Feature-Based Opinion Summarization for Arabic Reviews

Alaa M. El-Halees  
Faculty of Information Technology  
Islamic University of Gaza  
Gaza, Palestine  
alhalees@iugaza.edu.ps

Doaa Salah  
Faculty of Information Technology  
Islamic University of Gaza  
Gaza, Palestine  
do3a.it@gmail.com

Abstract— Opinion mining applications work with a large number of opinion holders. This means a summary of opinions is important so we can easily interpret holders’ opinions.

The aim of this paper is to provide a feature-based summarization for Arabic reviews. In our work, a system is proposed using Natural Language Processing (NLP) techniques, information extraction and sentiment lexicons. This provides users to access the opinions expressed in hundreds of reviews in a concise and useful manner.

We start with extracting feature for a specific domain, assigned sentiment classification to each feature, and then summarized the reviews. We conducted a set of experiments to evaluate our system using data corpus from the hotel domain. The accuracy for opinion mining we calculated using objective evaluation was 71.22%. We, also, applied subjective evaluation for the summary generation and it indicated that our system achieved a relevant measure of 73.23% accuracy for positive summary and 72.46% accuracy for a negative summary.

Keywords— Feature-based summarization; Feature extraction; Arabic Opinion Summarization; Sentiment Classification.

I. INTRODUCTION

Feature-based opinion mining is an important topic in Arabic. But, it only counts the quantity of opinions about a feature [1]. A better alternative is putting a summary for the positive and negative opinions.

Opinion mining applications work with a large number of opinion holders, this means some form of a summary opinion is needed. That is because simply using the ratings is not enough to get a description of what people think. Also, manually analyzing a huge quantity of data is hard and time-consuming [2]. The process of opinion mining would ease this task by providing summary of customer reviews regarding the features of product or service. In addition, opinion mining research mostly concentrates on extracting opinions from a large number of opinions’ holders. That is because getting an opinion from one holder is not sufficient [3].

There are two types of opinion summarization, one is based on features which are entitled feature-based opinion summary, and the other that does not rely on the presence of a feature and is called simply opinion summarization. In this paper, our work is based on feature-based summarization, which divides input text into aspects/features, and generates summary of each feature.

We developed a system using Natural Language Processing techniques (NLP), information extraction and sentiment lexicons generated by the data corpus to extract opinions on product features and summarize them. This provides user access to the opinion expressed in hundreds of reviews in a concise and useful manner.

Opinion summarization has the followings tasks: first, extracts the features of an object; second, identifies opinions that related to the object features in each review and finds the opinion orientation as positive or negative; and third, generates some sentences for each feature-opinion pair as the summary [4].

Therefore, in our work, we propose to develop a feature-based opinion summarization for Arabic reviews that start by identifying the features for the current domain then the summarization process is done to determine which features have satisfaction and dissatisfaction the customers.

The rest of the paper is structured as follows: the second section discusses the related works; the third section addresses our research methodology, the fourth section about experiments and results, while the fifth section implies the conclusion and future works of the paper.

II. RELATED WORKS

Although, there is no work in Arabic opinion summarization, there are some works in other languages mostly in English. In this section, we present the work done in feature-based opinion summarization.
Kim et al. in [4] presented a list of the most common approaches used for feature-based opinion summarization. The authors indicated that the process generally has three phases: The first phase is feature recognition, which is used to find important features in the reviews. The second phase is opinion orientation detection which is used to detect the opinion’s polarity, as positive or negative, on each selected feature. The third phase is the generation of features’ summary which is to generate summary to display results in a simple way.

