International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-3, Issue-2, May 2014

Review Mining: A New Approach using Modified NLP

Umang Sardesai, Aakash Makwana, Sagar Haria

Abstract: ***The Web has become an excellent source for gathering consumer opinions. There are now numerous Web sites containing such opinions, e.g., customer reviews of products, forums, discussion groups, and blogs. Nowadays we get all the technical specifications of a product on the Web, but what matters is what the customer feels about or what his opinions about the product are. This paper focuses on analyzing and summarizing online customer reviews of products. While analyzing we devise a new approach for NLP, by assigning a latent weight to each aspect/feature of a product. After extracting the sentiment in each sentence of the review, we summarize the opinions and express it graphically. This will not only help customers but also help the product manufacturers to get an indirect customer feedback.***

Key Words: NLP, Sentiment Analysis, Opinion mining, Latent weight, Visualization

I. 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. 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. For a product manufacturer, there are additional difficulties because many merchant sites may sell its products, and the manufacturer may (almost always) produce many kinds of products. In this paper, we study the problem of generating product-based summaries of customer reviews of products sold. Here, features broadly mean product features (or attributes) and functions. Given a set of customer reviews of a particular product, the task involves three subtasks: (1) identifying features of the product that customers have expressed their opinions on (called product features); (2) for each feature, identifying review sentences that give positive or negative opinions; and (3)producing a summary using the discovered information. Let us use an example to illustrate a product-based summary. Assume that we summarize the reviews of a particular digital camera, digital_camera_1. The summary looks like the following:

Product: Mobile_1
Feature: **picture quality**
Positive: 253
<individual review sentences>
Negative: 6
<individual review sentences>
Feature: **size**
Positive: 134
<individual review sentences>
Negative: 10
<individual review sentences>

Figure 1: An Example Summary

In Figure 1, picture quality and (camera) size are the product features. There are 253 customer reviews that express positive opinions about the picture quality, and only 6 that express negative opinions. The <individual review sentences> link points to the specific sentences and/or the whole reviews that give positive or negative comments about the feature. With such a product-based summary, a potential customer can easily see how the existing customers feel about the mobile. If he/she is very interested in a particular feature, he/she can drill down by following the <individual review sentences> link to see why existing customers like it and/or what they complain about.

II. RELATED WORK

Initially, the aspects in reviews were given orientation like positive , negative and neutral using NLP approach. An opinion is a quintuple, (e, a, s, h, t), where e is the name of an entity, a is an aspect of e, s is the sentiment on aspect a of entity e, h is the opinion holder, and t is the time when the opinion is expressed by h. The sentiment s is positive, negative, or neutral, or expressed with different strength/intensity levels, e.g., 1 to 5 stars as used by most review sits on the Web. Here, e and a together represent the opinion target. So, tuple formation consists of 6 steps, the first 5 steps being identification of each tuple item and 6th step being the formation of the quintuple.

Revised Manuscript Received on 30 May 2014.
* Correspondence Author
Umang Sardesai*, U.G. Student, Department of Computer DJSCOE, Vile-Parle (W), Mumbai, India.
Aakash Makwana, U.G. Student, Department of Computer, DJSCOE, Vile-Parle (W), Mumbai, India.
Sagar Haria, U.G. Student, Department of Computer, DJSCOE, Vile-Parle (W), Mumbai, India.

A basic example of the quintuple:
(Samsung, picture_quality, negative, bigJohn, Sept-15-2011)
So, basic format is
(Product, Aspect, Orientation, Author, Timestamp).

III. PROPOSED SOLUTION

Here, we give the idea of using the concepts of weight.
For our system, we consider mobile phone reviews. Hence all aspects will be related to aspect to mobile phones. Aspect of a product (here mobiles) is nothing but an attribute or feature of the product like phone camera, phone battery etc.
our model mainly is divided into two parts.

1. We assign weights to every aspect
2. After assigning weights to aspects, we start the actual sentiment analysis/ opinion mining using modified NLP.

