Building a New Tourism Sentiment Lexicon Containing Descriptive Words in Modern Standard and Colloquial Arabic

Mohammed Alkoli
Department of Studies in Computer Science, University of Mysore, Mysuru – INDIA. Mysuru-570006

B. Sharada
Department of Studies in Computer Science, University of Mysore, Mysuru – INDIA. Mysuru-570006

Sami A.M. Alquhali
English Dept., Amran Community College (ACC), Yemen

ABSTRACT
In tourism industry, sentiment analysis is emerging as a technology that can be used to assess the sentiments of the tourists based on their responses on different social media sites or platforms. Sentiment analysis is an important and helpful technique for decision makers to evaluate services and identify problems and deficiencies.

Too many studies have been done on this field in other languages, but in Arabic the number of studies is limited. In addition, such studies on Arabic examine each dialect of Arabic separately and no single study includes sentiment analysis examination of a group or some varieties of Arabic dialects along with Modern Standard Arabic (MSA). The previous studies also do not address the different Arabic dialects, therefore the researches here think there should be a study that includes sentiment analysis of a number of Arabic dialects along with Modern Standard Arabic, the current research paper is an example, as most Arabs (Arab people) express their opinions in their dialects of Arabic and few use Modern Standard Arabic.

The main goal of this research paper is to build a new sentiment analysis lexicon based on the opinions of Arab tourists visiting India. This lexicon includes lexemes (words or vocabularies) taken of three Arabic dialects (namely: Gulf, Levantine and Egyptian dialects of Arabic) along with Modern Standard Arabic. The lexicon will be also evaluated by comparing it to the existing one, namely SemEval2016, using a machine learning technique called Support Vector Classifier for obtaining better results. Thus, building a new dictionary will be effective in sentiment analysis in modern Arabic and most Arabic dialects.

Keywords
Arabic Dialects; Modern Standard Arabic; Sentiment Analysis; Tourism; Support Vector Classifier.

1. INTRODUCTION
The economic factor is the main concern of governments, private and public companies and the process of developing economic aspects is a great challenge to economic prosperity and income improvement, whether for governments or other institutions. One of these economic aspects is tourism, which is considered one of the most important factors for so many countries in the economic. So, the governments and companies need new ways and technologies to improve the tourism and its services.

In tourism industry, sentiment analysis is emerging as a technology that can be used to assess the sentiments of the tourists based on their responses on different social media sites or platforms [1]. So, the main role of sentiment analysis is to analyze the people's opinions that are like written blogs, comments, reviews or tweets, as a comprehensive sentiment, and categorize them as positive, negative or neutral [2], [3]. However, most of the existing studies focused on analyzing people's feelings written in English language and other languages while there are few studies interested in analyzing feelings or opinions and feelings written in Arabic language. Moreover, the existing studies on Arabic language focused on modern standard Arabic (MSA), and it is rare to find a study focusing on some Arabic dialects.

Arabic language has three main varieties: Classical Arabic; which is the language of the Qur'an (Islam's Holy Book); Modern Standard Arabic (MSA) and colloquial Arabic. MSA the most eloquent Arabic language variety used in writing and in most formal speech. Colloquial Arabic refers to all spoken and written varieties spoken in daily communication. These colloquial varieties vary from one Arab country to another and from one region of the same country to another [4]. Furthermore, most of the previous studies randomly classified Arabic colloquial dialects into five dialects; Gulf, Iraqi, Levantine, Egyptian and Maghrebi. There is a difference between these colloquial dialects. Sometimes, it is hard for the speakers of one colloquial dialect to communicate with speakers of different dialect, for example speaker of Egyptian Arabic cannot fully understand speakers of colloquial Moroccan Arabic and therefore, MSA is used for achieving such communication as it is the most understood form of Arabic by all Arabs[5, 6].

In this research, the main goal is to build an Arabic sentiment lexicon that can be called 'KoSeLex' to be used in social media applications that may help in understanding or realizing of Arab tourists' opinions (the written expressions) apart from their dialects. The main contribution in 'KoSeLex' is providing words expressing feelings in various Arabic dialects as well as Modern Standard Arabic (MSA).

The rest of this paper is organized as follows: section 2 provides an overview of existing Arabic sentiment lexicons, section 3 explains the process of building the presented lexicon, methodology presented in section 4, as experimental results and evaluation are presented in section 5, and finally section 6 concludes this paper and presents future plans.

2. PPREVIOUS WORK
In this section, the researchers discuss the important previous studies related to the field of sentiment analysis with a special

reference to some common existing lexicons of sentiment analysis in Arabic language.