The main contribution of Kamal in [5] is feature-based opinion mining from review at the finer-grained stage and classifies a sentence as subjective or objective. Then, he extracted feature-opinion pairs, and finally classify reviews. In addition, the author gave a summary for the customer review to represent features and their opinions in graphical form. The extracted knowledge and their scores are stored as JSON file which is used for graphical presentation. The visualizing of the summary can be presented for single or multiple reviews. Maharani et al. in [6] presented a summary of the last research in the aspect-based opinion summarization. They, also, described among different summary generation techniques, text and graphical opinion mining summarization, including the effectiveness of summarization in different fields. Finally, they introduced some research challenges in the area. Hu et al. in [7] considered ways of creating a feature-based summary for customers’ reviews. They proposed techniques to perform these tasks in three steps: defining features from reviews, appoint a polarity for each feature in reviews and finally, generating a summary for each feature in reviews. Zhuang et al. in [8] extracted features from movie reviews using another technique. They constructed a feature list of the cast of a movie by integrating the full cast of each reviewed movie. They used regular expressions to find if a word in current review is matched with another in the feature list. They used WordNet ontology to define the words polarity. They also used statistics to generate summarization. Lu et al. in [9] proposed an approach to summarize features using the following steps: step one to extract the main features; step two to predict a score for each feature; step three to extract a summary from the reviews. Finally, Eirinaki et al. in [10] identified the most representative features of each customer review. Then, in the next step, they assigned a rate for each feature. The research presents an algorithm to identify the semantic orientation of the review that can identify the polarity of the feature. In addition, the proposed algorithm produced the summary of the sentiments of the most important features.

III. RESEARCH METHODOLOGY

The methodology, which we followed in our research paper is divided into the following steps:

A. Customer Reviews

For our work, hotel corpus from [11] has been chosen, which consists of 2860 reviews of an equal number of positive and negative users’ reviews.

B. Preprocessing

For our collected dataset, we did the following preprocessing steps:
- Remove irrelevant data: this process applied by removing punctuation, numbers and non-Arabic words that are not useful for our method.
- Tokenization: the process applied by identifying the smallest units in the review content.
- Part-of-speech tagging: It is an important step as it obtains general patterns for the Arabic language. In our work, we use Stanford-POS tagger from [12]. The tagger parses each sentence and assigns a tag for each word (e.g. the word is a verb, noun … etc).
- Stemming process: The stemming process is important for our method since it brings together words based on their lexicon-semantic similarity. We used root stemming method for the following purpose: firstly, to produce better matching of features in the search engine. Secondly, the Arabic Sentiment lexicon contains only root words. We used the system described by [13].

C. Extract Product Features

In order to extract features, we used a modified version of the method proposed by [14]. The input of the feature extraction process is text and the output are a set of extracted features. The extraction process consists of three main steps:

Step 1: Identifying frequent nouns

Works such as [15] and [14] on opinion mining found that features in the text can be extracted from the most frequent nouns. For that, we counted the number of nouns in each review. A noun is frequent if it has a number more than a certain threshold. The selected nouns are collected in a list as candidates’ features list.

Step 2: Identifying relevant nouns

We collected the adjective which is adjacent to the nouns in the candidates’ features list. In this step, we discarded prepositions and irrelevant words and stop word for more accurate results.

Step 3: Remove unrelated nouns:

We used PMI-IR measure from [16] as in equation (1) to discard irrelevant nouns. PMI queries a Web search engine with the words under analysis and words represent positive or negative sentiment and
computing the number of returned hits (matching documents).
\[ PMI(t_1, t_2) = \frac{\log_2 (\text{Hits}(t_1) \cap \text{Hits}(t_2))}{\text{Hits}(t_1) \times \text{Hits}(t_2)} \] (1)

Where \( \text{Hits} (t_i) \) is the number of pages containing term \( t_i \), \( \text{Hits} (t_2) \) is the number of pages containing term \( t_2 \) and \( \text{Hits} (t_1 \cap t_2) \) is the number of pages with both terms. Using quires in Google, \( t_1 \) is the tested noun and \( t_2 \) is the entity domain. Using these results, we can eliminate unrelated nouns which have measure less than certain thresholds.

A. Sentiment Prediction

We need to know if the feature opinion is positive or negative. So, we used ArSenL lexicon from [17] to determine the polarity of these reviews. Lexicon is a list of Arabic terms and their associations with the sentiment. The association is shown in the form of a number (sentiment score) between -5.0 and 5.0. In general, if the positive score is greater than the negative score then it considers the opinion review as positive polarity otherwise negative polarity.

