

Apriori Algorithm - step 2

Minimum support count - 2

Transaction ID	Items
1	1 3 4
2	2 3 5
3	1 2 3 5
4	2 5
5	1 3 5

Candidate Itemset	Support
1,2	1
1,3	3
1,5	2
2,3	2
2,5	3
3,5	3

Frequent Itemset	Support
1,3	3
1,5	2
2,3	2
2,5	3
3,5	3

Apriori Algorithm - step 3

Minimum support count - 2

Transaction ID	Items
1	1 3 4
2	2 3 5
3	1 2 3 5
4	2 5
5	1 3 5

Candidate Itemset	In FI 2
1,2,3 {1,2},{1,3},{2,3}	No
1,2,5 {1,2},{1,5},{2,5}	No
1,3,5 {1,3},{1,5},{3,5}	Yes
2,3,5 {2,3},{2,5},{3,5}	Yes

Frequent Itemset 2	Support
1,3	3
1,5	2
2,3	2
2,5	3
3,5	3

Apriori Algorithm - step 3

Minimum support count - 2

Transaction ID	Items
1	1 3 4
2	2 3 5
3	1 2 3 5
4	2 5
5	1 3 5

Candidate Itemset	In FI 2
1,2,3 {1,2},{1,3},{2,3}	No
1,2,5 {1,2},{1,5},{2,5}	No
1,3,5 {1,3},{1,5},{3,5}	Yes
2,3,5 {2,3},{2,5},{3,5}	Yes

Frequent Itemset 3	Support
1,3,5	2
2,3,5	2

Apriori Algorithm - step 4

Minimum support count - 2

Transaction ID	Items
1	1 3 4
2	2 3 5
3	1 2 3 5
4	2 5
5	1 3 5

Candidate Itemset	In FI 3
1,2,3,5 {1,2,3},{1,2,5}, {2,3,5},{1,3,5}	No

Candidate Itemset	Support
1,2,3,5 {1,2,3},{1,2,5}, {2,3,5},{1,3,5}	1

Frequent Itemset 3	Support
1,3,5	2
2,3,5	2

Opinion Words Extraction

These are words that are primarily used to express subjective opinions.

Thus the presence of adjectives is useful for predicting whether a sentence is subjective, i.e., expressing an opinion. This paper uses adjectives as opinion words. We limit the opinion words extraction to those sentences that contain one or more product features, as we are only interested in customers' opinions on these product features.

Opinion Words Extraction

Definition: *opinion sentence*

If a sentence contains one or more product features and one or more opinion words, then the sentence is called an *opinion sentence*.

We extract opinion words in the following manner (Figure 3):

```
for each sentence in the review database
  if (it contains a frequent feature, extract all the adjective
      words as opinion words)
    for each feature in the sentence
      the nearby adjective is recorded as its effective
      opinion. /* A nearby adjective refers to the adjacent
      adjective that modifies the noun/noun phrase that is a
      frequent feature. */
```

Figure 3: Opinion word extraction

For example, *horrible* is the effective opinion of *strap* in "*The strap is horrible and gets in the way of parts of the camera you need access to*." Effective opinions will be useful when we predict the orientation of opinion sentences.

Orientation Identification

- Utilize Adjective Synonym set and antonym set to predict the semantic orientations
- adjectives share the same orientation as their synonyms and opposite orientations as their antonyms
- use a set of seed adjectives, which we know their orientations
- adjectives that WordNet cannot recognize, they are discarded as they may not be valid word

Bipolar Structure

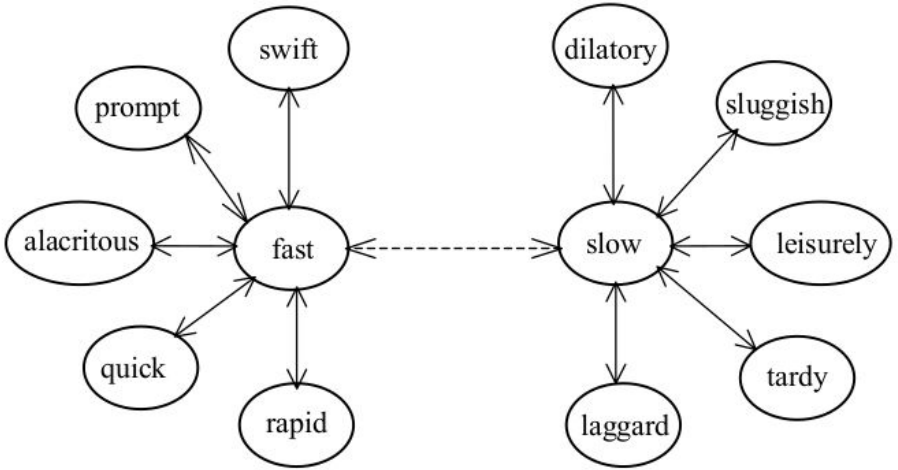

Figure 4: Bipolar adjective structure,
(→ = similarity; ----> = antonymy)

```
1.  Procedure OrientationPrediction(adjective_list, seed_list)
2.  begin
3.    do {
4.      size_1 = # of words in seed_list;
5.      OrientationSearch(adjective_list, seed_list);
6.      size_2 = # of words in seed_list;
7.    } while (size_1 ≠ size_2);
8.  end
```

```
1.  Procedure OrientationSearch(adjective_list, seed_list)
2.  begin
3.    for each adjective w_i in adjective_list
4.    begin
5.      if (w_i has synonym s in seed_list)
6.        { w_i's orientation= s's orientation;
7.        add w_i with orientation to seed_list; }
8.      else if (w_i has antonym a in seed_list)
9.        { w_i's orientation = opposite orientation of a's
                                orientation;
10.       add w_i with orientation to seed_list; }
11.   endfor;
12. end
```

Figure 5: Predicting the semantic orientations of opinion words

Feature Orientation

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.

Sentence Orientation

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.

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 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.

Feature Extraction

Products	Total Features (Manually)	Features Extracted by Program Correctly	Extra Features extracted
Samsung gear 3	41	29	7
Moto G5s plus	25	17	2
Sony Bravia (32)	45	25	7
Dell Inspiron	60	48	9

Feature Orientation

Product	Precision	Recall	Accuracy
Samsung Gear 3	0.901	0.861	0.82
Moto G5s Plus	0.73	0.77	0.69
Dell Inspiron	0.70	0.74	0.78
Sony bravia (32)	0.78	0.80	0.70

Sentence Orientation

Product	Precision	Recall	Accuracy
Samsung Gear 3	0.82	0.861	0.90
Moto G5s Plus	0.625	0.525	0.77
Dell Inspiron	0.68	0.72	0.72
Sony bravia (32)	0.701	0.692	0.68

Review Orientation

Product	Precision	Recall	Accuracy
Samsung Gear 3	0.80	0.80	0.90
Moto G5s Plus	0.80	0.86	0.80
Sony bravia(32)	0.88	0.80	0.70

Conclusion & Results

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.

Thank You