Towards an efficient opinion measurement in Arabic comments

Walid Cherifa\textsuperscript{a}, Abdellah Madani\textsuperscript{b}, Mohamed Kissi\textsuperscript{a,∗}

\textsuperscript{a}Laboratory LIMA, Department of Computer Science, Faculty of Sciences. B.P. 20, 24000, El Jadida, Morocco
\textsuperscript{b}Laboratory LAROSERI, Department of Computer Science Faculty of Sciences. B.P. 20, 24000, El Jadida, Morocco

Abstract

Arabic language is the fifth most widely used language on Internet\textsuperscript{1}. Every day, a huge volume of Arabic comments and reviews have been generated concerning different aspects of our life. In the light of the scarcity of systems to analyze this data, we propose in this paper a complete approach in order to identify and classify author’s opinions. It is conducted using a dataset consisting of 625 Arabic reviews and comments collected from Trip Advisor website which fall into five classes. We started first by choosing the appropriate stemming algorithm, and used it to introduce our new mathematical approach to formulate opinions. The classification which based on Support Vector Machines derived a scheme, which, in turn, often needed to be refined as some reviews remained unclassified. We opted then for a new similarity approach based on k-nearest neighbors and feature weighting to classify them. Finally, we compared our global approach with two recent works in terms of accuracy. The results obtained have met our expectations.

Keywords: Automatic language processing; Arabic text; Opinion mining; Low-level light-stemming; Support vector machines; K-nearest neighbors; Similarity measure;

1. Introduction

Arabic is one of the major languages of the United Nations\textsuperscript{2}. It is the mother language of more than 330 million people in more than 22 countries\textsuperscript{3}. It is also the language of Holy Quran.

∗ Corresponding author. Tel.: +212 5 23 34 23 25; fax: +212 5 23 34 21 87.
E-mail address: kissim@gmail.com; kissi.m@ucd.ac.ma
Recent years have been witnessing an explosive increase in the number of Arabic internet users. A tremendous number of opinions is shared every day concerning different daily aspects and events.

In this regard, opinion mining is the computational study of people’s opinions toward entities, individuals, issues, events and topics. However, distilling the information contained in Arabic texts is not always available. Only few researches have been done for Arabic language. We can group them into four main categories: keyword spotting, lexical affinity, statistical methods, and concept-level techniques.

Unfortunately, many typical challenges were noticed while stemming Arabic opinionated texts. Therefore, our work will primarily focus on resolving the vocabulary mismatch problem. We introduce for that a new text treatment method.

After that, we define our modeling function that will represent each given comment by a vector. Then, we shall proceed to the classification with support vector machines, which is a relatively recent supervised learning technique applicable to both classification and regression.

After the step of classification, the remaining uncertain comments will be sent for a new post-classification based on k-nearest neighbors and a distance that we will define in the coming chapters. Finally, the proposed approach outperforms two main methods for opinion recognition: SentiWordNet and SentiStrength.

The rest of this study is organized as follows: In section 2, we shall summarize our modeling method, while section 3 details the classification of our corpora. In section 4, we present a further optimization that extends our approach before concluding with results comparison and a brief presentation of our future works in the last part.

2. Background

2.1. Corpora

The dataset used in this study contains 625 comments on hotels collected from TripAdvisor website and classified into five categories: “excelente” (excellent); “buena” (very good); “normal” (middling); “débil” (weak) and “terrible” (horrible). We started by eliminating insignificant comments like “HMMMM.....” which can’t explicitly translate author’s opinion. Comments containing repetition of letters such as “_repeat_repeat?” or punctuations such as “!” were distinguished as they may increase the level of subjectivity. As publishing a comment can be subject to errors, we analyzed thoroughly the collected corpora and found that some of them can be matched in different classes. If the comment appeared in one class 80% more than in a distinct class, we attributed to it the label of the class with the dominant frequency. Finally, we gathered all tokens figuring in the remaining comments, and defined them as our variables for coming treatments.

2.2. Formulation

Stemming is a crucial pre-processing step in opinion mining as well as a very common requirement of Natural Language processing functions. Arabic stemming algorithms can be ranked, according to three categories: as root-based approach (ex. Khoja); stem-based approach (ex. Larkey); and statistical approach (ex. N-Garm).

