

Sentiment Analysis for Arabic Language: A Brief Survey of Approaches and Techniques

Mo'ath Alrefai¹, Hossam Faris² and Ibrahim Aljarah³

^{1,2,3}*Business Information Technology Department, The University of Jordan, Amman, Jordan*

¹*muathalrefai89@gmail.com*, ²*hossam.faris@ju.edu.jo*, ³*i.aljarah@ju.edu.jo*

Abstract

With the emergence of Web 2.0 technology and the expansion of on-line social networks, current Internet users have the ability to add their reviews, ratings and opinions on social media and on commercial and news web sites. Sentiment analysis aims to classify these reviews in an automatic way. In the literature, there are numerous approaches proposed for automatic sentiment analysis for different language contexts. Each language has its own properties that makes the sentiment analysis more challenging. In this regard, this work presents a comprehensive survey of existing Arabic sentiment analysis studies, and covers the various approaches and techniques proposed in the literature. Moreover, we highlight the main difficulties and challenges of Arabic sentiment analysis, and the proposed techniques in literature to overcome these barriers.

Keywords: Arabic, Sentiment analysis, Survey

1. Introduction

With the emergence of Web 2.0 technology and rising number of social media websites and web forums, current Internet users have the ability to add their reviews, ratings or opinions on web pages whether it is a commercial website or news website or any other website. Furthermore, people by nature often consult discussion forums before making investment decision by asking friends about their opinions, and they read consumers' reviews before making a big purchase like an electrical device or a car. So, the consumers tend to trust other consumers' opinions rather than advertisements, and this lead to a better decision for the consumer especially in case when he doesn't have a lot of experience or knowledge about the target product [1,2]. Not only from consumers side, the marketers will try to get the bad reviews about their products and report the issues to their developers and designers, so they can meet the public demands.

Extracting manually good or bad reviews for a specific product is extremely cumbersome and time consuming, and this in turn led to the emergence of a new technique for extracting the opinions related to a specific topic which is called Sentiment Analysis (SA). SA refers to use the Natural Language Processing (NLP) and text analysis to extract the sentiment from a given text that is related to specific topic. Sharda *et. al.*, in [3] defined SA as "a technique used to detect favorable and unfavorable opinions toward specific products and services using large numbers of textual data sources". By using sentiment analysis, we can answer some important questions like this "What do the consumers say about this device or this service" by collecting the reviews about the device or the service and then digging into opinions using automatic sentiment analysis tool to get the answer.

Received (April 18, 2018), Review Result (July 5, 2018), Accepted (July 14, 2018)

Arabic is currently ranked as the fourth language used in the web, and there are about 168 million of Arabic Internet users. Most of sentiment analysis works focus on English language and there is relatively less work on Arabic language compared with English. The complexity of Arabic language and the lack of the available resources of Arabic sentiment analysis like lexicons and datasets are the main obstacles in Arabic sentiment analysis. In the literature on SA, there are two major approaches to build SA system, machine learning approach and lexicon-based approach, also, in some literatures we found the two approaches were combined to build a hybrid one. In this work we review the existing Arabic Sentiment Analysis studies and cover the various approaches proposed in literature. In addition, we highlight the main challenges of Arabic Sentiment Analysis.

The rest of this work is organized as follows: Section 2 provides a brief background on Sentiment Analysis approaches; Section 3 talks about the main challenges in Arabic Sentiment Analysis; Section 4 reviews Sentiment Analysis process; Section 5 contains the conclusion and future work.

2. Challenges of Arabic Language

Arabic is one of the six official languages of the United Nations and it is a morphologically rich language. This language has two main forms, standard and dialectal. Modern Standard Arabic (MSA) used in formal speeches and writing like book and newspapers while dialect Arabic used in informal writing specially in social media and it vary from one country to another. Analyzing Arabic text is very complicated, the following points explain the complexity of applying sentiment analysis on Arabic language.