B. Summarization

In this step we generate features’ summary using the following steps:
1. For the selected features in features generated list, using the orientated of each feature, we computed a number of positive and negative opinions for each feature.
2. The extracted summary sentences ranked by three factors to extract the important of the sentence which are:

   The first factor is the strength of sentiment words in the sentence where opinions word varies in their intensity as Positive, Mildly Positive, or Strongly Positive. We used the same lexicon in the previous step to determine the strength of a sentence by summing the score of words of each sentence. The higher positive score considered the more strongly positive review.

   The second factor is the Term Frequency–Inverse Document Frequency (TF-IDF) value of the words in the sentence. TF-IDF from [18] is a weighting algorithm as given in equation (2), where a given term in a sentence weight directly proportional to the term frequency and inversely proportional to the document frequency. Then, this weight is used to inspect the content of a given sentence and differentiate between this sentence and another sentence. The result of TF-IDF is a term-score ranked list; the higher is the score the more related is the term for the sentence.

\[ tfidf = \sum_{w \in \text{term}} \frac{\text{strength}(w) \times \text{tf idf}(S)}{||S||} \] (2)

In the third factor, we used similarity measure where a number of users used the sentence or similar to it. So, if the reviewed sentence mentioned by more than one reviewer, it means that is important. In our work, we considered two sentences similar if they have some threshold similarity. To compute cosine similarity, we need two document vectors, where \( A \) and \( B \) are vectors represent each unique term with an index, and the value at that index is some measure of the importance of the term to the document as given in equation (3).

\[ \text{Similarity}(A, B) = \frac{\sum_{w \in \text{term}} A_B}{\sqrt{\sum_{w \in \text{term}} A^2} \sqrt{\sum_{w \in \text{term}} B^2}} \] (3)

Using the three factors, we can compute the importance of a sentence using a linear equation using equation (4).

\[ \text{Weight}(S) = \text{strength}(S) \times \text{tf idf}(S) \times \text{similarity}(S) \] (4)

C. Evaluation

To evaluate our proposed systems which based on evaluating the summary, we used two types of evaluation, subjective which is based on human evaluation and objective which is based on metrics.

1. Subjective Evaluation

Summarization systems have often been evaluated by comparing it with human-generated reference summaries. In some cases, the human summarizer constructs a summary by selecting relevant sentences from the original document, as in our case; the summaries are hand-written from scratch.

In our work, we evaluated our system using ROUGH. ROUGH is a number of metrics which is usually used to evaluate automatic summarization [19]. It compares the system summary with the human expert summary.

In our work, we evaluated our results with the help of a human judge. Therefore, we picked around 50 reviews for two product features from hotel corpus. Then we asked the human judge to give each sentence a positive/negative polarity. Then, we chose the best 5 sentences that express positive polarity, and the same goes for the negative. Then the final evaluation is measured by comparing the sentences chosen by our system with the manually chose by the judge.

2. Objective Evaluation

The measures evaluate the performance of classification is called confusion matrix, which is
also called a performance vector that contains information about realistic and predicted classifications. Precision and recall are basic measures used in evaluating our approach. The measures evaluate the performance of classification is confusion matrix. This evaluation depends on the comparison of real results and the effective results of the assessed system. Precision and recall are defined as follows: Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the dataset. On the other hand, precision is measured by the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved [20]. Accuracy is the proportion of the total number of predictions that were correct. F-Measure is the ratio between recall and precision measurements.

IV. EXPERIMENTS AND RESULTS

We conducted a set of experiments to get our results; next section gives the experiments and the results:

A. Collected data set

In order to do our experiments, we used a hotel review dataset from [11]. Table 1 gives snapshot of collected data set.

Table 1: A snapshot of collected reviews

<table>
<thead>
<tr>
<th>No.</th>
<th>Reviews</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Negative</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Positive</td>
</tr>
</tbody>
</table>

B. Feature Extraction

In order to calculate the accuracy of this step, relevant features were extracted manually from the reviews and stored in a file. Then, another list was generated from of proposed system. The accuracy is calculated by comparing the two lists.