ASSIGNING OF WEIGHTS:

We take reviews of various mobile products from flipkart.com. Considering an average of ten reviews per mobile product and considering 50 mobiles from different manufacturers we get approximately 500 reviews. We consider these reviews as a sort of training data. First we decide on the aspect keywords for mobile phones. This has to done manually. Going through whole list of reviews, we made a list of aspects which are related to mobile phones. The list of aspects includes p rice, camera, battery etc. and the list goes on. We also make a list of sub keywords for every aspect. For e.g. the camera aspect of a mobile phone will include image, pic, video picture quality etc and price will contain value, money, costs, as its sub keywords. Depending on keywords and sub keywords we search how many times an aspect has been mentioned in the reviews. Depending on the frequency of occurrence of aspect, we assign weights to it. If the aspect has a high frequency, it means it has been considered more as compared to the aspects having lower frequency. This means the aspect having higher frequency will have higher weight compared to the one having lower frequency. We use a sort of percentile system for assigning weight. The weights can have range from 0 to 1. Suppose the words and sub words related to aspect PRICE are mentioned 400 times (we are considering arbitrary values), the ones related to BATTERY are 300 and CAMERA 200. Thus considering PRICE has been mentioned maximum number of times, we will assign it a weight of '1'. BATTERY will be assigned a weight of 300/400 = 0.75 Thus the general formula to determine the weight will be

$$\textit{Weight for aspect } A = \frac{\textit{Frequency of occurrence of aspect } A}{\textit{Frequency of aspect with } \max \textit{ occurrence}}$$

Similarly weight for aspect CAMERA will be 200/400 = 0.5. Thus the aspects which are given higher importance by the customers will get a higher weight and vice versa.

MODIFIED NLP:

Thus we have modified our basic NLP tuple using the weight aspect like this.

(e, a, w, s, h, t)

where,
e is the name of an entity,
a is an aspect of *e*,
w is the weight of aspect a,
s is the sentiment on aspect *a* of entity *e*,
h is the opinion holder,
t is the time when the opinion is expressed by *h*.

WHY DO WE NEED WEIGHTS?

While writing where the user may give unnecessary importance to certain aspects of a product which are not required. Thus by assigning weights to every aspect, the important ones will contribute more to the overall rating of the product whereas unimportant ones will not be completely excluded and by assigning them a lesser weight ,their contribution to the overall rating will be reduced. In the earlier models when weights were not assigned, the unimportant aspects were either completely exempted or included with same weightage as that of important ones. This led to inaccurate interpretation of review. By assigning weights we have optimized the basic NLP algorithm and have achieved better and accurate results.

SENTIMENT ANALYSIS:

It consists of following steps
Step 1:
Download or Crawl through the opinions.
Step 2:
POS (Part of speech Tagging)
Involves using of NLP parser i.e. breaking sentence into parts.
e.g.: I am in complete awe of this camera.
Here I is a noun, complete is an adverb and so on.
Step 3: Feature Identification
1. The picture are very clear.
2. While light, it easily fits into pocket.
Here 1. directly implies that the reference is made to the camera whereas 2. represents that an indirect reference is made to the size of the product.
Step 4: Opinion word extraction
For each feature a nearby adjective is recorded as its effective opinion.
e.g.: The strap is horrible.
This sentence implies that the adjective 'horrible' is assigned to the noun strap.
Step 5: Orientation Identification
We identify whether the adjective stated implies positive or negative effect. Here we assign the aspect rating depending on the degree
We use senti-word net for this.
Step 6: Overall Rating
Now the aspect rating obtained is used along with aspect weight to obtain overall rating. If *Wi* and *Ri* is the aspect weight and aspect rating respectively for a particular aspect *Ai*, then

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$Overall\ Rating = \sum_{i=0}^{n} WiRi.$$

where *n* is the number of aspects for a product

Step 7: Summarization This is the final step which involves graphical representation. Using the graphs which we show, the user can get overall rating or aspect level rating of the product.

