A Lexicon-based Approach to Build Service Provider Reputation from Arabic Tweets in Twitter

Haifa Al-Hussaini and Hmood Al-Dossari
Information Systems Department, College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia

***Abstract*—Nowadays Social media has become a popular communication tool among Internet users. Many users share opinions and experiences on different service providers everyday through the social media platforms. Thus, these platforms become valuable sources of data which can be exploited and used efficiently to support decision-making. However, finding and monitoring customers' opinions on the social media is difficult task due to the fast growth of the content. This work focus on using Twitter for the task of building service providers' reputation. Particularly, service provider's reputation is calculated from the collected Saudi tweets in Twitter. To do so, a Saudi dialect lexicon has been developed as a basic component for sentiment polarity to classify words extracted from Twitter into either a positive or negative word. Then, beta probability density functions have been used to combine feedback from the lexicon to derive reputation scores. Experimental evaluations show that the proposed approach were consistent with the results of Qaym, a website that calculates restaurants' rankings based on consumer ratings and comments.**

***Keywords*—Reputation; Sentiment Analysis; Arabic Language; Saudi Dialect; Social Media**

I. Introduction

Nowadays Social media has become a popular communication tool among Internet users. Many users share opinions on different service providers everyday through the social media platforms such as Twitter [a] and Facebook [b]. These platforms become valuable sources of data that can be exploited and used efficiently to support decision-making. For example, it is worthy for a customer who wants to buy a product to search information on the social media trying to find other consumers' opinions on the product.

However, finding and monitoring customers' opinions on the social media is difficult task due to the fast growth of the content. A study conducted in [1] shows that 85% of residents of Saudi Arabia use social media and the highest rates of Twitter usage are in the Middle East (based on a survey of more than 152,000 internet users across 31 markets). Figure 1 illustrates percentage of total users who subscribed in Twitter per country in 2015. It is clear that Saudi Arabia has the highest Twitter users by 53%.

The aim of this study is to develop a reputation approach based on sentiment analysis of Arabic tweets in Twitter. Such an approach is important because the huge amounts of valuable information contained in the social media cannot be readily analyzed manually. In this paper, a lexicon-based approach has been proposed to build service providers' reputation from social media. More specifically, the Beta reputation system [2] has been adapted to calculate service provider's reputation from tweets in Twitter. To do so, a Saudi dialect lexicon has been developed as a basic component for sentiment polarity. The proposed lexicon is used to classify words extracted from Twitter into either a positive or negative word. Then, the beta probability density functions have been used to combine feedback from the lexicon to derive reputation scores. Experimental evaluations show that the proposed approach were consistent with the results of Qaym, a website that calculates restaurants' rankings based on consumer ratings and comments.

The remainder of the paper is organized as follows. A review of the literature is presented in Section II. A Saudi dialect lexicon is introduced in Section III. The proposed lexicon-based approach for reputation calculation using Arabic tweets in Twitter is explained in Section IV. Experimental results are reported in Section V and discuss some relevant issues in Section VI. Finally, Section VII concludes the paper.

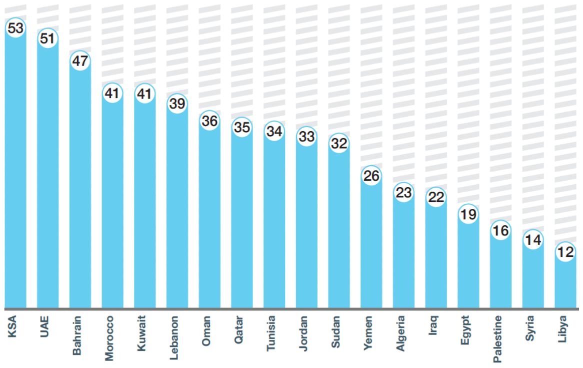


Fig. 1: Arabic subscribers in Twitter per country [7]

II. Literature Review

A Reputation system provides a promising way for building trust between service providers and consumers. Most reputation systems gather user feedback, such as ratings or reviews, then aggregate them into a single value to represent a reputation score [8], [9]. This type of system is called a feedback-based reputation system [4]. One main obstacle for feedback-based reputation models is sparsity of ratings [2], [3], [19]. That is, there is insufficient data to build the reputation score. This is because usually users have no direct incentive

[a] https://twitter.com

[b] https://www.facebook.com

for providing ratings, thus they abstain to leave their ratings after service usage [2].

SocialTrust is a proposed reputation system that has three chief components. First, social networks which is a network of friends and partners. Consequently, each node preserves two lists of mutual-trusted nodes. Each node needs to keep its friendship with other by producing no harm to them and like to have more partners to gain more benefits. Second, a trivial reliable server selection is made to choose a client of the highest local ranking offered by other services. Third, reputation evaluation is based on the amount of credits that each node has collected through rating. Reputation evaluation is used to adjust reputation reward or punishment after a transaction [10]. This reputation system is implemented for peer to peer networks.

A probabilistic graphical reputation model is created to embody the relationship between social brands and users [11]. It collects the network information as well as the feedback of the users. This model reduces unfair outcome from a single user and a single comment. It jointly concludes the brand reputation. The implementation of this model is based on a parallel block-based Markov Chain Monte Carlo (MCMC) sampling method. The model is assessed by using a large amount of Facebook data. However, this model is very complicated and impractical.

Another reputation model is based on mining textual feedback of services to build an expectations model for each consumer is proposed in [12]. This model is created for e-commerce settings and the research is based on real data gained from eBay. It is based on textual feedback not rating.

To easily monitor the reputation of a company in the Twitters, a strategy that arranges a set of tweets into diverse clusters based on the tweets topics is developed in [13]. The acquired clusters are prioritized into different priority ranks. A cluster with high priority represents a topic which may affect the reputation of a company, and deserves instant attention.