- Every Arabic country has its own dialect that vary from other country, and people tend to use their dialect instead of using MSA.
- The same word can be written by different users in different ways like the words that ends with Ta' marbootah (ة), for example, (المؤثرة) can be written as (المؤثره).
- The negation words can invert the meaning of the sentence, so if the sentence had a positive sentiment then after adding the negation word in the sentence then it will have had a negative sentiment. In Natural Language Processing (NLP), negation words are considered stop words and it will be removed from the sentence as a preprocessing step.
- The same verb can be written in different ways based on the subject of the verb, singular or plural, feminine or masculine. For example, " هو يحب السيارات " (He likes cars) and " هي تحب السيارات " (She likes cars).
- Some Arabic names are derived from adjectives, and the name itself does not have a sentiment while the adjective may have a sentiment. For example, the name Jameelah and the adjective Beautiful have the same form in Arabic (جميلة).
- Arabic speakers could use idioms to express their opinions, and the idioms have implicit opinion. For example, the idiom حسبي الله ونعم الوكيل has a negative opinion while it does not have any negative word.

3. Sentiment Analysis Approaches

Sentiment analysis is considered as text mining techniques that can be used to detect favorable and unfavorable opinions toward specific services or products, and it can be applied on different areas such as brand Management, politics, healthcare, financial markets, or in e-mail filtration by prioritizing received e-mails.

[4, 10]. Khoja stemmer [11] is an open source for stemming Arabic text, and it has already used in many Arabic works.

- Emoticons converter: replacing the emoticon with their respective meaning in language.
- Dialect converter: mapping the slang word to MSA.
- Negation and intensification: detecting negation and intensification phrases in dataset. This technique is very important for sentiment extraction in order to give the opposite polarity in case of negation and to give more weights to the current polarity in case of intensification. This technique was used in [4, 7], *e.g.*, negation: هذا المطعم غير جيد the word غير will reflect the output polarity knowing that the word جيد will get a positive polarity but the final output should be negative, intensification: هذا المطعم جيد جدا the word جدا will increase the weight of word جيد.
- Part-Of-Speech (POS) tagging: Getting the type of the word such as noun, verb, adjective, *etc.*, and give the word of these types more weight. This technique was used in [9].
- Feature extraction: For training machine learning classifiers, the annotated data must converted to feature vector, and a combination of specific features will give a specific class. The single feature is produced by converting a piece of text to a feature. In sentiment analysis studies, there are many features have been used, the most common features: Term Frequency, Term Unigrams, Parts Of Speech (POS), and Negation. And to make the job easier, many tools have been used in Sentiment Analysis researches like Rapidminer [12] as in [4], MATLAB [13] as in [14], R [15] as in [16], and WEKA [17] as in [18].

3.2. Sentiment Classification

The third step in Sentiment analysis process depends on the used approach for classification. As mentioned in the introduction, sentiment classification approaches can be categorized into machine learning approach, lexicon-based approach, and hybrid approach which each of which can be categorized further into sub-categories as shown in Figure 1.

In the following subsections we describe each of the three approaches in details.

3.3.1. The Machine Learning Approach

The machine learning approach involves getting computers the ability to act without being programmed. Computer programs uses the exposed data to detect patterns and then adjust program actions and make intelligent decisions accordingly. In sentiment classification, this approach relies on the usage of famous machine learning techniques on the text. Most of researches on Arabic sentiment analysis