IV. IMPLEMENTATION

1. Extraction:

This is a very important section when it comes to BIG DATA. The reviews are to be extracted from flipkart.com. The process of extracting data from the web is called as "Web Scraping" or "Screen Scraping". There are various languages with which we may do scraping but the best are Python and Perl. I decided to go with Python. Python has wonderful library called "Beautiful Soup", which makes screen scraping a breeze. Beautiful Soup provides a few simple methods for navigating, searching, and modifying a HTML page which is essential during scraping. Initially, we specify a mobile phone on the GUI of python shell. Then, the job is to find the product on Flipkart. Once that done we search for a link having certified product reviews. Once we reach the certified review section, extraction is done page by page. We only fetch certified reviews and not all reviews so that spams and fake reviews are neglected. All the extracted reviews are stored in a text file. However all these reviews extracted from the web pages are in raw and unprocessed form. So proper formatting is done to remove the redundant data such as HTML tags, null sentences etc. All the extraction and formatting of the raw file is done in Python.

2. Sentiment Analysis:

We perform sentiment analysis, sentence by sentence. This sentence is POS tagged to identify what part-of-speech every word belongs to. The POS tagging is done using Stanford NLP parser. After POS tagging, we have to find the sentiment of the words in the sentence especially the adjectives and adverbs. The sentiment is every word can be found using SentiWordNet API. This API has assigned sentiments to nearly every word in the dictionary. Along with finding the sentiment we also find what aspect the sentence is about by comparing the sentence with a set of sub-keywords assigned to every keyword. Depending upon the adjectives or adverbs, the sentiment is analyzed and appropriate rating is given to the aspect ranging on a scale of 0 to 10 for every sentence. There are special cases that we have taken care of. For example, review containing false negatives like "the phone is not so good". Also there are some sentences describing the phone and not its aspects. We put such review sentences in general category. Apart from these, we have also fetched the timestamp of each review thus allowing to assign a higher weight to the latest reviews and a comparatively lower weight to obsolete or outdated reviews as the technology they mention about may not be of a higher priority on the current date. The final rating for a particular aspect is calculated as an average of ratings of every sentence having that aspect. The whole sentiment analysis is implemented using Java.

3. Summarization:

The final aspect rating obtained, is used along with aspect weight to obtain overall rating. If Wi and Ri is the aspect weight and aspect rating respectively for a particular aspect Ai, then Basically, we take weighted average to find the overall rating of the phone and scale it from 0 to 10. This aspect wise rating is expressed in the form of bar graphs and pie-charts. We have used Google Charts API to display the above mentioned charts. The overall rating is compared with the existing traditional NLP using Gauge meters which makes it visually pleasing. This summary is implemented on HTML pages with the help of CSS and JavaScript.