E-Bay [c] and Amazon [d] are two common reputation systems in online e-commerce. In the e-Bay reputation system, a buyer can give feedback about the sellers service quality after each transaction and the system stores these ratings in a centralised manageable data storage. It calculates the feedback score by subtracting the negative ratings from the positive ratings and displays this score on the website.Amazon.com is one of the worlds leading online retailers. Amazon allows its users to review each product or service they receive. The reviews can then be accessed by all users. An average score for each product is calculated based on the reviews.

Foursquare [e] is a discovery service application which provides search results for its users. The application reads the users location and his request (e.g. restaurant) and provides data on nearby services meeting his requirements that other users trust. The application provides recommendations based on the user's current location.

Qaym.com [f] is a website that allows Saudi consumers to evaluate restaurants. It allocates rankings for each restaurant based on the customers comments, which are then used to define tags that describe the characteristics of the restaurant and supplemented by consumer ratings. The website then calculates the restaurants ranking based on consumer ratings and comments.

In this study, a reputation approach based on sentiment analysis of Arabic tweets in Twitter is proposed. More specifically, service provider's reputation is calculated from the collected Saudi tweets in Twitter. To do so, a Saudi dialect lexicon has been developed as a basic component for sentiment polarity. The proposed lexicon is used to classify words extracted from Twitter into either a positive or negative word. Then, beta probability density functions have been used to combine feedback from the lexicon to derive reputation scores.

III. Developing a Lexicon for Saudi Dialect

A. *Lexicon Definition*

A lexicon is a list of words in a language. The Cambridge Dictionary [g] defines lexicon as ”a list of all the words used in a particular language or subject, or a dictionary.” It may be general or domain-specific. The interest words are usually open-class or content words, such as nouns, verbs, and adjectives, rather than closed-class or grammatical function words, such as articles, pronouns, and prepositions. It may also include multi-word expressions such as fixed phrases, phrasal verbs, and other common expressions (Happy End). Each word or phrase in the lexicon is described in a lexical entry; exactly what is included in each entry depends on the purpose of the particular lexicon [14].

B. *Importance of Lexicon*

Broadly, there are two major categories for sentiment analysis: supervised and unsupervised approaches. In the supervised approach, data marked with its class (positive or negative) and used as training data for classification using one of the machine learning algorithms like Nave Bayesian Classifier, Support Vector Machine (SVM), Maximum Entropy [15]. One of the flaws of supervised approaches is that they require a carefully selected training set with highly accurate annotations. The unsupervised approach, on the other hand, is based on building a lexicon to infer class of each word. The advantage of this approach is that it is domain independent. However, building a high quality lexicon usually requires significant effort [6].

People on social media use their local dialect rather than Modern Standard Arabic (MSA) [16]. Therefore, in this paper, a Saudi dialect lexicon from Saudi tweets is developed to help calculate the reputation score for an entity.

C. *Saudi Dialect Lexicon*

Lexicon Structure: In this paper, a lexicon of Saudi dialect is developed. The developed Lexicon is composed of Saudi tweets as it is specialized in the Saudi dialect. A lexical entry in our lexicon is either a word or phrase which demonstrates how Saudis express their opinions on an entity. The word can

[c] http://www.ebay.com
[d] https://www.amazon.com
[e] https://foursquare.com
[f] http://www.qaym.com
[g] http://dictionary.cambridge.org

be a noun, verb, pronoun, adjective, adverb or interjection. Examples of word in Saudi dialect are نقمة, نعمة, زينة and افضل. A phrase is a group of words that express a concept and is used as a unit within a sentence. Examples of Saudi phrases are شي يصيح, نقطة ضعفي and آخر مرة. Some common phrases are used with clear sentiments, for example لن اكرر التجربة, to handle negating sentences.

The proposed lexicon is a two column file: a word/phrase and a score. The score captures the polarity of the word/phrase and it takes one of the two values: positive (+1) or negative (-1). Table 1 shows some examples about the Saudi dialect with their polarities.

Lexicon Building: Extracting sentiment from text is a complex task due to the significant amount of Natural Language Processing (NLP) required [6]. This task becomes even more difficult when dealing with morphologically rich languages such as Arabic language [15], [16] and when processing brief, noisy texts such as ”tweets” or ”Facebook statuses”.

Broadly, there are two lexicon building techniques, manual building and automatic building [17]. In this paper, a semi-automatic building technique was adopted. Initially, the lexicon was built manually by collecting tweets and classifying ther words in positive or negative polarity. After that, the proposed algorithm will enrich the lexicon by new words automatically each time a reputation calculation request submitted to the system (explained in details in Section 4). Finally, the new words will be classified manually by the lexicon moderator.

Given the limited work done for Arabic text in the field of sentiment analysis, especially for the Saudi dialect, two lists were manually built about Saudi dialect: one for the most occurring positive sentiment words, and one for the most occurring negative sentiment words. A NodeXl [h], a free open source to explore network graphs, was used to collect tweets written in the Saudi dialect from Twitter. Then, every word that has a sentiment (positive or negative) have been extracted. Consequently, 762 positive words/ phrases and 662 negative words/phrases have been collected. For each positive and negative sentiment, a PHP code scans the tweets and uses two different mechanisms, one for the extraced words and the other for the extracted phrases (see Algorithm 1 in Section 4).

The sentiment of each word was identified using crowd sourcing technologies [18]. That is, the list of extracted words were sent to three workers (Saudi people) to give a coarse sentiment label (positive, negative, or neutral) to each word. Then, a voting was used for final classification. Since each word was labeled by three workers, it is easy to determine if the target word represent a positive, negative or neutral word (i.e. if two out of three chose the same polarity).

IV. A Lexicon-based Approach to Build Reputation from Social Media

The proposed system consists of four main components: data collection, data pre-processing, sentiment analysis and reputation calculation. They are illustrated below in Figure 2.