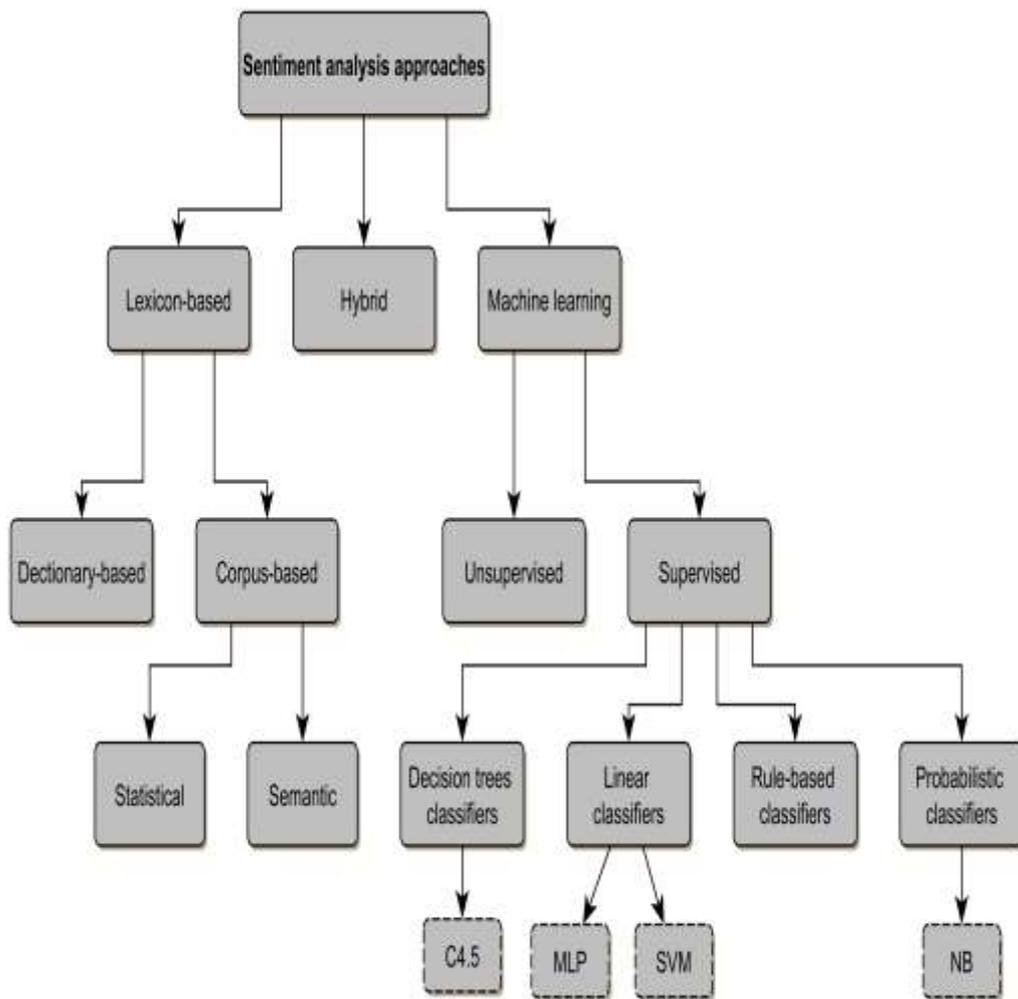


Figure 1. Classification of Sentiment Analysis Approaches

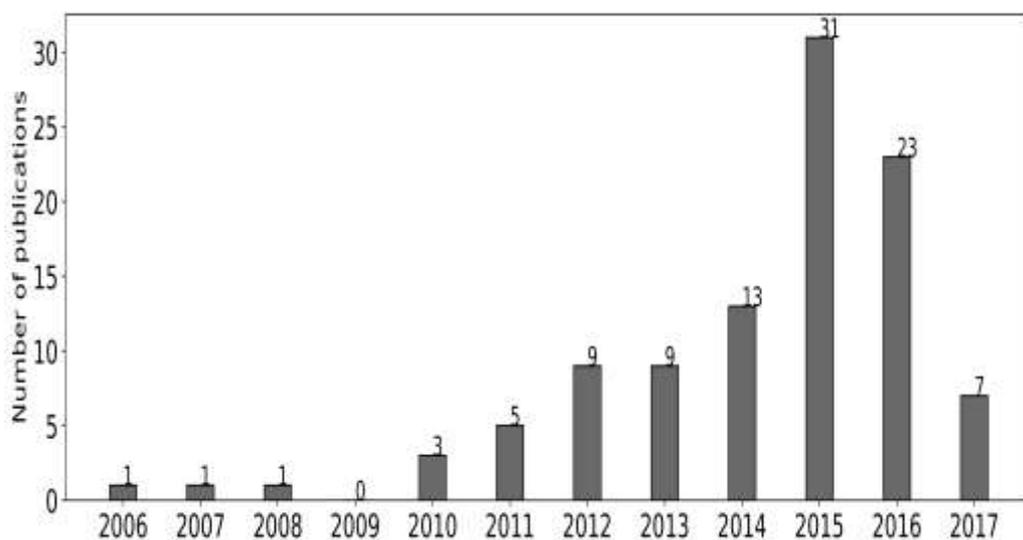


Figure 2. Count of Surveyed Publications of Sentiment Analysis for Arabic Language per Year

