Developing Lexicon-based Algorithms and Sentiment Lexicon for Sentiment Analysis of Saudi Dialect Tweets

Waleed Al-Ghaith
Department of Information Systems
Shaqra University
Imam Muhammad Ibn Saud Islamic University
Riyadh, Saudi Arabia

***Abstract*—Majority of studies on sentiment analysis field, specifically Arabic lexicon-based approach, are focused on doing preprocessing methods on targeted dataset text or collected textual data from Twitter (Twitter dataset) rather than dealing with lexicon itself. This study proposes a new method, we constraint firstly on building a new sentiment lexicon with reasonable number of words and then doing adequate preprocessing methods on the lexicon's words in addition to the (Twitter dataset). The study presents Saudi Dialect Sentiment lexicon called SaudiSentiPlus contains 7139 words which mostly generated from Saudi tweets and other dictionaries. Moreover, this study also presents two lexicon-based algorithms for Saudi dialect to deal with (prefixes and suffixes) letters in order to increase performance of proposed Saudi dialect lexicon. The experiment which has been conducted in this study to evaluate the performance of SaudiSentiPlus comprises four phases. The precision, recall, accuracy, and F-Score are measured in every phase. We built our testing dataset from twitter by focusing on Saudi dialect hashtags (971 thousands tweets from 162 hashtags). The results, show that accuracy of SaudiSentiPlus with the two lexicon-based algorithms reached to 81%.**

Keywords*—Sentiment analysis; opinion mining; lexicon-based; Arabic text mining; Saudi Arabia

I. Introduction

A Social Network Site (SNS) is a platform enables people to share their opinions on any issue and to build social relations with individuals within and beyond their social circle [1].

Twitter as a one of the most popular SNSs that has been growing rapidly in recent years. Twitter's users increased by more than 500% since 2009 [1]. Twitter's users express their feeling, opinions or spreads news or facts about 200 billion times annually via their tweets, 500 million of them per day, 350,000 per minute, and 6000 per second [2].

In 2014, total number of active Twitter users in the Arab world reached 5,797,500 users and the country with the highest number of active Twitter users in the Arab region is Saudi Arabia with 2.4 million users, accounting for over 40% of all active Twitter users in the Arab region. The estimated number of tweets produced by Twitter users in the Arab world in March 2014 was 533,165,900 tweets, an average of 17,198,900 tweets per day [3] (see Fig. 1).

Currently, in 2019, Saudi Arabia was ranked the fourth in the world with around 10 million active users after the United States, Japan and the United Kingdom (see Fig. 2) [4].

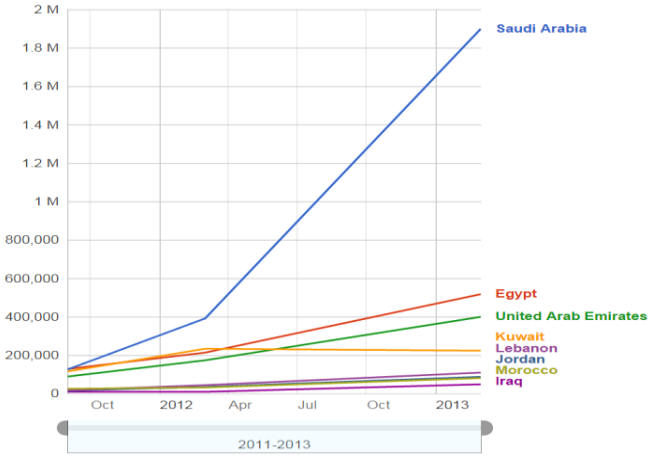

Fig. 1. Number of Active Twitter users in the Middle East [3].

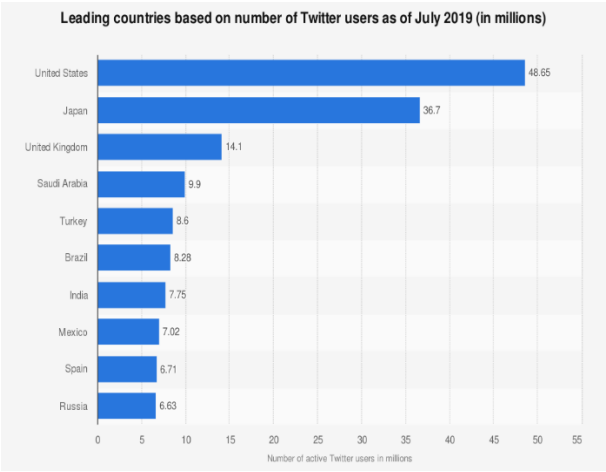

Fig. 2. Number of Active Twitter users in the World [4]..