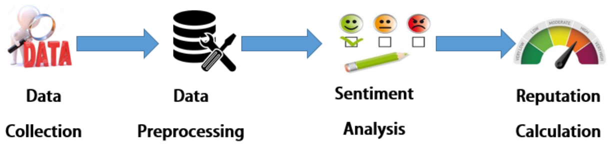


Fig. 2: Proposed system components

A. Data Collection

The collected data in the proposed system is tweets. They are collected automatically using Apache Nutch [20]. The Apache Nutch is a complete open source Web crawler written in Java. It is highly scalable and built over Hadoop Map/Reduce. It is used in distributed architecture and can automatically grab webpage hyperlinks. In the proposed system, the Apache Nutch was modified to crawl only Arabic tweets.

The Apache Nutch can copy all of the visited pages for searching and reduce the maintenance work. It is highly modular and most of its functions can be changed via plugins. It contains a data repository with two databases: web page status database and link database. The link database can be symbolized as web graph to grab web content and parses data. For good performance, it supports multi-threaded and multi-protocol.

B. Data Preprocessing

For data preprocessing, Apache Solr [21] is used. It is a fast open-source Java search server that allows easy creation of search engines for websites, databases and files. The basic component of the Apache Solr that makes it able to perform searches rapidly is the index it creates and uses to search the text. The index inverts a keyword-centric data structure based on the words in the pages. This index is stored in an index folder in the data folder.

When data is added to the Apache Solr, it passes through a series of transformations before being added to the index. This is regarded as the ”analysis phase”. Examples of these transformations are removing suffixes and prefix of Arabic words, removing stop-words, splitting words into sub-words that help to match delimiters with words.

The first step to work on the Apache Solr is to define kinds of fields using a file called schema.xml. The supported field types in the Apache Solr are: float, long, double, date, and text. This study is dealt with text filed since the collected data are ”tweets”. Each filed has an analyzer and tokenizer [23]. While the analyzer examines the field text and generate a ”token stream”, the tokenizer breakdowns the filed data into lexical units.

The Apache Solr offers support for the Light-10 (PDF) stemming algorithm. This algorithm describes character normalization and stemming. To provide flexibility, the stemming algorithm divided into two filters as in Figure 3:

1) solr.ArabicNormalizationFilterFactory, and

[h] http://nodexl.codeplex.com

TABLE I: Examples of the Saudi Dialect

(a) Word Sentiment

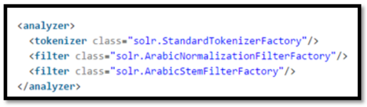
Word	Polarity
ابغا	+1
تحفه	+1
بايخ	-1
ادمان	+1
حسافه	-1
اطلق	+1
سخيف	-1
تعيس	-1
اردا	-1
ابداع	+1

(b) Phrase Sentiment

Phrase	Polarity
لا يعلا عليه	+1
الله يديم النعمه	-1
لا يفوتك	+1
اموت فيه	+1
يحوم الكبد	-1
ما فيه مثله	+1
ما ينفع بريال	-1
مالت عليهم	-1
لا يعلا عليه	+1
لا طعم و لا ريق	-1

2) solr.ArabicStemFilterFactory

Note that light stemming eliminates only the common affixes (prefixes and suffixes) without altering the origin (root) of a word.



```
<analyzer>
  <tokenizer class="solr.StandardTokenizerFactory"/>
  <filter class="solr.ArabicNormalizationFilterFactory"/>
  <filter class="solr.ArabicStemFilterFactory"/>
</analyzer>
```

Fig. 3: Indexing Arabic Content in Solr [22]

C. Sentiment Analysis

Sentiment analysis has recently become one of the growing areas of research related to text mining and natural language processing [5], [6]. It is a type of natural language processing for tracking the mood of the public about a particular product or topic. Not surprisingly, the most important indicators of sentiments are sentiment words. These are words that are commonly used to express positive or negative sentiments. For example, in Saudi dilaect زين, نعمه and أعشق are positive sentiment words, while شين, مخيس and معفن are negative sentiment words. Apart from individual words, there are also phrases and idioms, e.g., مالت عليه. Sentiment words and phrases are instrumental to sentiment analysis for obvious reasons. A list of such words and phrases is called a sentiment lexicon.

Since sentiment words are often the dominating factor for sentiment classification, it is not hard to imagine that sentiment words and phrases may be used for sentiment classification in an unsupervised manner. It performs classification based on some fixed syntactic patterns that are likely to be used to express opinions. The syntactic patterns are composed based on part-of-speech (POS) tags.

Another unsupervised approach is a lexicon-based which uses a dictionary of sentiment words and phrases with their associated orientations [24]. This method was originally used in sentence-level sentiment classification [25].

In this study, the lexicon-based approach has been adapted for sentiment analysis. That is, the developed Saudi dialect lexicon is used to classify the extracted words into either positive or negative word. To do so, a PHP plugin was integrated into the Apache Solr to enter the keyword (i.e. the targeted entity) and submit it to the Apache Solr to search for the relevant tweets. The different ways of typing any product in Saudi society are written in an excel file which is converted to a text file for use by the Apache Solr during the search and query processes. This file is added to the Apache Solr configuration to be used for synonym purposes. Finally, the PHP code scans the tweets to identify the sentiment words based on the collected sentiment words in the Saudi Lexicon. This way the meaning of each sentiment word is interpreted, which is the core of sentiment analysis.

D. Reputation Calculation

In this paper, the Beta reputation system that proposed by Josang and Ismail in [2] has been adapted to calculate the reputation of service provider. The reputation function is based on the Beta Probability Density Function. It is used to show the probability distributions of binary events. The binary events in this study are positive or negative sentiment. It is a mathematical base that is used to represent the reputation ratings based on feedback (tweets). It is based on posteriori probabilities of binary events to convey beta distributions [2]. It is a continuous function which uses two parameters α and β.

In the proposed approach, the extracted words from the collected tweets are viewed as a set of Bernoulli trials: positive word/phrase or negative word/phrase, and then modeled as Beta distributions:

$$R = \frac{\alpha}{\alpha + \beta} \tag{1}$$

Where α = r + 1 and β = s + 1, and r is the observed number of positive words and s the observed negative ones.

The ratio of α and β determines where in the interval [0,1] the reputation score peaks, and a high α will cause the score to occur close to 1.