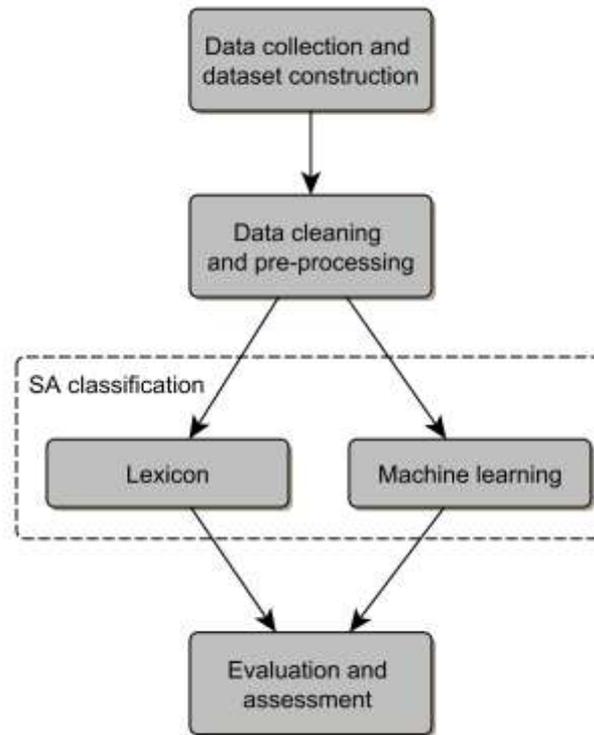


Figure 3. Flowchart of Main Processes in Sentiment Analysis Approaches

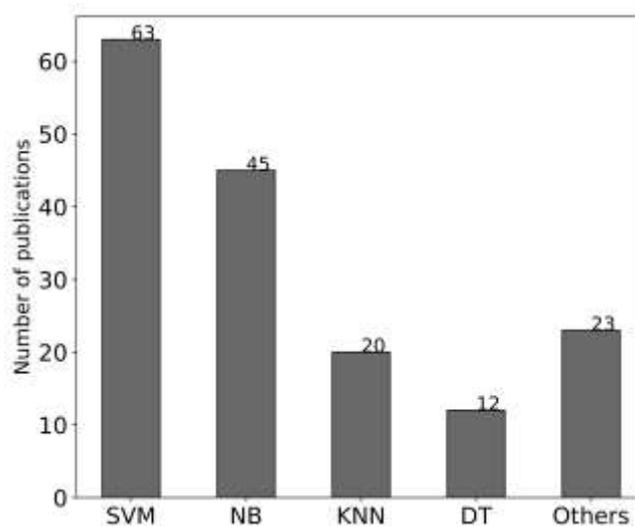


Figure 4. The Count of Publications for Most Applied Machine Learning Classifiers in Arabic Sentiment Analysis

used ML approaches because it was reported that they are more accurate than the lexicon approaches.

Machine learning can be classified into two main categories: unsupervised learning and supervised learning. In unsupervised learning, the learning algorithm does not require annotated documents and it will automatically discover some structure in the given documents. In contrast to unsupervised learning, in supervised learning, the algorithm will use the annotated documents for training. Supervised

learning involves using already annotated documents and machine learning algorithm in order to train a supervised classifier. After preparing and annotating the dataset, it will be divided into training and testing datasets. Training dataset will be used for training the classifier and testing dataset will be used for testing and evaluated the classifier. In this training process, the classifier will learn from the annotated documents and this will allow it to make a prediction about the documents that might come in the future. The supervised learning is widely used for building sentiment analysis system and it can be categorized further into four types: Decision Tree Classifiers, Linear Classifiers, Rule-based Classifiers, and Probabilistic Classifiers.

Based on our survey, we found that the most used supervised classification algorithm in Arabic sentiment analysis is Support Vector Machine (SVM) which belongs to Linear Classifiers category. SVM is an algorithm that transforms training data into higher dimension by using a non-linear mapping function, and within the new dimension, it determines the best linear separator between different classes. It was reported in many studies that SVM outperforms other classifiers [5].

Naive Bayes (NB) which belong to Probabilistic Classifier is the second applied ML classifier in Arabic SA. NB is a statistical classifier based on Bayes' theorem, and it uses the probability theory in order to find the most likely class [19]. It is considered as one of the simplest classifiers.

Figure 4 shows the most commonly used classifiers for Arabic SA and the number of papers that applied each of them. As we can see, there are four main classifiers which are SVM, NB, KNN, and DT. SVM is most applied where 63 papers used SVM which forms around 39% of the total number of papers that applied ML classifiers. NB comes next with 45 papers. KNN and DT are less common with 20 and 12 papers, respectively. Other classifiers were also applied for Arabic SA in few studies like Hybrid SVM & KNN in [20], and Random Sub Space (RSS) & SVM in [21]. All the other classifiers form around 14% of all papers.