This means that a massive amount of contents and opinions of Saudis toward phenomenon, topic, institution or individuals can be obtained and studied via twitter. This content can be either objective contents (e.g. news, facts) or subjective

contents such as (opinions or sentiments about entities). Opinions mining or sometimes alternatively mentioned as sentiment analysis is the research discipline which aims to analyze individuals' sentiments or opinions toward entities such as topics, people, issues, organizations or events [5] and classifying them as negative, positive or neutral opinions.

The Saudis speak and write in Arabic language and few of them are fluent in English. Arabic language is the fastest growing language on the web (8,917.3 %), it is ranked the fourth among languages on the web as illustrated in Table I [6].

Arabic language has many variants, however we can categorized it to three categories. The first is the Qur'an language which is classical Arabic; the second is Modern Standard Arabic (MSA) which it used in formal speech and writing. The third is informal or dialectical Arabic. Dialectical Arabic refers to all oral diversities spoken in daily communication for 27 Arabic countries and from one area of the same country to another [7].

According to Darwish and Magdy [8] Arabic social media's users tend to use Arabic dialects online rather than MSA. Likewise, Saudis use their colloquial language in social media and in Twitter in particular, which makes study their opinions or doing sentiment analysis based on their tweets a challenging task. In social media, Arabic colloquial or dialect are changeable and has word elongations with nonstandard spellings. Consequently, doing sentiment analysis based on standard formal-dependent lexicon is inefficient since that it will be unable to capture colloquial or dialect language in social media text. Thus, there is a need to develop another efficient method considering create a dialect-dependent lexicon for sentiment analysis of social media.

In this study, we think beyond of the box, we constraint firstly on building a new sentiment lexicon with reasonable number of words and phrases, and then conducting adequate preprocessing methods on the lexicon's words and phrases in addition to the (Twitter dataset). To the best of our knowledge, no such effort (doing preprocessing methods on the lexicon and collected dataset) has been made in prior studies. This study presents Saudi Dialect Sentiment lexicon (SaudiSentiPlus) contains 7139 words and can be used for sentiment analysis of Saudi dialect tweets. Moreover, in this paper we propose a new method based on presenting two lexicon based algorithms to deal with (prefixes and suffixes) letters of lexicon's words and phrases. This new method has a positive significant effect on increasing the performance or accuracy of (SaudiSentiPlus) lexicon. We evaluated the performance of SaudiSentiPlus through four phases. The precision, recall, accuracy, and F-Score are measured in every phase. We built our testing dataset from twitter by focusing on Saudi dialect hashtags (971 thousands tweets from 162 hashtags). We asked three annotators to classify the dataset's tweets randomly and manually to three classifications (positive, negative, and neutral) as presented in evaluation section.

Next section presents the proposed methodology in details. Followed by evaluation, results and discussion, and then we conclude this study in Section 5.

TABLE. I. TOP TEN LANGUAGES USED IN THE WEB

Top Ten Languages Used in the Web - April 30, 2019 (Number of Internet Users by Language)					
TOP TEN LANGUAGES IN THE INTERNET	**World Population for this Language (2019 Estimate)**	**Internet Users by Language**	**Internet Penetration (% Population)**	**Internet Users Growth (2000 - 2019)**	**Internet Users % of World (Participation)**
English	1,485,300,217	1,105,919,154	74.5 %	685.7 %	25.2 %
Chinese	1,457,821,239	863,230,794	59.2 %	2,572.3 %	19.3 %
Spanish	520,777,464	344,448,932	66.1 %	1,425.8 %	7.9 %
Arabic	**444,016,517**	**226,595,470**	**51.0 %**	**8,917.3 %**	**5.2 %**
Portuguese	289,923,583	171,583,004	59.2 %	2,164.8 %	3.9 %
Indonesian / Malaysian	302,430,273	169,685,798	56.1 %	2,861.4 %	3.9 %
French	422,308,112	144,695,288	34.3 %	1,106.0 %	3.3 %
Japanese	126,854,745	118,626,672	93.5 %	152.0 %	2.7 %
Russian	143,895,551	109,552,842	76.1 %	3,434.0 %	2.5 %
German	97,025,201	92,304,792	95.1 %	235.4 %	2.1 %
TOP 10 LANGUAGES	5,193,327,701	3,346,642,747	64.4 %	1,123.0 %	76.3 %
Rest of the Languages	2,522,895,508	1,039,842,794	41.2 %	1,090.4 %	23.7 %
WORLD TOTAL	7,716,223,209	4,386,485,541	56.8 %	1,115.1 %	100.0 %