The idea of adding 1 to r and s (and thus 2 to $r + s$) follows Laplace's rule of succession for applying probability to inductive reasoning [26]. This rule reflects the assumption of an equi-probable prior, which is commonly adopted in probabilistic reasoning. That is, having no information about an entity, (i.e. $r = 0$ and $s = 0$), R is calculated as: 1 / (1+1) =0.5, suggesting that a positive and negative word about the entity is equally likely. For example, suppose that α equals to 8 and β equals to 2. In this case, 10 words were extracted from the collected tweets and the reputation of the given entity can be calculated as follow: $R = 8/(8 + 2) = 8/10 = 0.8$

Note that in this study rating nearer to "1" donates good reputation score and vice versa. Algorithm 1 shows pseudo-Code for a lexicon-based approach to calculate the reputation score in detail.

Algorithm 1 Reputation Calculation Algorithm

```
 1: Input: $T$ a set of tweets extracted from Twitter,
           $L$ a Saudi Dialect Lexicon.
 2: Output: $R$ Reputation Score.
 3: Process:
 4: $\alpha$ = 0
 5: $\beta$ = 0
 6: for each $t_i$ do
 7:    for each $w_i$ do
 8:       if $w_i \in \gamma$ then
 9:          if $w_i$.Polarity $=$ +1 then
10:             $\alpha$ = $\alpha$ + 1
11:          else
12:             $\beta$ = $\beta$ + 1
13:          end if
14:       else
15:          $\xi \leftarrow w_i$
16:       end if
17:    end for
18: end for
19:
20: for each $p_i \in \psi$ do
21:    for each $t_i$ do
22:       if $p_i \in t_i$ then
23:          if $p_i$.Polarity $=$ +1 then
24:             $\alpha$ = $\alpha$ + 1
25:          else
26:             $\beta$ = $\beta$ + 1
27:          end if
28:       end if
29:    end for
30: end for
31: $R$ = $\frac{\alpha}{\alpha+\beta}$
32: Return $R$
```

In this research, the lexicon-based approach is used to compute the reputation score of a given service provider. The developed lexicon L:=(γ, λ , ψ , ξ) where γ is a set of positive and negative words in the Saudi dialect lexicon, λ is a set of neutral words (i.e. neither positive nor negative), ψ is a set of positive and negative phrases, and ξ is a set of non-known words that need to be checked by the lexicon moderator and added to one of the first two groups.

The input for the proposed algorithm are T and L (Line 1) and the output is the reputation score R (Line 2). T is a set of tweets that have been collected from Twitter about a target entity and L is the developed Saudi dialect lexicon.The algorithm uses two variables: α and β, the former is for calculating total number of positive ratings (Line4) and the later is for calculating total number of negative ratings (Line5). The algorithm then scans all the words in T (from Line 6 to Line 18). That is, for each tweet t_i, it checks each single word w_i (Line 7) and if it has a positive sentiment (Line 9) it adds 1 to α (Line 10), otherwise it adds 1 to β (Line 12). In case of having a word that does not exist in the lexicon, the word will be added to ξ (Line 15). The second part of the algorithm (from Line 20 to Line 30) scans all the phrases in ψ (Line 20) to check whether the phrase is existed in the collected tweets or not. That is, for each tweet (Line 21), it checks if the given phrase (p_i) exists in the tweet (Line 22) α will be increased by 1 if the phrase's polarity equals to +1 (lines 23 and 24), otherwise it adds 1 to β (Line 26). The lines from 20 to 30 are repeated for all phrases in the lexicon. Once the algorithm has scanned all of the collected words and the lexicon phrases, it moves to calculate the reputation score by dividing the total number of positive ratings (α) by the sum of positive and negative ratings (Line 31) and then returns the result to the requester (Line 32).

To explain the importance of having list of neutral words in our lexicon, a specific example is introduced. Suppose that 500 words were extracted from 100 tweets about service provider $SP1$. According to Algorithm 1, the 500 words will be compared with the constructed lexicon. Assume that only 100 words exist in the lexicon. According to the proposed algorithm, 400 words will be added to the waited list (ξ) in order to be compared with the neutral list (λ). Now, assume that 300 words out of the 400 words have already existed in the neutral list. These words will be consequently removed from the waited list. Finally, the remaining 100 words will be manually classified by the lexicon moderator as either positive or negative. Thus, having the neutral list in the proposed lexicon decreased the moderators effort by almost 75% (i.e. instead of classifying 400 words, the moderator classified only 100 words). Note that, in general, the proposed lexicon will build itself over each time a new reputation calculation request submitted to the system.

V. Experimental Evaluation

This section presents an empirical evaluation of the lexicon-based approach. A set of experiments were conducted to evaluate performance our proposed approach against the sentiment analysis approach.

A. Data Set

The data set used in this experiment was automatically collected from Twitter using Apache Nutch [20]. In this study, Apache Nutch was modified to crawl only the Arabic tweets on Twitter. To identify Saudi tweets, the collected tweets were filtered based on user location.

For the purposes of this study, 550 tweets were collected from Twitter regarding 9 restaurants (see Table 2). The collected data was injected into Solr to implement the data preprocessing. The entered tweets were indexed for fast and efficient retrieval by the Solr indexer.

TABLE II: Number of Collected Tweets for Each Restaurant

Entity	Number of tweets
Restaurant 1	56
Restaurant 2	59
Restaurant 3	58
Restaurant 4	64
Restaurant 5	55
Restaurant 6	72
Restaurant 7	68
Restaurant 8	66
Restaurant 9	52
Overall tweets	550

B. Evaluation Metrics

In this study, the lexicon-based approach has been evaluated using three metrics: precision, recall and reputation score.