4. Screenshots:

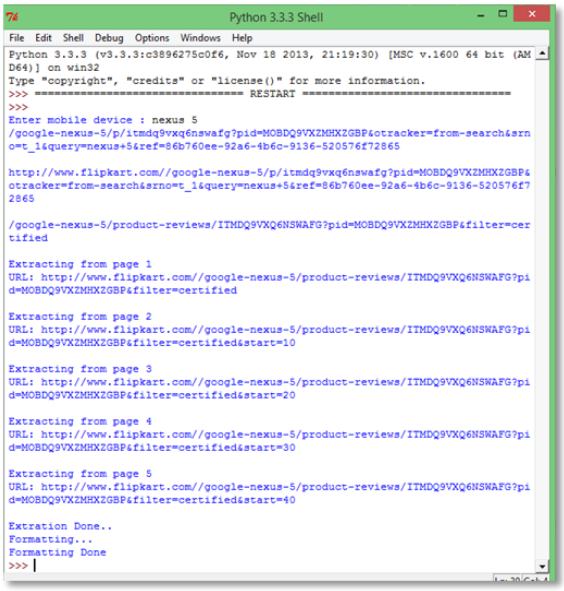

Figure 2: Python Shell Running Extraction Code

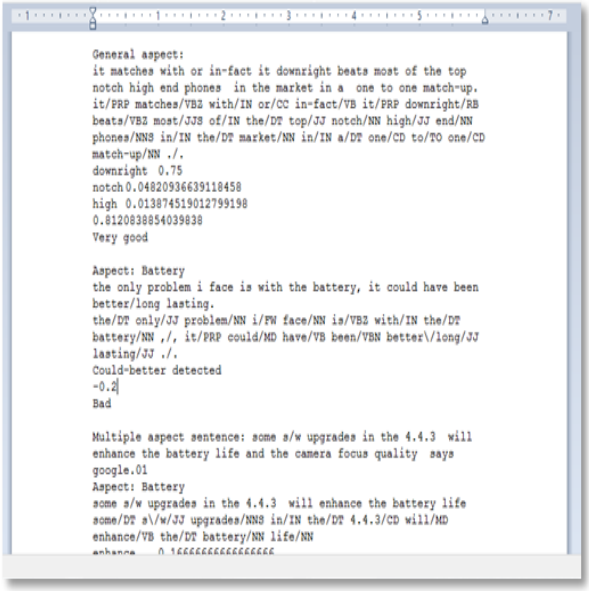

Figure 3: Log File Generated

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Figure 4. Overall Rating

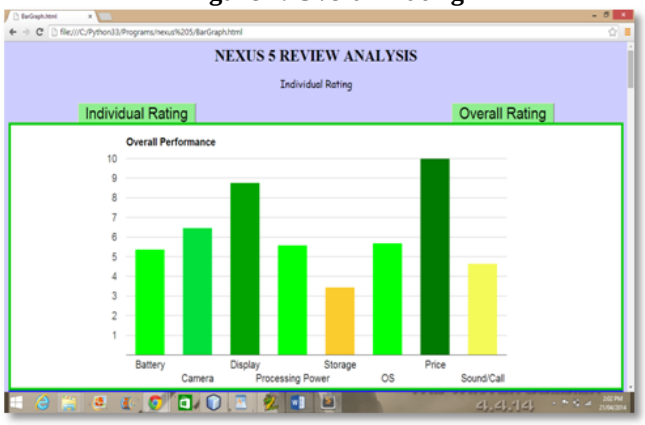

Figure 5. Overall Performance

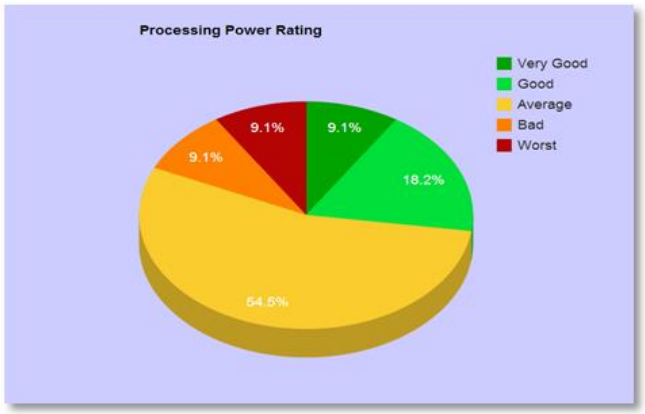

Figure 6. Individual Aspect Rating

V.CONCLUSION

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. For many of the mobile devices it is seen that the analysis using our algorithm which uses modified NLP is much more accurate and realistic as compared to analysis by Traditional NLP. Summarizing the reviews is not only useful to customers, but also crucial to product manufacturers.

ACKNOWLEDGMENT

We would like to thank our Project Guide Mrs. Sindhu Nair for her constant support and suggestion during the course of our BE project.

REFERENCES

1. Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. KDD'04, 2004.
2. B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In WWW '05, pages 342{351, 2005.
3. Hongning Wang, Yue Lu, Chengxiang Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach 2010.
4. Bing Liu. Sentiment Analysis and Subjectivity. To appear in Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.
5. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 201

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)