Source: [6]

II. Proposed Methodology

Variant methodologies were wildly adapted among researchers to produce sentiment lexicons [9, 10], however, three approaches of them are broadly accepted in Arabic lexicon construction process. The first approach is building sentiment lexicon by taking words from Arabic dictionaries or from other sentiment lexicon and divided these words based on their polarities (see [11], [12]). The second approach is based on translation of English lexicons (see [13]). The third approach, based on selecting seed sentiment words and then finding the words that occur in conjunction with the seed words (see [14]).

In this study, we started with building the study's sentiment lexicon, firstly, by using the second approach which is automatic translation of English sentiment lexicons that already created by two prior studies [15] and [16]. Then, in order to enrich the study's sentiment lexicon with Saudi dialect words, we manually extracted all the sentiment Saudi dialect words from the twitter data (datasets). This approach partly inlines with the third approach.

Finally, more words (4431 words) have been taken from Saudi dialect sentiment lexicon (SauDiSenti) for sentiment analysis of Saudi dialect tweets [17] and next we deleted the repeated words and divided all these words based on their polarities. This approach is consistent with the first approach. To determinate the polarity, two-way classification (positive or negative) on the datasets has been adopted. We asked three annotators to classify the words manually. All the annotators are Saudi and Arabic native speakers and two of them are Arabic language teachers. If there is any disagreement among the three annotators, we solved it by voting. We called the study lexicon SaudiSentiPlus, it contains (7139 words).

A. Preprocessing

Applying sentiment analysis directly to collected textual data from the Twitter (Twitter dataset) could lead to inaccurate outcomes [18]; thus, collected textual data or (Twitter dataset) need to be prepared for another processing procedure. This called data preprocessing which means transform the collected textual data or dataset into a format that be more adequate to the purpose of the study. Applying preprocessing techniques before doing the sentiment analysis processes is highly recommended in many Arabic sentiment analysis studies, particularly in the Arabic dialectal dataset because it is commonly written in an unstructured shape [19, 20, 5, 11].

Thus, in this study, to enhance sentiment analysis results for the collected Twitter dataset, we applied some of preprocessing techniques which are as follows:

1) Tweet cleaning. This step to remove irrelevant data such as user names, Twitter characters, URLs and all non-Arabic letters.

2) Elimination of redundant letters. In a moment of emotion, some twitter's users tend to repeat some word's letters when they want to emphasize something such as Gooooooal, COOOOL, WOOOW, In the same way, it happens to Arab or Saudi twitters for example "أليييييييييم" or "رووووعة".

B. The Proposed Algorithms

In order to increase performance of proposed Saudi dialect lexicon (SaudiSentiPlus), we developed two lexicon based algorithms to deal with (prefixes and suffixes) letters of the lexicon's words (see Fig. 3 and Fig. 4).

Due to that Saudi dialect words originally and mostly are extracted from Arabic language words and Arabic language is a morpho¬logical language and their words might be varied depending on the presence and position of some well-known letters in a word. Moreover, some of these letters come at the beginning (prefixes) or end (suffixes) of a word. Furthermore, these letters also have different shapes depending on their word appearance in the text or context. For instance, (A) letter in Arabic (ا) which pronounce (Alif) might be drawn or written in four ways (آ ، إ ، أ ، ا). This has its effects in (The) in Arabic which is a combination of two letters "ا" and "ل" to give us in its turn (The) in four forms (ال ، ألَ ، إل ، آل) [21].