Precision: The frequency with which retrieved documents or predictions are relevant or 'correct', and is properly a form of Accuracy. It is also known as Positive Predictive Value (PPV) or True Positive Accuracy (TPA). In the field of information retrieval, precision can be defined as the fraction of retrieved documents that are relevant to the query [27]. More specifically, precision can be calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Where TP is the true positive and FP is the false positive. In our context, the retrieved documents are sentiment words that were extracted from the collected tweets. For example, suppose that 100 words were extracted from Twitter and 25 were mistakenly retrieve (they either do not express sentiment or were misclassified). In this scenario, the precision is as follows: 75/(75 + 25) = 0.75.

Recall: The frequency with which the relevant documents are retrieved or 'recalled' by a system. It is also known as Sensitivity or True Positive Rate (TPR). In the field of information retrieval, recall is the fraction of the documents that are relevant to the query that are successfully retrieved [27]. More specifically, recall can be calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Where TP is the true positive and FN is the false negative. For example, suppose that the collected tweets contain 100 sentiment words and only 60 were extracted and retrieved. In this case, the recall can be calculated as follows: 60/(60 + 40) = 0.60.

Reputation Score: In this work, the ratings are represented by the extracted words from the collected tweets. More specifically, the reputation score is calculated using equation1. For example, if 95 positive words and 15 negative words were extracted from the collected tweets, then the reputation score is computed as: 95/(95 + 15) = 0.86.

C. Experimental Results

In this section, a set of experiments were conducted to evaluate the performance of the proposed approach. First, for validity purposes, calculating reputation based on words extracted from the collected tweets was examined against a sentiment analysis approach (i.e. using sentence-level rather than word-level). Second, the proposed system was used to calculate the reputation scores of the restaurants. Finally, how the reputation of the restaurants might be fluctuated over time and how this fluctuation may affect the user's decision was demonstrated.

Experiment 1: Word-level vs. Sentence-level: To prove the concept, an evaluation of the proposed approach is presented in this section. A comparison between the proposed approach (word-level) against a sentiment analysis (sentence-level) was conducted. For simplicity, in the sentiment analysis approach, each collected tweet was manually classified as positive, negative or neutral.

To explain how to calculate the reputation score using the two approaches, a specific example is introduced. In this example: ”مطعم لوسين يجنن وأفضل مطعم” means ”Lusin restaurant is amazing and it is the best restaurant”. Two sets of calculations were performed to determine the reputation score for the restaurant.

- Sentence-level: this approach attempts to label the whole sentence as positive or negative sentiment. Thus, it assigned a score of +1.
- Word level: this approach counts the number of positive and negative words in the sentence. In our example, there are only two positive words (”amazing” and ”best”), hence, it assigned a score of +2.

For the sake of this experiment, tweets about 5 different restaurants were collected from Twitter. On average, 60 tweets were collected about each restaurant. The reputation scores of the four restaurants were calculated based on the two approaches and the result is shown in Figure 4.

Figure 4 shows results of the word-level approach against the sentence-level approach (sentiment analysis approach). It is clear that the calculated reputation scores for the five restaurants are close to those scores calculated by the sentence-level approach. For example, restaurant 1 has the highest scores in both approaches (0.96 using our word-level approach and 0.98 using sentence-level approach). More importantly, the order of the five restaurants is equivalent by the two approaches. In addition, how the Foursquare Application rates these restaurants was investigated and restaurant 1 was founded as the best one, which supports the results of the two approaches. Furthermore, the order of the restaurants was similar to their order in the Foursquare application. Note that since the calculation of the

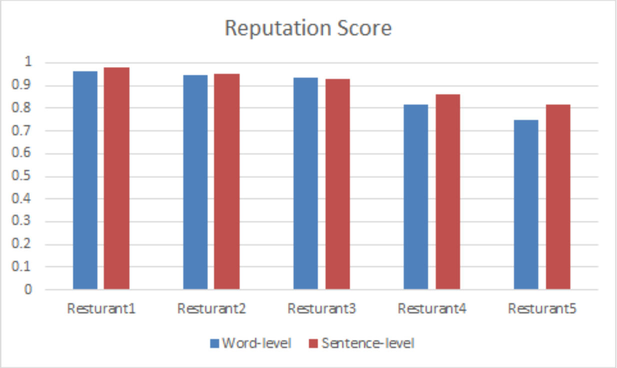


Fig. 4: Reputation scores of word-level vs. sentence-level

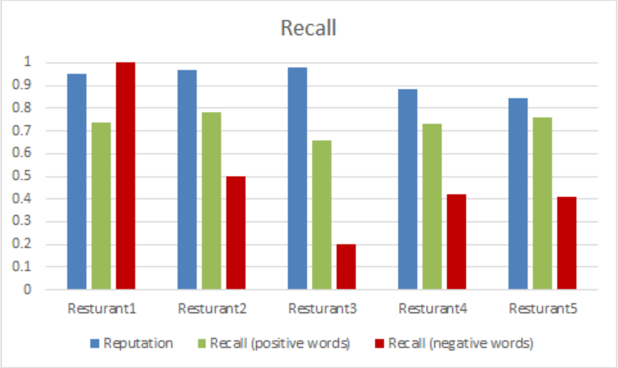


Fig. 6: Recall of positive and negative words against reputation scores

two approaches was conducted manually, the precision and recall for the two approaches was 100%.

In addition, compared to other websites which evaluate restaurants in Saudi Arabia, the proposed reputation system gave similar results to Qaym. For example, it was found that restaurant 2 is better in evaluation than restaurant 3, which supports the results of the proposed approach.

Experiment 2: Reputation calculation using a lexicon-based approach: In this experiment, reputation scores of the five restaurants were calculated using a lexicon-based approach. The lexicon-based approach was implemented using PHP language [i]. The proposed approach scans the collected tweets to extract the sentiment words and uses the Saudi Lexicon to classify them as positive or negative. The reputation scores of the five restaurants were recalculated using the lexicon-based approach and the results shown in Table 3 and Figure 5.