In our experiment, the number of manually extracted features was 27 features, and the number of features extracted by the system was 22 features, the two lists overlapped between 22 features from 27 features. Results are presented in table 2; it achieved 81.48% accuracy.

In the first step, we extracted only the 35% most frequent features from the reviews. It is obvious that the results are not acceptable. In step two, which used the adjectives adjacent to frequent features identified in the first step, significantly increased the percentage to 81.48%.

Table 2: Accuracy rate for feature extraction step

<table>
<thead>
<tr>
<th>Number of features extracted by system</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>81.48%</td>
</tr>
</tbody>
</table>

C. Lexicon Based Sentiment Classification

An Arabic sentiment dictionary is taken to retrieve the polarity of each review contains the feature. In this method, a list of positive and negative words is required attached with the sentiment score assigned to each of the words. Then, it sums up the total number of positive and negative polarity of all the words of the sentence; if the total number of positive polarities is greater than total negative polarity value, then the positive polarity assigned to the document and vice versa. As seen in table 3, an example of the extracted (polarity, feature) for the review, some of the reviews were classified as positive reviews, but it should have a negative sentiment and vice versa. The results for this method for hotel corpus shown in table 4, the f-measure of 70.08%.

Table 3: Example of system extracted (Feature, Polarity) in reviews

<table>
<thead>
<tr>
<th>Review</th>
<th>Feature</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>فندق موفنبيك ومساكن برج هاجر مكة الموقع قريب من الحرم والنظافة جيدة</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>المكان رائع ولقد سعدت بالإقامة فيه وهو في عالية النظافة والطابعية كما أن الموانئ وتعامله جيد وودوين من هو جيد رأى القمة اختيارات الموفين ودودين وحاولوا من هو في غاية الفخامة والنظافة كما أن الموفين ودودين</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>المزيد من النظافة بالظافة</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>الخدمة. الخدمة لا تستجيب من أول مرة لا أعلم بما</td>
<td>Negative</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Performance of classifying opinions for hotel dataset

<table>
<thead>
<tr>
<th>Domain</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotels</td>
<td>71.22%</td>
<td>67.48%</td>
<td>72.90%</td>
<td>70.08%</td>
</tr>
</tbody>
</table>
D. Subjective evaluation:

Sample results were given to group human experts consist of 10 people to assess them. The raters worked independently to undertake two tasks: in the first task, identify the features in each sentence and give the sentence a Positive/ Negative polarity. After that, we asked them to choose the best 10 reviews that express positive reviews for each feature, and the worst 10 reviews that express the negative reviews for each feature.

We calculated the accuracy for the summary generation step, by calculating the matches sentences that selected by the human judge, and the sentences that generated by the system. The average accuracy for the positive extracted summary for both features was 73.23%; the average accuracy for the negative extracted summary was 72.46%. The overall average is 72.84%.

We noticed that the summary result depends only on the strength of the sentiment, which in our work were the overall polarity of the sentence, gives us a approximately 60% of the sentences that match the extracted human sentences, where after applying the TF-IDF factor the result enhanced by 67%. Finally, when we calculate the cosine similarity for the sentences that have the same features, the average accuracy was increased to 72.84%.

V. CONCLUSION

There are no previous efforts existed in extracting feature-based opinion summary from Arabic. In this paper, we proposed a system that extracts features from reviews and summarizes customer opinions. To do that, we formulated a dataset of reviews to assess the approach, then the results compared to human subjects’ opinion results. The accuracy for opinion mining is 71.22% using subjective evaluation. The results indicated that our system achieves highly relevant measure with 73.23% accuracy for positive summary, and 72.46% accuracy for negative summary, with an overall average is 72.84%.

In the future works, we plan to handle the corpus more efficiently by dealing with the reviews that have more than 300 words and perform a sentence splitting in other efficient ways to a better result. We aim to do other summarization methods for more accurate results as well as incorporate our method with supervised classification approaches for sentiment classification. We can use the same process described in our work for feature-based opinion summarization in different corpora, covering other kinds of domains (books, physician’s offices, computers…. etc). We can use more than one public Arabic lexicon to improve the performance of our sentiment classification process.

REFERENCES