And because this letters is frequently repeated in most Saudi dialect words, we have counted and confined most of these repeated letters (prefixes) and (suffixes). Prefix letters contain ("ال", "آل", "إل", "أل", "ا") and Suffix letters are (" ", "ة", "ه", "ني", "نهن", "نهم", "وهم", "وهن", "هم", "هن", "ين", "ون", "وا"). We applied the first algorithm (see Fig. 3) to remove these letters (prefixes and suffixes) from most of lexicon words and saved all these words in new lexicon with words have no (prefix or suffix) letters. Next step we applied the second algorithm (see Fig. 4) which used the new lexicon with words have no (prefix or suffix) letters. The second algorithm used words from new lexicon and added (prefix or suffix) letters (one after one) step by step on these words and compared it to the texts taken from Twitter. In other words, we applied the algorithm to each word to give us all possible options by adding firstly (prefixes) alone to the same word one by one and then compared it one by one with all Twitter dataset's text and then added (suffixes) to the same word and compared them (one by one) with all dataset's text again and then added some of (prefixes and suffixes) together one by one to the same word and compared them with the all dataset's text and this has been repeated until last word in the new SaudiSentiPlus lexicon (see Fig. 4) . For example, the word "غبي" which means "stupid for male" when we applied light stemming algorithm the chances of finding or matching this word against the text taken from Twitter is increased. Since that this algorithm, for instance, will add the letter "ة" as a suffix to the word "غبي" to give us new word "غبية" which means "stupid for female".

III. Evaluation

To evaluate the effectiveness of SaudiSentiPlus, we compared it with SauDiSenti. We built our testing dataset from twitter by focusing on Saudi dialect hashtags (971 thousands tweets from 162 hashtags). We asked three annotators to classify the dataset's tweets randomly and manually to three classifications (positive, negative, and neutral). All the annotators are Saudi and Arabic native speakers and two of them are Arabic language teachers. They labeled 300 tweets for each classification.

These tweets have been used to evaluate the lexicon's accuracy for SaudiSentiPlus and SauDiSenti. Four of the most

widely used accuracy measures in the literature are utilized ([17, 13]). They are precision (P), recall (R), F measure (F), and accuracy (Acc) and their mathematical equations are as follow:

Precision (P) = TP / (TP+ FP)

Recall (R) = TP / (TP + FN)

Accuracy (Acc) = (TP + TN) / (TP + FP +TN +FN)

F-Score (F) = 2TP / (2TP + FP + FN)

Where TP or True Positive indicates to number of tweets that are correctly predicted as a positive, TN or True Negative are number of tweets that are correctly predicted as a negative, FP or False Positive indicates to number of tweets that are incorrectly predicted as a positive, FN or False Negative are number of tweets that are incorrectly predicted as a negative.

IV. Results and Discussion

The purpose of this experiment is to study the effect of increasing the size of the lexicon and to find whether there is any effect when we applied the first (see Fig. 3) and the second algorithms (see Fig. 4). As aforementioned above, these two lexicon based algorithms were developed to deal with (prefixes and suffixes) letters in order to increase performance of proposed Saudi dialect lexicon.

Table II illustrates the performance results of the experiment. Better accuracy (74%) has been achieved when the lexicon size is increased (from 4431 words to 7138 words). Moreover, accuracy has been increased to reach (81%) when we applied the two algorithms with the lexicon-based approach. Table II lists the precision, recall, accuracy, and F-Score results of the experiment.

As aforementioned, the lexicon construction accomplished through four phases. In the first phase, the lexicon was at its smallest size with 4554 words taken from automatic translation of English sentiment lexicons that already created by two prior studies [15] and [16] and other more sentiment Saudi dialect words which were manually extracted from the twitter data (datasets). The lexicon (SaudiSentiPlus 1) performance or its accuracy reached 61% which is better than 54% of SauDiSenti with its 4431 words [17]. In the second phase, or (SaudiSentiPlus 2) as shown in Table II, no new words have been added to the lexicon however we applied the two lexicon based algorithms on the lexicon (SaudiSentiPlus 1) to yield better accuracy (68%).

In the third phase, more words (4431 words) have been taken from Saudi dialect sentiment lexicon (SauDiSenti) for sentiment analysis of Saudi dialect tweets [17] and next we deleted the repeated words and divided all these words based on their polarities to reach to 7139 words. In this stage (see SaudiSentiPlus 3 in Table II) the accuracy has been enhanced to reach to (74%). Finally, we noticed that accuracy has been increased to reach (81%) when we applied the two algorithms with the lexicon-based approach (see SaudiSentiPlus 4 in the Table II).