Light-stemming is the process of reducing inflected words to their stems by removing prefixes and suffixes, while stemming reduces these stems to their original roots. However, affixes are not always easily deciphered; we are subject to two kinds of errors: under-stemming and over-stemming. In what follows, we will use a rules-based stemmer we introduced in our previous works that corrects these mistakes.

Although stemming reduce significantly text size; in contrast, light stemming preserves more meaning of the Arabic word processed. However, as our work is purely on semantics, we need then to consider the complete meaning. Thus, we introduce the low level light stemming that keeps particles (“من أفضل” one of the best / “أفضل” best), punctuation (“،” wonderful ! / “،” wonderful) and superlative adjectives (“الأفضل” the best / “أفضل” best) as described in the following table:
Table 1 shows that low level stemming is the most compatible approach with our study. We obtain a group of 665 low level stems.

Some of these low level stems intensify the opinion like: “قَرِينَ” (beautiful) /”قَرِينَ” (very beautiful) and “كَرِينَ” (I liked it) /”كَرِينَ” (how I liked it). These intensifiers will be represented by constants $c_i$ to be multiplied by next or previous word. We form then a list of pre-intensifiers (“قَرِينَ”: very) that affect words coming after, a second list of post-intensifiers (“قَرِينَ”: a lot) affecting previous words, and a third one of those that can affect both sides.

The second faced mistake concerns contradicted combined sentences such as “قدِينَ” (I disliked the service). While the majority of studies on Arabic opinion mining often employs supervised learning techniques; other works resort to fuzzy logic in a try to deal with the previous mistake.

As for intensifiers, we multiply each part of the sentence by a constant while giving priority to the last one. For the two previous rules, stopwords and punctuation are kept in order to reinitialize the multiplication coefficient.

The time also has its impact on an opinion, temporal verbs and conjunctions are used to lower opinions level (“كَرِينَ” : the hotel used to be beautiful).

Finally, for a given comment $comm$, we formulate the representing vector by:

$$f(comm) = \sum_{i=1}^{\text{665}} \frac{Cf_i}{\sum_{i=1}^{\text{665}} Cf_i} \cdot e_i$$  

Where $Cf_i$ is the product of constants $c_i$ affecting the variable $X_i$.

The following figure shows an example of our modeling approach and resumes our global corpora preparation:

Figure 1: (a) Modeling example; (b) Corpora preparation.
3. Classification

Since opinion mining stands at the cross junction to information retrieval and machine learning, we will enhance the modeling approach seen above by the suitable classification models in this part.

3.1. Support Vector Machines

SVM were introduced by Vapnik\textsuperscript{18} as a class of supervised machine learning techniques. They are based on two key ideas: the notion of maximum margin and the concept of kernel function. In linear classification, SVM create a hyper plane that separates the data into two sets with the maximum margin\textsuperscript{19}. For the cases where the data are not linearly separable, SVM map the data representation space into an area of larger dimension in which it is probable that there is a linear separator (Figure 2 (b)).

Mathematically, in linear case, SVM learn the sign function $f(x) = \text{sign}(\omega \cdot x + b)$ where $\omega$ is a weighted vector in $\mathbb{R}^n$. They find the hyper plane $y = \omega \cdot x + b$ by separating the space into two half spaces with the maximum-margin. Elements on both sides of the hyper plan are well classified. Uncertain ones are those remaining in the middle as shown in the following figure:

![Figure 2: (a) SVM classification (b) nonlinear SVM mapping.](image)

The figure above highlights the fact that more the element is near the center $\omega \cdot x + b = 0$, less it is well classified. Hence the need for other considerations to classify it.

The SVM have been successfully used on Arabic opinion mining\textsuperscript{20, 21, 22} and they derived better accuracy results than other machine learning techniques.

3.2. Computation

Results accuracy is the result of a monitoring step for approximating the reality with reliable data. It is defined as the difference between the desired value and the value obtained by the considered model. The overall performance was evaluated in terms of root mean-square error (RMSE) and Mean Absolute Error (MAE) (2):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_{i}^{d} - y_{i}^{c})^2}{N}} \quad \text{; MAE} = \frac{\sum_{i=1}^{N} |y_{i}^{d} - y_{i}^{c}|}{N}$$

Where $y_{i}^{d}$ is the desired output, $y_{i}^{c}$ the calculated output and $N$ is the total number of comments.