3.3.2. The lexicon-based Approach

Lexicon-based approach depends on lexicon which contains a collection of sentiment words, each word has a polarity value, positive words have values greater than zero, negative words have values less than zero, and any word that does not exist in the lexicon is considered as a neutral word. For example, a sentiment classification task can be performed based on this approach by looking for sentiment words in a given text or a document then adding weights or tags to these words and then counting the weights and tags to detect the overall sentiment. Sentiment words list with its polarity value exist in a sentiment lexicon, and to prepare the sentiment lexicon there are two approaches: dictionary-based approach and corpus-based approach [22].

Dictionary-based approach works as follows, starting from initial seed set of sentiment words with known positive and negative orientation. Then exploit available thesaurus and corpora like WordNet [23] to find synonyms and antonyms for each word in the list. The newly found word are added to the seed list and the next iteration starts. The process will have completed when no new words can be found. The major disadvantage of this approach which that it does not find opinion words with domain orientation, for example, if we say: the phone speaker is quiet, this indicates a negative opinion, but if we say: the car is quiet, this indicates a positive opinion.

Unlike the dictionary based approach, corpus-based approach can find domain and context specific opinion words. This approach rely on statistic or syntactic patterns with seed list of opinion words with known polarity to find new sentiment words with their polarity in a large corpus. For statistical pattern, the new sentiment

word can be found by its occurrence frequency in a large annotated corpus, so if the word appears more frequently in positive documents than the negative documents then it will be added to words list as a positive word, and if it appears more frequently in negative document then it will be added as a negative word. In other words, the word will be added as a positive word if it occurs more frequently in a positive document, and it will be added as a negative word if it occurs more frequently in a negative document. In syntactic pattern, the words with similar opinion appear together in the corpus, and this pattern suppose that if the words appear frequently together within the documents then they are likely having the same polarity. So the word that does not have a polarity value and it appears frequently with another word with known polarity then it should have the same or opposite polarity value of the known word based on the connected word between them like word (AND), for example, the word (spacious) in this sentence (the car is comfortable and spacious) will have the same polarity value of the word (comfortable).

For constructing and expanding Arabic lexicon, there are two ways to do that in the literature [8, 24]: Manual and Automatic. In manual techniques, the lexicon is built by translating the content of SentiWordNet [25] or SentiStrength [26], and for each translated word finds its synonyms and add the word with its synonyms to lexicon. In automatic technique, the lexicon is built by starting from manually collected and annotated lexicon (base lexicon) and then increase the lexicon size by adding synonym and antonyms. The manual method for lexicon creation is more accurate than the automatic one, but the automatic method required less time and work, and according to the study [7], the lexicon with bigger size has a better accuracy, so the lexicon should contain most possible sentiment words in order to increase the accuracy of the classifier that uses the Lexicon.

Lexicon-based approach is domain independent and the lexicon is built for all domains on the contrary of the ML approach which does not work well on data that differ from the training data. Also, in this approach there is no need to have an annotated data for training process.

An interesting example of a work that applied both approaches the corpus-based and the lexicon-based approach for Arabic SA was conducted by Abdulla et al. in [7]. In order to build the lexicon, they collected 3479 negative and positive words, positive words are weighted with +1 value and negative words with -1 value. The lexicon was built in three phases, in each phase they increased number of words in the lexicon, and they proved that the bigger lexicon will get better results. The text was classified based on the total weights, if the total is greater than zero then it is positive, and if it is equal to zero then it is neutral and otherwise it is considered as a negative. The achieved accuracy using the lexicon was 58.6 %. For the corpus-based approach, 2000 labeled tweets were collected by using a tweet crawler, 1000 positive and 1000 negative, and after applying the preprocessing steps they used four classifiers: SVM, NB, D-Tree, and KNN, the best result was 87.2 % by using SVM classifier which is better than the lexicon based accuracy.