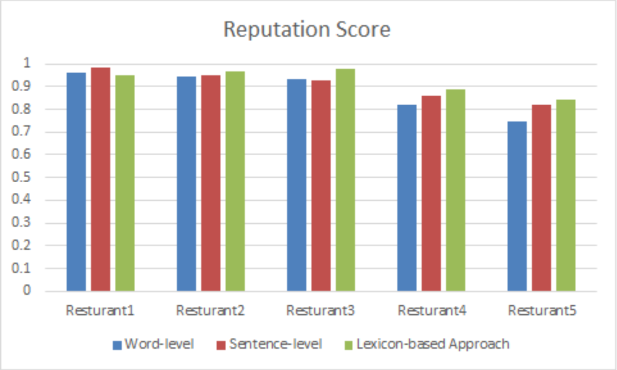


Fig. 5: Reputation scores using the lexicon-based approach

In Figure 5, it is clear that the reputation scores of the five restaurants using our proposed approach is similar to the reputation scores calculated by the two other approaches. However, looking closely at the scores reveals that the order of the restaurants has changed. That is, while restaurant 1 was recommended as the best one by the two other approaches, it was ranked third by the lexicon-based approach. It is important to emphasize that the lexicon-based approach is highly sensitive to the content of the Saudi dialect lexicon. The next experiment explains the reasons behind the change of the restaurants order between the proposed approach and the two other approaches.

Experiment 3: Calculation of Recall and Precision: While recall is the frequency with which the relevant documents are retrieved or 'recalled' by a system, precision is the frequency with which the retrieved documents or predictions are relevant or 'correct'. In the proposed approach, the reputation score will be directly affected by the frequency of the extracted words (recall) and their correctness (precision). Table 4 and Figure 6 show recall values that are associated with each restaurant's reputation score. Note that precision and recall of positive and negative words were calculated separately in order to clearly understand the difference in the restaurants' ranking between the proposed approach and the other approaches.

Figure 6 shows that the recall of negative regarding restaurant 1 is slightly higher than that of the recalls of other restaurants (i.e. all negative words against the restaurant have been retrieved). In addition, approximately 19 positive words were missed, making the recall of the positive words relatively poor (74%). However, although the recall of the positive words for the claimed best restaurant (number 3) was the lowest (66%), the majority of the negative words were not discovered (almost 80% of the negative words) leading to the highest reputation score. Overall, it is evident that the reputation scores were significantly influenced by the recall values for positive and negative words. That is, if the recall of positive words reaches 1 while the recall of negative words drops towards 0, then the reputation score will be increased and vice versa. In an optimal case where the recall of positive and negative words equals 1, the reputation score computed by the lexicon-based approach will be equivalent to the word-level approach. This will happen when the Saudi dialect lexicon becomes complete and accurate.

Table 5 and Figure 7 illustrate the precision of positive and negative words for the five restaurants. It is clear that the precision values for the five restaurants are quite high compared to the recall values, therefore, the reputation scores were influenced by the low recall values. As mentioned previously, once the Saudi dialect lexicon becomes complete and

[i] http://php.net/manual/en/langref.php

TABLE III: Reputation Scores of Lexicon-based Approach vs. Sentiment Analysis Approach

Restaurant	Manual Calculation		Automated Calculation
	Word-Level	Sentence-Level	Lexicon-based Approach
Restaurant 1	0.961	0.981	0.949
Restaurant 2	0.945	0.949	0.966
Restaurant 3	0.932	0.928	0.979
Restaurant 4	0.818	0.859	0.885
Restaurant 5	0.746	0.818	0.844

TABLE IV: Recall of Positive and Negative Words for the Five Restaurants

Metric	Res1	Res2	Res3	Res4	Res5
$TP + FN$	75	71	69	63	50
TP	56	56	46	46	38
Recall (positive words)	0.74	0.78	0.66	0.73	0.76
$TP + FN$	3	4	5	14	17
TP	3	2	1	6	7
Recall (negative words)	1	0.5	0.2	0.42	0.41
Overall Recall	0.76	0.77	0.64	0.68	0.67

TABLE V: Precision of Positive and Negative Words for the Five Restaurants

Metric	Res1	Res2	Res3	Res4	Res5
$TP + FN$	55	56	45	43	35
TP	56	56	46	46	38
Precision (positive words)	0.98	1	0.97	0.93	0.92
$TP + FN$	3	2	1	5	7
TP	3	2	1	6	7
Precision (negative words)	1	1	1	0.83	1
Overall Precision	0.983	1	0.979	0.923	0.93

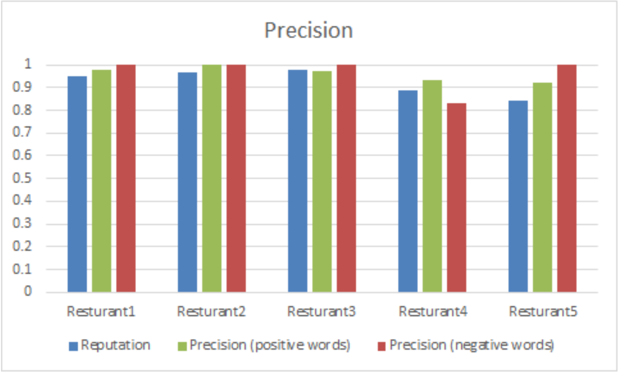


Fig. 7: Precision of positive and negative words against reputation scores

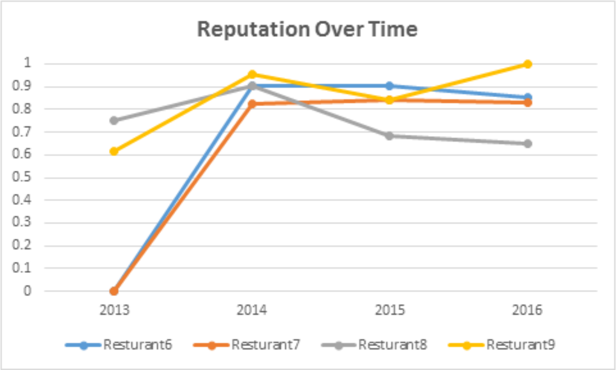


Fig. 8: Classification Accuracy

accurate the reputation scores of the given entities will be more accurate.