Input: lexicon L

Output: new lexicon NL with words have no (prefix or suffix) letters

1. Create array of new lexicon's words NLW
2. initialize Prefix = ("آل", "إل", "أل", "ال")
3. initialize Suffix = ("نهن", "ني", "ه", "ة", "وا", "ون", "ين", "هم", "هن", "وهن", "وهم", "تهم")
4. **for** each w ∈ L **do**
5. **if** w contains any (prefix or suffix) letters **then**
6. **Remove** (prefix or suffix) letters
7. **Add** new word with no (prefix or suffix) letters to the NLW array
8. **else**
9. **Add** w to NLW array
10. **end if**
11. **end for**
12. **Save** NLW array in NL **file**

Fig. 3. Algorithm#1 to Remove (Prefixes and Suffixes) from most of Lexicon Words.

```
Input: new lexicon NL, set of texts T
Output: Sentiment for each text t ∈ T
initialize Prefix = ("ال", "أل", "إل", "آل")
initialize Suffix = ("ه", "ة", "وا", "ون", "ين", "هم", "هن", "وهن", "وهم", "نهم", "نهن", "ني")
Create a data structure to hold the lexicon. lexicon = dict()
# In this study, we used Python diction. The key of the dictionary was the word
# and the value was the word's score.
for each w ∈ NL do
    lexicon[w] = polarity of the same word (w)
 end for
Create a data structure to hold the tweet's text. tweet = dict()
for each t ∈ T do
   initialize score to 0
   Clean t (remove any non-Arabic letters from the text)
   for each w ∈ NL do
      initialize foundMatch to false
      if w = any word in the text t then
            foundMatch is true
      else if  w with any letter of Prefix  = any word in the text t then
            foundMatch is true
      else if  w with any letter of Suffix  = any word in the text t then
            foundMatch is true
      else if  w with any letter of Prefix and Suffix  = any word in the text t then
            foundMatch is true
      end if
      if   foundMatch  then
            score = score + int(lexicon[w])
       end if
     end for
   tweet['text'] = t
   tweet['score'] = score
   if (score > 0)  then
      tweet['sentiment'] = 'positive'
      num_pos += 1
   else if  (score < 0) then
      tweet['sentiment'] = 'negative'
      num_neg += 1
   Else:
      tweet['sentiment'] = 'neutral'
      num_neu += 1
    End if
End for
```

Fig. 4. Algorithm#2 to Increase the Chances of Finding Saudi Dialect Words.

TABLE. II. PERFORMANCE OF THE SAUDISENTIPLUS LEXICON COMPARED WITH SAUDISENTI

Lexicon	Method	P	R	Acc	F
SauDiSenti (4431 w)	lexicon-based	55%	53%	54%	54%
SaudiSentiPlus 1 (4554 w)	lexicon-based	61%	59%	61%	60%
SaudiSentiPlus 2 (4554 w)	lexicon-based + Algorithms	67%	68%	68%	67%
SaudiSentiPlus 3 (7139 w)	lexicon-based	73%	73%	74%	74%
SaudiSentiPlus 4 (7139 w)	lexicon-based + Algorithms	81%	82%	81%	80%

V. CONCLUSION

Majority of researchers on sentiment analysis field, specifically Arabic lexicon-based approach, are focused on the dataset text preprocessing methods rather than dealing with lexicon itself.

In this study, we think beyond of the box, we constraint firstly on building a new sentiment lexicon with reasonable number of words and then doing adequate preprocessing methods on the lexicon's words in addition to the (Twitter dataset). The study presents Saudi Dialect Sentiment lexicon called SaudiSentiPlus contains 7139 words.

Due to that Saudi dialect words originally and mostly are extracted from Arabic language words and Arabic language is a morphological language and their words might be varied depending on the presence and position of some well-known letters in a word. Moreover, some of these letters come at the beginning (prefixes) or end (suffixes) of a word. Furthermore, these letters also have different shapes depending on their word appearance in the text or context.

In order to increase performance of proposed Saudi dialect lexicon (SaudiSentiPlus) we developed two lexicon based algorithms to deal with (prefixes and suffixes) letters of the lexicon's words (see Fig. 3 and Fig. 4).

The experiment which has been conducted to evaluate the performance of SaudiSentiPlus comprises four phases. The precision, recall, accuracy, and F-Score are measured in every phase. We built our testing dataset from twitter by focusing on Saudi dialect hashtags (971 thousands tweets from 162 hashtags). We asked three annotators to classify the dataset's tweets randomly and manually to three classifications (positive, negative, and neutral). All the annotators are Saudi and Arabic native speakers and two of them are Arabic language teachers. They labeled 300 tweets for each classification.