Although the lexicon based approaches have their own advantages, however, they still suffer major drawbacks such as lack of availability, detecting the idioms, limited size, and requiring to a large amount of linguistic resources. Therefore, many researchers proposed different methods and techniques to overcome these drawback. For example, in [14] the authors proposed combining the available resources: English SentiWordNet (ESWN), Arabic WordNet (AWN), and the Standard Arabic Morphological Analyzer (SAMA), the combination has done by merging two lexicons in order to create a Large Scale Arabic Sentiment Lexicon (ArSenL), the first lexicon was created by matching Awn to ESWN and the second lexicon was developed by matching lemmas in the SAMA lexicon to ESWN,

ArSenL has 28,780 lemmas and it is publicly available. Another example was presented in [27] where the authors presented a large scale Arabic idioms/proverbs sentiment lexicon of MSA and colloquial for sentiment analysis tasks, the idioms/proverbs lexicon was manually collected and annotated, it contained 3632 idioms and proverbs, the authors have proved that the idioms/proverbs lexicon can improve the sentiment classification process.

3.4. The Hybrid Approach

The hybrid approach combines both the machine learning and the lexicon-based approaches. This approach has been reported as a better approach than the machine learning and the lexicon-based approaches. The combination can be done by adding some extracted features from lexicon into machine learning classifier as well as other features such as sentence-level and linguistic features. Another way to combine both approach is by removing each word that does not exist in the lexicon from all instances so only sentiment words will be remaining. The removed words do not affect the sentiment but keeping them will confuse the classifier and decrease the accuracy. Therefore, by removing these words the accuracy is expected to improved.

Compared with previous approaches, few works have used the hybrid approach for SA for Arabic language. For example, the work in [28] has presented a hybrid sentiment analysis approach for classifying Egyptian dialect tweets. In this work, the authors have combined the two sentiment analysis approaches, Machine learning and lexicon-based approaches. In Machine learning approach, three features were used: unigram, bigram and tira-gram. And for lexicon-based approach, they have collected the sentiment words from their corpus and gave each word a weight based on its frequency. The combination between the two approaches was done by adding the sentiment words as a features into Machine learning approach. And in the results, the accuracy of the hybrid approach was better than each single one.

Another work in this research line was presented in [29]. This work presented a hybrid approach for Arabic opinion question answering by adding lexicon features to three classifiers, SVM, NB, and KNN. After preparing the data and applying different pre-processing techniques they have tested the three classifiers without lexicon features, and after that they have tested the same classifiers with lexicon features. In all experiments there was a remarkable enhancement by adding lexicon features to the three classifiers.

Another interesting work can be found in [18]. This work presented a hybrid sentiment analysis approach by removing all unsentimental words from training dataset so the training dataset contained sentimental words, instances and their label. The authors in this work have used two classifiers, SVM and KNN, and they have compared the proposed approach with the machine learning approach by using the same classifiers. The results have confirmed that the proposed approach has better accuracy than the machine learning approach.

4. Conclusions and Possible Research Avenues

Arabic sentiment analysis is considered one of the growing fields that concern with uncovering the feelings of the individuals toward specific event, brand, or something else. It is clear that the field of Arabic Sentiment Analysis is in the early stage and the research on this field has rapidly increased over the last few years.

In this survey, we have investigated the Arabic sentiment analysis, and made comprehensive literature review about this important topic. All the previous studies are categorized into three main approaches: machine learning approach, lexicon-based approach, and hybrid approach. We found that most of researchers have

focused on machine learning approach by applying different classification models such as SVM, and NB, which are the most used algorithms in the literature. The main disadvantage of the machine learning approaches is the domain dependent, which does not work well on data that differ from the training data. On the other hand, lexicon based approaches resolve this issue, where there no need for training data and it is considered a domain independent. Lexicon approach also has some drawbacks such as lack of availability, detecting the idioms, limited size, and requiring to a large amount of linguistic resources. Moreover, there is no stemming algorithms for dialect words of Arabic and Arabic dialects has more stopwords than the standard.

To resolve the shortcomings of the previous approaches, some researchers proposed hybrid approaches, which combine machine learning and lexicon approaches by adding some extracted features from lexicon into machine learning classifiers or by removing all words from the dataset, where these words do not exist in the lexicon. This approach has been reported as a better approach for Arabic Sentiment Analysis. In spite of this advantage, the number of publications that investigated the hybrid approach for Arabic language is much less than the other approaches.

Regarding the lack of available datasets, most of works used tweets to do their experiments, and other works have used OCA dataset [5]. Therefore, the research on SA for Arabic language lacks the publications related to other contexts and social networks.

WordNet, SentiWordNet, and Multi-Perspective Question Answering (MPQA) [30] are the most common linguistic sources used in English Sentiment Analysis, and building linguistic resources for Arabic language is still needed.