Experiment 4: Reputation fluctuation over time: There are many factors that might positively or negatively affect the reputation score of a given entity, which are reflected in people's opinions towards the entity. For example, it is expected that the reputation of a restaurant that offered a delicious food might be affected when the restaurant's chief moves to another restaurant.

The aim of this experiment is to examine how the reputation score can be affected over time. Therefore, 303 tweets were collected about an additional four restaurants (restaurant 6, restaurant 7, restaurant 8 and restaurant 9). Tweets were collected from a period of four years for each one and the annual reputation score was calculated. Note that in this experiment the lexicon with additional positive and negative words was updated to improve the recall and precision of our proposed approach. Figure 8 presents the results of this experiment.

In Figure 8, it is obvious that the reputation scores of the four restaurants fluctuated over the four years. While the restaurant 8 was the best one in 2013, its reputation started decaying in 2015 (it dropped to become the worst one). Conversely, restaurant 9 started with low reputation and gradually built its reputation to become the best one in 2016. In addition, it is evident that restaurants 6 and 7 deliver consistent services over time. This is reflected by the stability of their reputations over 2014, 2015 and 2016. Note that there was no data for these two restaurants in 2013 and zero values in this year simply mean that they had no reputation.

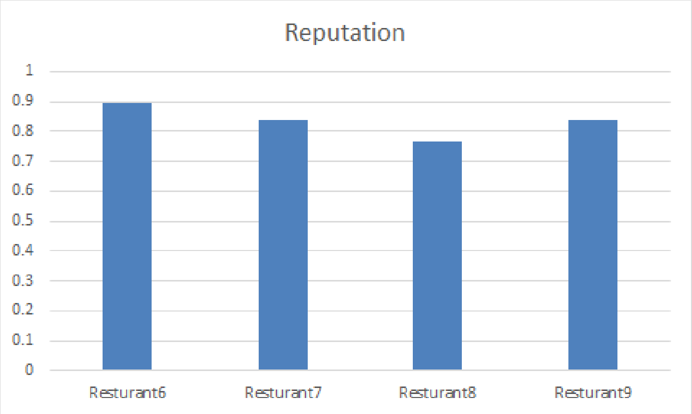


Fig. 9: Reputation of the four restaurants over the four years

Old feedback may not always be relevant to the reputation score, because the quality of the restaurant, for example, may change over time. One possible way to handle this issue is by introducing a forgetting factor. The main idea for the forgetting factor is to give the old feedback less weight than the new one when calculating the reputation score. The forgetting factor can be adjusted according to the expected rapidity of change in the observed entity. As Figure 9 shows, if the historical data that collected over the four years was considered, restaurant 6 would be nominated as the best restaurant (its reputation is 89%). However, if only the recent feedback that gathered on the last year (2016) was considered, then restaurant 9 will be the best one. It is important to emphasize that the data on 2016 only covers three months (from January to March).

VI. Discussion

In this paper, the word polarity was represented in the same level (+1 for positive word and -1 for negative word), and the reputation score is calculated by counting the number of positive and negative words using the beta probability density function. However, categorizing all the words at the same level is rather limited and does not allow to specify the level of polarity effectively. For example, it is unfair to assign the two words رائع and زين at the same level of polarity (i.e. +1). Also, it was approved that users tend to repeat the letter more than once to emphasize the meaning or feeling in social media [24]. Saudi users, for example, often duplicate letters to express overstatement of their opinion. They repeat the letter of ي in the word لذيييذ which means delicious. This word is exactly similar to the word لذيذ, but with more emphasize of feeling. Thus, It might be more appropriate to consider different levels of polarity instead of categorizing all the words at the same level. That is, in categorizing the polarity of the two words, the word with repeated letter would be assigned higher level of polarity. A method proposed in [28] can be used for inferring the polarity degree of a word from its statistical association with a set of positive and negative paradigm words.

In reputation calculation, the beta probability function was adopted. However, it is limited to work with only two values (positive or negative). In the other words, this model excludes the possibility of providing word polarities with graded levels such as praise, positive, slightly positive, ..etc. In principle this model is unable to distinguish between polarized ratings. One possible solution is to build the reputation score based on the Dirichlet probability distribution which is a multinomial Bayesian probability distribution as in [29]. The multinomial aspect of Dirichlet probability distribution means that any set of discrete rating levels can be defined. Hence it provides great flexibility and usability in dealing with multilevel of word polarity.

Although sentiment words and phrases are important for sentiment analysis, only using them may lead to inaccurate classification. For example, in our lexicon احب is a positive word but negation لا احب will change it to a negative one. Also, words in Arabic can have a different meanings based on their context. For example the word يم has two meanings in the Saudi dialect; delicious or towards a direction or place. These words influence the precision of our lexicon and increase the probability of a false or misleading detection. Approaches proposed in [34]–[36] can be applied to determine the correct polarity of the words by take into account the context of the words .

The proposed approach establishes service provider’s reputation from extracted peoples opinions from Twitter. However, people can be misused by organizations to write fake opinions (i.e. spammers tweets) [32]. For example, a seller of musical instruction DVDs was penalized $250,000 by the U.S. Federal Trade Commission after being accused of using partners to write positive reviews of the companys products on various websites [33]. Such fake opinions for sure will affect the calculated reputation score and mislead the consumer to select unreliable service provider. The proposed techniques in [30] and [31] can be used to detect and remove spam tweets to produce more reliable reputation scores.

VII. Conclusions

In this paper, a lexicon-based approach to calculate reputation score from Twitter was proposed. A Saudi dialect lexicon from Saudi tweets was developed to improve addressing the sentiment of the Arabic tweets. The experimental results show that the proposed approach was efficient and gave similar results to sentiment analysis approach. Also, the result was consistent with the result of Qaym in evaluating the same selected restaurants.

A possible extension of the proposed approach is to consider multilevel of word polarity rather than binary level. Also, spam detection techniques can be applied to detect and remove spam tweets in order to produce more reliable reputation scores.