As the results of SVM approach lie largely on the choice of a kernel\textsuperscript{23}, we mention here that we have compared main kernel functions:
Table 2: SVM kernel functions comparison

<table>
<thead>
<tr>
<th>Kernel</th>
<th>dot</th>
<th>radial</th>
<th>Epahnenikov</th>
<th>Polynomial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>$x \cdot y$</td>
<td>$\exp(-\gamma |x - y|^2)$</td>
<td>$(3/4)(1 - u^2)$</td>
<td>$(x \cdot y + 1)^d$</td>
</tr>
<tr>
<td>MAE</td>
<td>0.20</td>
<td>0.35</td>
<td>0.36</td>
<td>0.83</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.30</td>
<td>0.56</td>
<td>0.59</td>
<td>1.08</td>
</tr>
<tr>
<td>Max Margin</td>
<td>0.72</td>
<td>0.83</td>
<td>0.86</td>
<td>3.6</td>
</tr>
<tr>
<td>Nb Margin$\geq 0.4$</td>
<td>35</td>
<td>198</td>
<td>211</td>
<td>369</td>
</tr>
</tbody>
</table>

Results in table 2 allowed us to choose the dot product kernel for coming steps as it has a the minimal values deviation and the smallest number of uncertain classified comments.

3.3. Variables contribution

One of the main purposes of SVM classification in general, and our study in particular is the determination of the descriptors influencing opinion recognition. The evaluation of their relevance proved quite interesting and useful. In this part, we chose to estimate their relative contribution as follows:

The contribution of each descriptor $i$ to the classification ability is measured by removing its corresponding values from database. Then the SVM calculate the output as usual. This process is reiterated for each descriptor. Finally, the contribution $\text{Cont}_i$ of descriptor $i$ is given by:

$$\text{Cont}_i = 100 \frac{\Delta m_i}{\sum \Delta m_j}$$

Where $\Delta m_i$ is the deviation absolute value between the observed and the estimated values for all compounds. Table 3 summarizes the ten most influencing variables for each class of comments:

Table 3: Most influencing variables on classification.

<table>
<thead>
<tr>
<th>Variables</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>10th</th>
</tr>
</thead>
<tbody>
<tr>
<td>متزاژ</td>
<td>(3,9%)</td>
<td>متزاژ</td>
<td>(3,0%)</td>
<td>متزاژ</td>
<td>(2,7%)</td>
<td>متزاژ</td>
<td>(2,3%)</td>
<td>متزاژ</td>
<td>(2,0%)</td>
<td>متزاژ</td>
</tr>
<tr>
<td>جيد جدا</td>
<td>(3,3%)</td>
<td>جيد جدا</td>
<td>(3,0%)</td>
<td>جيد جدا</td>
<td>(2,7%)</td>
<td>جيد جدا</td>
<td>(2,6%)</td>
<td>جيد جدا</td>
<td>(2,0%)</td>
<td>جيد جدا</td>
</tr>
<tr>
<td>متوسط</td>
<td>(3,3%)</td>
<td>متوسط</td>
<td>(3,0%)</td>
<td>متوسط</td>
<td>(2,6%)</td>
<td>متوسط</td>
<td>(2,1%)</td>
<td>متوسط</td>
<td>(1,8%)</td>
<td>متوسط</td>
</tr>
<tr>
<td>ضعيف</td>
<td>(3,3%)</td>
<td>ضعيف</td>
<td>(3,0%)</td>
<td>ضعيف</td>
<td>(2,7%)</td>
<td>ضعيف</td>
<td>(2,3%)</td>
<td>ضعيف</td>
<td>(2,3%)</td>
<td>ضعيف</td>
</tr>
<tr>
<td>مروع</td>
<td>(3,8%)</td>
<td>مروع</td>
<td>(3,8%)</td>
<td>مروع</td>
<td>(2,7%)</td>
<td>مروع</td>
<td>(2,1%)</td>
<td>مروع</td>
<td>(2,0%)</td>
<td>مروع</td>
</tr>
</tbody>
</table>

This table presents lists of most influencing variables on the classification. We note that some of them monitor different classes. A scalability of their contribution will be considered in next chapter.

4. Optimization

In this section, we will consider comments that gave, by SVM, values 0.4 far from their class as uncertain in the classification process. We will try here to optimize their classification according to different features:
4.1. K-Nearest Neighbors

K-Nearest Neighbor is a potent method for opinion mining\textsuperscript{28}. It compares a given review vector to each labeled vector in the training set. Then its $k$ nearest neighbors are determined by measuring similarity which may be computed by, for example, Minkowski, Hamming or the Euclidean distance\textsuperscript{29}.

