International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA

SAMIKSHA - Sentiment Based Product Review Analysis System

\(^a\)Aarti Potdar, \(^b\)Pranav Patil, \(^c\)Raunak Bagla, \(^d\)Rohitashwa Pandey, \(^e\)Prof. Nagesh Jadhav

\(^a,b,c,d\)Student, Department of Computer Engineering, MIT College of Engineering Pune, India
\(^e\)Department of Computer Engineering, MIT College of Engineering Pune, India

Abstract

As the technology is advancing, everything is getting more digital and automated. Today, we can buy everything on e-commerce websites and get it delivered without visiting the stores. From a screw driver to a high performance motor, there is hardly anything that we cannot find on e-commerce websites. But judging a product just by its pictures and reviews is a difficult and requires a thorough analysis. For a given product there are tens of thousands of reviews. To analyze all these reviews and to make a decision is a tiresome task. To make the task easier for the buyers we are proposing “Samiksha”, a review bot which will generate a factual summarization of all the user reviews. The proposed software will produce an average numerical rating of all the specific features of the particular product to help the buyer get a detailed overview of the product. Thus Samiksha will prove to be convenient medium of analyzing all the reviews and be beneficial to the buyers on e-commerce websites. Keeping that in mind, security and privacy of the users should not be hindered.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer-review under responsibility of organizing committee of the ICISP2015

Keywords: NLP; Parser; Preprocessing; Relation Extraction; Feature Extraction.

1. Introduction

Every potential customer goes online in search for that perfect product that matches all his or her requirements. Considering the scenario where the buyers’ requirements are met, what is the next step the person takes? He or she looks into its quality and performance. Today, the technology has taken over the market and e-commerce websites

* Corresponding author.
E-mail address: potdaraarti94@gmail.com

1877-0509 © 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer-review under responsibility of organizing committee of the ICISP2015
doi:10.1016/j.procs.2016.02.096
have become the major hub for buying and selling products. The average rating is an average rating for the product and user reviews. The average ratings allow the buyer to deduce from it, whether he

But then, how does one analyze the efficiency and usability of the product on an e-commerce website? The answer or she should consider the product as a worthwhile option or not. The major decision by the buyer is taken after reading the reviews and point of view of various other users which gives a detailed overview of the product. Reviews today are used in several researches for the purpose of mining and gaining relevant information which acts as a base for users analysis.

A plethora of reviews for a single product are available for the buyers' reference. Reading numerous reviews and making the right choice from the pool of similar products is a hectic task. It requires a detailed analysis of every review.

Our proposed model will help the buyers to reduce their work load of scrolling through and analyzing thousands of reviews. Samiksha, the review bot, will generate a factual summarization of reviews of any particular product, making the task of the buying convenient. Buyer doesn’t have to read, think and analyze; you just sit back and let Samiksha to get work done. In e-commerce environment reviews posted by various personnel are either by their hands-on experience or by their technological knowledge. Our research focuses on extraction of useful information from all the reviews of a particular product which will give the buyer its direct knowledge in a precise manner.

On these words, this research presents an approach for analyzing and summarizing multiple reviews. Samiksha, our review bot will have four main phases:

1. Fetching of user reviews of a particular product.
2. Preprocessing of fetched reviews using Natural Language Processing.
3. Analysis of preprocessed reviews based on various factors.
4. Generation of factual and summarized review of the product.

Fig.1: Main phases of Samiksha

Remaining sections of this paper is organized as follows: section II is literature survey related to our work. Section III discusses about the proposed model, section IV gives an insight on the future scope of our research.

2. Literature Survey

Research has been done in the field of sentiment analysis, software quality analysis and summarization of customer reviews. The work here is closely related to Jamroonsilp and Prompoon on analysis of software reviews and ranking them on based of their qualities. In another work by Hu and Lui, they summarized customer reviews by
polarities.

For the purpose of relation extraction the major focus is on extraction of sentimental words relating to a particular feature of the desired product. Using the sentiment words and the feature extraction, a relation is to be developed. In Natural Language processing by Stanford Parser\(^6\), a relation is represented as following:

\( \text{nsubj(<sentimental word>, <feature>)} \)

For example: This camera is very good.