References

[1] X. Dimitrios, A. S. Alali . Investigating the attitude of the average Saudi towards the Social Media. ACACOS, Vol. 14, 86-94. (2014).

[2] A. Josang, R. Ismail, C. Boyd. A survey of trust and reputation systems for online service provision. Decision support systems, 43(2), 618-644. (2007).

[3] Y. M. Afify, I. F. Moawad, N. L. Badr, and M. F. Tolba. A personalized recommender system for SaaS services. Concurrency and Computation: Practice and Experience, Wiley Online Library, (2016).

[4] Y. L. Sun, Y. Liu. Security of Online Reputation Systems: The evolution of attacks and defenses. IEEE Signal Process. Mag., 29(2), 87-97. (2012).

[5] B. Pang, L. Lee. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135. (2008).

[6] B. Liu. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167. (2012).

[7] Arab Social Media Report. TNS, first report. (2015): Available at:http://dmc.ae/img/pdf/white-papers/ArabSocialMediaReport-2015.pdf, Accessed May 15th, 2016.

[8] I. Lin, H. Wu, S. Li, and C. Cheng. A fair reputation system for use in online auctions. Journal of Business Research, Elsevier, 86(4): 878-882, (2015).

[9] H. Al-Dossari. A Reputation-based Approach for Consumer Protection in Saudi Arabia. 4th International Conference on Computational Aspects of Social Networks (CASoN), IEEE, 321-326, (2012).

[10] K. Chen , H. Shen, K. Sapra and G. Liu. A Social Network Integrated Reputation System for Cooperative P2P File Sharing. 22nd International Conference in Computer Communications and Networks (ICCCN), Nassau, (2013).

[11] K. Zhang, D. Downey, . Z. Chen , Y. Xie , Y. Cheng , A. Agrawal , . W.-k. Liao and A. Choudhary. A probabilistic graphical model for brand reputation assessment in social networks. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, New York, (2013).

[12] A. M. ElMessiry, X. Gao and M. P. Singh. Incorporating Expectations as a Basis for Business Service Selection. In Service-Oriented Computing, Springer Berlin Heidelberg, 486-500, (2010).

[13] A. Qureshi, C. O'Riordan and G. Pasi. Concept Term Expansion Approach for Monitoring Reputation of Companies on Twitter. In CLEF (Online Working Notes/Labs/Workshop), Rome, (2012).

[14] G. Hirst. Ontology and the Lexicon. In Handbook on Ontologies, Springer Berlin Heidelberg, 269-292, (2009).

[15] A. Shoukry, A. Rafea. Sentence-level Arabic sentiment analysis. 2012 International Conference on Collaboration Technologies and Systems (CTS), IEEE, 546-550, (2012).

[16] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat. Sentiment Analysis in Arabic tweets. 5th International Conference on Information and Communication Systems (ICICS), IEEE, 1-6, (2014).

[17] N. Abdulla, N., Ahmed, M. Shehab, M. Al-Ayyoub, and M. Al-Kabi. Automatic Lexicon Construction for Arabic Sentiment Analysis. International Conference on Future Internet of Things and Cloud (FiCloud), Barcelona, IEEE, 547-552, (2014).

[18] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing systems on the World-Wide Web. Communications of the ACM, 54(4), 8696, (2011).

[19] S. Linda, and K. Bharadwaj. A Fuzzy Trust Enhanced Collaborative Filtering for Effective Context-Aware Recommender Systems. Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2, Springer, 227-237, (2016).

[20] Nutch. Welcome to the Apache Nutch Wiki. Retrieved from https://wiki.apache.org, (2015).

[21] Kuć, Rafał. Apache Solr 4 Cookbook. Packt Publishing Ltd, (2013).

[22] Solr Wiki. Language Analysis. Retrieved from: https://wiki.apache.org/solr/LanguageAnalysisArabic, (2014).

[23] C. Targett. Understanding Analyzers, Tokenizers, and Filters. Retrieved from:https://cwiki.apache.org, (2014).

[24] M. Taboada, B. Julian, T. Milan, V. Kimberly, and S. Manfred. Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267-307, (2011).

[25] X. Ding, B. Liu, and S. P. Yu. A holistic lexicon-based approach to opinion mining. in Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008). (2008).

[26] E. S. RISTAD. A Natural Law of Succession. TR 495-95, Department of Computer Science, Princeton University, July, (1995).

[27] D. Olson, and D. Delen. Advanced Data Mining Techniques. Springer-Verlag Berlin Heidelberg, (2008).

[28] P. D. Turney, D. Peter, and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4), 315346, (2003).

[29] A. Josang, and J. Haller. Dirichlet Reputation Systems. The Second International Conference on Availability, Reliability and Security, IEEE, 112-119, (2007).

[30] M. Daiyan, S. K. Tiwari, and M. A. Alam. Mining product reviews for spam detection using supervised. International Journal of Emerging Technology and Advanced Engineering. 4(8), 619-623, (2014).

[31] S. Xie, G. Wang, S. Lin, and P. S. Yu. Review spam detection via temporal pattern discovery. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 823-831, (2012).

[32] R. Nithish, S. Sabarish, M. N. Kishen, A. M. Abirami, and A. Askarunisa. An ontology based sentiment analysis for mobile products using tweets. 5th International Conference on Advanced Computing (ICoAC), IEEE, 342,347, (2013).

[33] FTC. Firm to pay ftc $250,000 to settle charges that it used misleading online ”consumer” and ”independent” reviews. Retrieved from: http://www.ftc.gov/news-events/press-releases/2011/03/firm-pay-ftc-250000-settle-charges-it-used-misleading-online (2011).

[34] R. Gonzlez-Ibez, M. Smaranda, and W. Nina. Identifying sarcasm in Twitter: a closer look. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:short papers (ACL-2011). (2011).

[35] A. Kennedy, and I. Diana. Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22(2), 110-125, (2006).

[36] B. Li, Z. Lanjun, F. Shi, and W. Kam-Fai. A Unified Graph Model for Sentence-Based Opinion Retrieval. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010). (2010).