N. Al-Twairesh et al [7] conducted a study on the Saudi dialect, which is considered to be within the Gulf dialect or the dialect of southern Arabia, according to several classifications of Arabic dialects. They collected data from Twitter, tweets posted from Saudi Arabia in particular. While N. A. Abdulla et al [8] addressed lexicon-based approach to sentiment analysis for the Arabic language where they decided to work with tweets which include opinions written in both Modern Standard Arabic (MSA) and the Jordanian dialect, one of the Levantine Arabic dialects. M. Nabil et al [9] collected the data from Twitter, tweets posted by the Egyptian users using Egyptian dialect.

S. R. El-Baltagy and A. Ali [10], presented a case study the goal of which is to investigate the possibility of determining the semantic orientation of Arabic Egyptian tweets and comments given in limited Arabic resources. They proposed a lexicon-based approach (Unsupervised Approach) to establish a sentiment classification of Egyptian dialect texts. The authors achieved good results using the two algorithms on a Twitter dataset (83.8% accuracy) and (63% accuracy).

Regarding building lexicons in Arabic sentiment analysis, with the development of sentiment analysis techniques and their reliance on them in many aspects, some scholars showed interest to build the lexicons that meet the needs required for sentiment analysis. But most of the existing lexicons are introduced in English language and there are only few lexicons in Arabic language which have many shortcomings too. Building new lexicons in Arabic sentiment analysis has followed two main methods: First, they translated the English lexicons of sentiment analysis into Arabic ones. Second, they extracted the polarity terms by applying supervised and unsupervised techniques on Arabic resource.

Samhaa R. El-Beltagy [11], presented NileULex as an Arabic sentiment lexicon which has Arabic words and compound phrases where 45% of the terms and expressions in the lexicon are from Egyptian dialect, while 55% of the words are Modern Standard Arabic. The study-results were of less than 86% in accuracy. Karima Abidi & Kamel Smaili[12] built a lexicon for the Algeria dialect which is a part of Maghrebi dialect. Each entry of this lexicon is composed of a word, written in Arabic script (modern standard Arabic or dialect) or Latin script (Arabizi, French or English). Their proposed method on a test lexicon, in such study, scored 73% in accuracy. Kiritchenko et al [13][14] generated a new lexicon called 'SemEval-2016' where the authors presented several sentiment lexicons that were automatically generated using two different methods: (1) by using distant supervision techniques on Arabic tweets, and (2) by translating English sentiment lexicons into Arabic. The result obtained has accuracy of 66.6%.

3. BUILDING THE LEXICON

The new words have been collected from many sources of social media including Facebook, Twitter and YouTube and websites used by Arab people visiting India. The new lexicon is built to include Modern Standard Arabic and some Arabic dialects (Gulf, Levantine and Egyptian dialects). The researchers have expanded the new lexicon by merging the old lexicon called "SemEval2016" which consists more than 1600 positive and negative words with more than 2650 new positive and negative special words for sentiment that describe the tourists' opinion.

Table 1 below shows only words selected as sample from the words that are used in the new lexicon in Modern Standard Arabic and the other different Arabic dialects under study. These words are spoken by Arab tourists visiting India coming from various Arab countries where these words express different feelings and opinions in their local dialects.

Table 1. Sample Words Used in Sentiment Analysis in MSA and Colloquial Arabic Dialects

Word	Polarity	MSA	Gulf Dialect	Levantine dialect	Egyptian Dialect
Amazing	Pos.	رائع	زين	منيحه	حلوه
Bad	Neg.	سيئ	مدحدره	خربانه	وحشه
Good	Pos.	جيده	طيبه	كويسه	مليحه
Dirty	Neg.	متسخه	مقرفه	مكركبه	مءفنه

4. METHODOLOGY

The dataset has been collected from two sources: Twitter and YouTube. From Twitter, the tweets of Arabic tourists who visited India have been extracted. From YouTube, the comments by Arab tourists who visited India have been extracted. Then, the dataset has been classified using two methods. In the first method, classification was based on the old lexicon called "SemEval2016" as shown in Figure 1. The second method, dataset was based on the new lexicon "KoSeLex" as shown in Figure 2. The collected dataset classification was achieved by using Support Vector Classifier (SVC) to obtain the best accuracy for polarity.

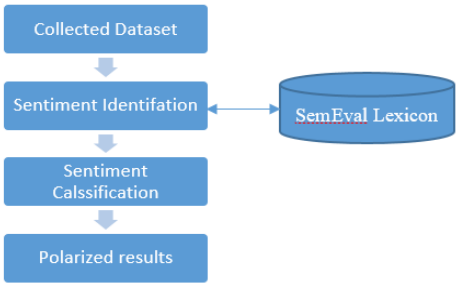

Fig 1: Sentiment Classification based on SemEval2016

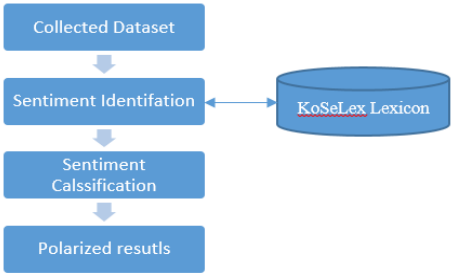

Fig 2: Sentiment Classification based on KoSeLex

5. EVALUATION

Two evaluations have been carried out to determine the best lexicons and accuracy. The KoSeLex lexicon was compared to SemEval2016 lexicon. The first evaluation tested the collected dataset with the SemEval2016 lexicon and the second evaluation has tested the same collocated dataset with the KoSeLex lexicon. In both evaluations, the researchers have used Support Vector Classifier.