\( \text{nsubj(good-5, camera-2)} \)

But, one of the drawbacks regarding the use of Stanford parser is the absence of code and the infrastructure which can directly support relation extraction. NLProcessor\(^7\) can be used as an alternative to Stanford Parser for creating parts-of-speech and relations.

In the process for classification of sentimental words, we need to analyze the sentimental base of the word related to a particular feature. Hearst\(^8\) and Sack\(^9\) in their works discussed classification of entire documents using cognitive linguistic models.

There are two main types of sentiment classification, term sentiment and sentence sentiment. Classification based on term sentiment is closely related to the work of Hu and Liu\(^4\). Hu and Liu constructed a set of positive and negative terms in order to classify the sentiment of a feature. In another process of classification by SentiWordNet\(^10\), the words are given ‘positive’ or ‘negative’ values based on its ‘objective’. We would combine both the approaches for this phase.

Sentence sentiment classification gives the actual semantic behavior of the two words which together gives the actual sentiment in a sentence. For example “not bad” signifies a positive sentiment of the sentence. Jamroonsilp and Prompoon\(^5\) have used the work of Ganapathibhota and Liu\(^11\), in which they introduced the set of following rules for classification of sentiments. Refer table 1.

<table>
<thead>
<tr>
<th>“Increasing comparative”</th>
<th>Negative Opinion</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Increasing comparative”</td>
<td>Positive Opinion</td>
</tr>
<tr>
<td>“Decreasing comparative”</td>
<td>Positive Opinion</td>
</tr>
<tr>
<td>“Decreasing comparative”</td>
<td>Negative Opinion</td>
</tr>
</tbody>
</table>

In our work, we will use these rules for sentiment classification. Summary generation is in form of text in most of the previous works. Here we will use numerical values instead for summarization. Frameworks\(^12,13\) have certain core facts and entities packaged in a template for a document. At the same time it also requires background knowledge to instantiate the template. Our work will not require any initialization and will be independent of the domain. Hu and Liu\(^4\), in their work summarized reviews and displayed all the positive as well as negative reviews for any particular feature of a product. Samiksha would generate numerical rating of specific features giving a factual summary of reviews related to a product.
3. Proposed Architecture

Figure 2 gives the architectural model of our proposed system “Samiksha”.

![Proposed Architecture Diagram]

The input to the system is the URL of the product whose reviews are to be analyzed. The output will be feature based numerical summary of the all the reviews fetched by the system.

The system works in 4 main steps: 1) Fetching of user reviews. 2) Preprocessing of fetched reviews. 3) Analysis of preprocessed reviews. 4) Generation of factual summarized review.

3.1 Fetching of user reviews

The main purpose of this phase to gather all the user reviews available for a particular product. The system will fetch the URL of the e-commerce website for the product desired by the user. It will then use the page source to extract the user reviews related to that product. XML, HTML and PHP are most widely used languages for creating the page source. Using the tags in XML and HTML, we can extract the user reviews for further analysis.

3.2 Preprocessing of Fetched Reviews

In this phase, the fetched reviews will be preprocessed and a relation between the feature and sentiment related to it is created. For this purpose, we use Stanford parser [6] which directly generates relations as discussed before. But some pronoun based relations in Stanford parser may not be clear. For example, a certain review regarding the camera feature of a phone:

“Talking about the camera, it is good”
Universal Dependencies generated by the parser:

```
advcl(good-8, Talking-1)
case(camera-4, about-2)
det(camera-4, the-3)
nmod(Talking-1, camera-4)
subj(good-8, it-6)
cop(good-8, is-7)
root(ROOT-0, good-8)
```

In this example the parser is not able to create a direct relation between the “camera” and the sentiment which is “good”. Here we need to pre-process the review and replace “it” with the feature itself that is “camera”. Now the sentence will be “Talking about the camera, camera is good”. Hence our final relation will be

Universal Dependencies generated by the parser:

```
advcl(good-8, Talking-1)
case(camera-4, about-2)
det(camera-4, the-3)
nmod(Talking-1, camera-4)
subj(good-8, camera-6)
cop(good-8, is-7)
root(ROOT-0, good-8)
```

After pre-processing, we can see that the direct relation between the “camera” and “good” can be established.