In our proposed method, given an uncertain classified comment $Comm$, the system has to find the $k$ nearest neighbors among adjacent classes which are classified in the previous phase ($k$ is fixed priori $=5$). As we achieved the contribution of most salient variables, we can use these weights to classify $Comm$ the most relevant instead of taking commonly terms frequency and presence. The similarity can then be measured by:

$$Sim = \sum_{i=1}^{665} \Delta m_i \left| Cont(X_i) \right|$$

(4)

Where $\Delta m_i$ is the difference between both coefficients of variable $x_i$ in the two compared comments and $Cont(X_i)$ is the contribution of the variable $X_i$ into the classification. We can now clearly distinguish between similarity and relevancy.

4.2. Global approach

The following figure shows our main algorithm that classifies a given comment:

![Figure 3: Global approach Flowchart.](image)

The figure above illustrates the integration of our similarity measurement into k-nearest neighbors to improve first classification results. Thus, we have completed an Arabic opinion mining model which deals with Arabic morphological complexity on the one hand and the comments syntax on the other hand.

4.3. Evaluation

In this section, we will compare our work with the two popular approaches: SentiStrength\textsuperscript{30} which is based on dominant polarities, and SentiWordNet\textsuperscript{31} which uses synsets of WordNet to assign sentiment scores to the comment. For this, we will need three evaluation measures: Recall, Precision and F-measure $FM$\textsuperscript{32}. 


Where $a$ is the number of comments correctly classified, $b$ : the number of false positives and $c$ is the number of false negatives. The overall results are shown in the following table in term of F-measure:

$$Recall = \frac{a}{a + c}; \quad Precision = \frac{a}{a + b}; \quad FM = \frac{2 \times Recall \times Precision}{Recall + Precision}$$ (5)

<table>
<thead>
<tr>
<th>Approach</th>
<th>SentiWordNet</th>
<th>SentiStrength</th>
<th>SVM</th>
<th>Hybrid SVM kNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ممتاز</td>
<td>0.74</td>
<td>0.79</td>
<td>0.86</td>
<td>0.98</td>
</tr>
<tr>
<td>جيد جدا</td>
<td>0.64</td>
<td>0.75</td>
<td>0.82</td>
<td>0.97</td>
</tr>
<tr>
<td>متوسط</td>
<td>0.47</td>
<td>0.68</td>
<td>0.87</td>
<td>0.97</td>
</tr>
<tr>
<td>ضعيف</td>
<td>0.69</td>
<td>0.69</td>
<td>0.87</td>
<td>0.98</td>
</tr>
<tr>
<td>مروع</td>
<td>0.70</td>
<td>0.81</td>
<td>0.86</td>
<td>0.97</td>
</tr>
<tr>
<td>Average</td>
<td>0.65</td>
<td>0.77</td>
<td>0.85</td>
<td>0.97</td>
</tr>
</tbody>
</table>

As we can see in the table above, the modeling approach combined with kNN provides the best result. In addition, the 6 comments not classified even after optimization lead us to analyze the reasons that can be considered in our future works. We found that they contain antiphrasis, and can be interpreted in different ways: (example: “كان فندق حليوب للأماكن وسمع نحوم”: it was a disappointing hotel and seven stars). Hence the need to an eventual use of fuzzy logic to solve this kind of mistakes by transforming definite values into grades of opinions.

Overall, the global presented approach offers satisfying results and remains a promising field of investigation.

5. Conclusion and perspectives

The research in opinion mining is expanded recently to cover new aspects and languages. In this paper, we conducted a study to build up an efficient Arabic opinion recognizer. We started first by introducing low level light stemming and compared it with the two popular treatments: stemming and light stemming. Then, we collected decisive rules in author’s comments to correctly model each comment by an n-dimensional vector. We used SVM to classify our dataset into the five considered classes. The contribution of our most influencing low level stems was also evaluated. To further optimize this process, we based on k-NN and most influencing low level stems to classify uncertain vectors and obtain finally satisfying results. As future works, to alleviate comments interpretation problem, we intend to incorporate fuzzy logic. In fact, the Arabic language is not a single linguistic variety; rather, it is a collection of different dialects, we can then take into consideration in our future analysis local languages to obtain reliable and useful information through social networks and other shared data. Finally, this work has opened the door to further investigations as it is yet far from being exhaustively explored.

References