Table 2 shows the results of first evaluation using the SemEval2016 lexicon that the system got a low accuracy for

positive polarity with greater than 60%. The accuracy of the negative polarity was less 55%. However, table 3 shows the results of the second evaluation using the KoSeLex lexicon that the system got a high accuracy for positive polarity with greater than 92% and the accuracy for the negative polarity was more than 75%.

Table 2. Results of First Evaluation Using the SemEval2016 lexicon

Polarity	Accuracy
Negative	0.53
Positive	0.645

Table 3. Results of First Evaluation Using the KoSeLex lexicon

Polarity	Accuracy
Negative	0.53
Positive	0.645

The comparison made between the KoSeLex lexicon and the SemEval2016 lexicon showed that there is a big difference in the results. The accuracy in the results of the KoSeLex lexicon is much better. This is because the collected data has specificity in expressing the opinions of Arab tourists written in several Arabic dialects because the SemEval2016 lexicon lacks words expressing opinions of Arab tourists. Table 4 shows a sample of words and how they are classified according to the two lexicons.

Table 4. Sample of Words That Shown In Both Lexicons With Different Polarities

Word	KoArLex	SemEval2016
مدحدره	Negative	Neutral
زينه	Positive	Neutral
مكركبه	Negative	Neutral
رهيبه	Positive	Negative

6. CONCLUSION AND FUTURE WORK

This paper aims at building a new lexicon called KoArLex that is built with many Arabic dialects including Modern Standard Arabic. This lexicon has improved the results in Sentiment Analysis for Arab tourists who write their opinions using different colloquial Arabic dialects. This study used KoArLex lexicon which revealed results with higher accuracy in comparison to the previous ones (SemEval2016), when using the SVC classifier, with a percentage of greater than 92% accuracy.

To sum up, the researchers recommend future work to be done on developing and expanding the dictionary so that it might include the more dialects as well as more dataset or words expressing wider scope of feelings and opinions in order to obtain better and more accurate results.

7. REFERENCES

[1] Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: capitalizing on big data. Journal of Travel Research, 58(2), 175-191.

[2] Altawaier, M. M., & Tiun, S. (2016). Comparison of machine learning approaches on arabic twitter sentiment analysis. International Journal on Advanced Science, Engineering and Information Technology, 6(6), 1067-1073.

[3] Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528).

[4] Boudad, N., Faizi, R., Thami, R. O. H., & Chiheb, R. (2018). Sentiment analysis in Arabic: A review of the literature. Ain Shams Engineering Journal, 9(4), 2479-2490.

[5] Al-Kabi M. N., Gigieh A. H., Alsmadi I. M., Wahsheh H. A., and Haidar M. M., "Opinion Mining and Analysis for Arabic Language", International Journal of Advanced Computer Science and Applications (IJACSA), 5: 181-195, (2014)

[6] Hamed O. and Zesch T. "The Role of Diacritics in Designing Lexical Recognition Tests for Arabic", In: Proceedings of the 3rd International Conference on Arabic Computational Linguistics, ACLing 2017, Dubai, United Arab Emirates, 119-128, (2017)

[7] Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., & Al-Ohali, Y. (2017). Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. Procedia Computer Science, 117, 63-72.

[8] Abdulla, N. A., Ahmed, N. A., Shehab, M. A., & Al-Ayyoub, M. (2013, December). Arabic sentiment analysis: Lexicon-based and corpus-based. In 2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT) (pp. 1-6). IEEE.

[9] Nabil, M., Aly, M., & Atiya, A. (2015, September). Astd: Arabic sentiment tweets dataset. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 2515-2519).

[10] El-Beltagy, S. R., & Ali, A. (2013, March). Open issues in the sentiment analysis of Arabic social media: A case study. In 2013 9th International Conference on Innovations in Information Technology (IIT) (pp. 215-220). IEEE.

[11] El-Beltagy, S. R. (2016, May). Nileulex: A phrase and word level sentiment lexicon for egyptian and modern standard arabic. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 2900-2905).

[12] Abidi, K., & Smaïli, K. (2018, May). An automatic learning of an algerian dialect lexicon by using multilingual word embeddings. In 11th edition of the Language Resources and Evaluation Conference, LREC 2018.

[13] Kiritchenko, S., Mohammad, S., & Salameh, M. (2016, June). Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016) (pp. 42-51).