### 3.3 Analysis of pre-processed reviews

After the pre-processing phase where the reviews are converted into relation between the feature and sentiment, these relations need to be analyzed. We create a set of positive and negative sentiment words rating them between -1 to 1 based on the intensity of the word used. The ratings of each word will be defined with the help of SentiWordNet [10] and the objective related to it. For example refer figure 4

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Word</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Worst</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>Bad</td>
<td>-0.75</td>
</tr>
<tr>
<td>3</td>
<td>Average</td>
<td>+0.375</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
<td>+0.75</td>
</tr>
<tr>
<td>5</td>
<td>Excellent</td>
<td>+1</td>
</tr>
</tbody>
</table>

The graphical representation of the SentiWordNet table is represented by figure 3:

![Fig.3. SentiWordNet graphical representation.](image)

Now there may be certain cases where a positive sentiment word may be expressed using a negative sentiment word preceded by a negation or vice versa. For example “not good” actually gives a negative connotation to the
sentence. In such cases the value assigned to the term sentiment will be changed by a certain factor. A similar case has also been explained by Ganapathibhota and Liu [11]. There may be cases where certain sentiment words may not be a part of our word set. For such words we will refer Big Huge Thesaurus API [14] to generate its closest synonym which is part of our word set. There may be cases where certain sentiment words may not be a part of our word set. For such words we will refer Google API to generate its closest synonym which is part of our word set.

3.4 Generation of factual summarization review

In addition to the summary generated by Hu and Liu [4], we will generate a summary based on a numerical value for every specific feature of a product. In the previous step, we assigned values to the sentiment between -1 to 1. We will then calculate the average rating of a particular feature from all related sentiment mentioned in numerous reviews using the following formula.

\[
\text{Avg rating} = \frac{\sum_{i=1}^{n} X_i}{n}
\]

where, \(X_i\) is the sentiment rating of the word related to that feature for all reviews.

\(n = \text{number of reviews containing that feature.}\)

For example, Let us consider two reviews

Review 1: "the camera is good"
Review 2: "the phone has an excellent camera"

In the first review, the value attached to "good" will be 0.75 and in the second review the value attached to "excellent" is 1. Hence the average rating for the feature “Camera” will be:

\[
R_x = \frac{0.75 + 1}{2} = 0.875
\]

Where \(R_x = \text{Avg rating (Camera) on a scale of -1 to 1}\)

**Input:** Product reviews fetched from the website and stored in a database.

**Output:** Feature based factual ratings generated by the agent.

```plaintext
feature[x] : // All features
Y = 0; // summation of all the ratings of a feature on the scale of -1 to 1
Rx = 0; // average rating of a feature on a scale of -1 to 1
r = 0; // converted rating of a feature to a scale of 0 to 5
// Iteration for all specifications
For j 0 < m
// Iteration for all reviews
For i 0 < n
If feature found
create relation between feature and words
Y = SentiWordNet( );
Y+=Y;
Rx = Y/i;
end
r = (Rx + 1) * 2.5;
end
```
Now we convert this value from the scale of “-1 to 1” into a scale of “0 to 5”. The formula for the conversion is:

$$r = (R_x + 1)* 2.5$$

where, \( r \) = rating on a scale of 0 to 5

\( R_x \) = rating

Therefore, based on our above example of camera 0.875 will yield a value of 4.6875 to the nearest second decimal. In Table 4, we used our algorithm on reviews of a laptop to calculate average feedback of reviewers for its “graphics” performance. The detailed procedure is described in Table 3.

4. Conclusion

In this paper, we have discussed a set of methods in order to generate a highly detailed mathematical summarized value for all the specific features associated with a particular product using all the user reviews posted for that product. We believe that this research would be of great importance as people have highly inclined themselves towards shopping on e-commerce websites. In future, we plan to make our system a better learning agent based on its
experiences from the new words encountered. Along with that we plan to add a comparative base to the reviews such as “product A is better than product B”. We would also plan to change the input scenario where user need not provide the URL of a specific website to our system. Instead, the buyer can directly mention the product name and our system will generate results from all the major e-commerce websites. Along with this some kind of encryption should be added to ensure security and privacy of data.

References