A comparison has been made among SauDiSenti with its 4431 words [17] and the study proposed lexicon (SaudiSentiPlus). The results, as illustrated in Table II, show that SaudiSentiPlus with the two lexicon-based algorithms achieved 81% accuracy which outperformed SauDiSenti with its 54% of accuracy.

REFERENCES

[1] Alghaith, W. (2015). Understanding Social Network Usage: Impact of Co-Presence, Intimacy, and Immediacy. International Journal of Advanced Computer Science and Applications (IJACSA), 6(8), 99-111.

[2] Alharbi, A, & Donckera, E. (2019). Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. Cognitive Systems Research, 54 (5), 50–61.

[3] Arab Social Media Report. (2014). Citizen Engagement and Public Services in the Arab World: The Potential of Social Media. Mohammed bin Rashid School of government, 1(6). Retrieved from Arab Social Media Report Website: http://www.arabsocialmediareport.com/.

[4] Statista. (2019, July). Leading countries based on number of Twitter users as of July 2019 (in millions). In Statista - The Statistics Portal. Retrieved from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ In-text Citation: (Statista, 2019).

[5] Adayel, H. & Azmi, A. (2016). Arabic tweets sentiment analysis – a hybrid scheme. Journal of Information Science, 42(6), 782–797.

[6] Internet World Stats. (2019, July). Top Ten Languages Used in the Web - April 30, 2019 (Number of Internet Users by Language). In Internet World Stats Portal. Retrieved from https://www.internetworldstats.com/stats7.htm/ In-text Citation: (Internet World Stats, 2019).

[7] Boudad, N., Faizi, R., Haj Thami, R., & Chiheb, R. (2018). Sentiment analysis in Arabic: A review of the literature. Ain Shams Engineering Journal, 9 (2018) 2479–2490.

[8] Darwish, K., & Magdy, W. (2014). Arabic Information Retrieval. Foundations and Trends in Information Retrieval, 7(4):239–342.

[9] Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., & Abdullah, S. (2018). Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis. Journal of Information Science, 44(3), 345–362. https://doi.org/10.1177/0165551516683908

[10] Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. Journal of Information Science, 44(4), 491–511. https://doi.org/10.1177/0165551517703514

[11] Abdul-Mageed, M, & Diab, M. (2014). SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In LREC (pp. 1162–1169).

[12] Ibrahim, H., Abdou,S. & Gheith, G. (2015). Automatic expandable large-scale sentiment lexicon of modern standard Arabic and colloquial. In 2015 First International Conference on Arabic Computational Linguistics (ACLing) (pp. 94–99). Cairo, Egypt.

[13] Assiri, A., Emam, A., & Al-Dossari, H. (2018). Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. Journal of Information Science 44(2): 184–202.

[14] AlNegheimish, H., Alshobaili, J., AlMansour, N., Bin Shiha, R., AlTwairesh, N., & Alhumoud, S. (2017). AraSenTi-Lexicon: A different approach. In International Conference on Social Computing and Social Media (pp. 226–235). Springer, Cham.

[15] Hu, M. and Liu, B. (2004). Mining and Summarizing Customer Reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004). Seattle, Washington: ACM DL.

[16] Liu, B., Hu, M. and Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In: Proceedings of the 14th International World Wide Web conference (WWW-2005). Chiba, Japan: ACM DL.

[17] Al-Thubaity, A., Alqahtani, Q., & Aljandal, A. (2018). Sentiment lexicon for sentiment analysis of Saudi dialect tweets. In: Proceedings of the 4th International Conference on Arabic Computational Linguistics (ACLing 2018). Dubai, UAE: Procedia Computer Science 142. 301–307.

[18] Duwairi, R., & El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. Journal of Information Science, 40(4), 501–513. https://doi.org/10.1177/0165551514534143

[19] Froud, H., Lachkar, A., & Ouatik, SA. (2012). A comparative study of root-based and stem-based approaches for measuring the similarity between Arabic words for Arabic text mining applications. Advanced Computing: An International Journal, 3(6), 55–67.

[20] Althobaiti, M., Kruschwitz, U., & Poesio, M. (2014). AraNLP: a Java-based library for the processing of Arabic text. In: Proceedings of the international conference on language resources and evaluation (LREC). Reykjavik, 26–3.

[21] Al-Kabi, M.N., Alsmadi, I.M., Gigieh, A.H., Wahsheh, H.A., & Haidar, M.M. (2014), Opinion mining and analysis for Arabic language. International Journal of Advanced Computer Science and Applications (IJACSA), 5(5), 181-195.