References

- [1] S. E. Shukri, R. I. Yaghi, I. Aljarah and H. Alsawalqah, "Twitter sentiment analysis: A case study in the automotive industry", Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on IEEE, (2015), pp. 1-5.
- [2] H. Hakh, I. Aljarah and B. Al-Shboul, "Online social media-based sentiment analysis for us airline companies", New Trends in Information Technology, (2017), pp. 176.
- [3] R. Sharda, D. Delen, E. Turban, J. Aronson and T. Liang, "Business Intelligence and Analytics: Systems for Decision Support", Always learning, Pearson Education, Limited, (2014).
- [4] R. M. Duwairi, R. Marji, N. Sha'ban and S. Rushaidat, "Sentiment analysis in arabic tweets", 2014 5th International Conference on Information and Communication Systems (ICICS), (2014) April, pp. 1-6.
- [5] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López and J. M. Perea-Ortega, "Oca: Opinion corpus for arabic", J. Am. Soc. Inf. Sci. Technol., vol. 62, (2011) October, pp. 2045-2054.
- [6] M. Nabil, M. A. Aly and A. F. Atiya, "Astd: Arabic sentiment tweets dataset", EMNLP, (2015), pp. 2515-2519.
- [7] N. A. Abdulla, N. A. Ahmed, M. A. Shehab and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based" 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), (2013) December, pp. 1-6.
- [8] N. Abdulla, R. Majdalawi, S. Mohammed, M. Al-Ayyoub and M. AlKabi, "Automatic lexicon construction for arabic sentiment analysis", 2014 International Conference on Future Internet of Things and Cloud, (2014) August, pp. 547-552.
- [9] D. Al-Hajjar and A. Z. Syed, "Applying sentiment and emotion analysis on brand tweets for digital marketing", 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), (2015) November, pp. 1-6.
- [10] R. Duwairi and M. El-Orfali, "A study of the effects of preprocessing strategies on sentiment analysis for arabic text", J. Inf. Sci., vol. 40, (2014) August, pp. 501-513.
- [11] S. Khoja, "Shereen khoja- research", (2012).
- [12] R. Klinkenberg, "RapidMiner: Data mining use cases and business analytics applications", Chapman and Hall/CRC, (2013).
- [13] M. U. Guide, "The mathworks", Inc., Natick, MA, vol. 5, (1998), pp. 333.
- [14] G. Badaro, R. Baly, H. Hajj, N. Habash and W. El-Hajj, "A large scale arabic sentiment lexicon for arabic opinion mining", ANLP 2014, vol. 165, (2014).

- [15] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics", *Journal of computational and graphical statistics*, vol. 5, no. 3, (1996), pp. 299-314.
- [16] S. Alhumoud, T. Albuhaireh and W. Alohaideb, "Hybrid sentiment analyser for arabic tweets using r", 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), vol. 01, (2015) November, pp. 417-424.
- [17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The weka data mining software: An update", *SIGKDD Explor. Newsl.*, vol. 11, (2009) November, pp. 10-18.
- [18] S. Alhumoud, T. Albuhaireh and M. Altuwaijri, "Arabic sentiment analysis using weka a hybrid learning approach", 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), vol. 01, (2015) November, pp. 402-408.
- [19] J. Han, J. Pei and M. Kamber, "Data Mining: Concepts and Techniques", *The Morgan Kaufmann Series in Data Management Systems*, Elsevier Science, (2011).
- [20] W. Cherif, A. Madani and M. Kissi, "Towards an efficient opinion measurement in arabic comments", *Procedia Computer Science*, vol. 73, (2015), pp. 122-129.
- [21] A. Ziani, N. Azizi and Y. T. Guiyassa, "Combining random sub space algorithm and support vector machines classifier for arabic opinions analysis", *Advanced Computational Methods for Knowledge Engineering*, Springer, (2015), pp. 175-184.
- [22] Z. Hailong, G. Wenyan and J. Bo, "Machine learning and lexicon based methods for sentiment classification: A survey", 2014 11th Web Information System and Application Conference, (2014) September, pp. 262-265.
- [23] G. A. Miller, "Wordnet: a lexical database for english", *Communications of the ACM*, vol. 38, no. 11, (1995), pp. 39-41.
- [24] N. Abdulla, R. Majdalawi, S. Mohammed, M. Al-Ayyoub and M. AlKabi, "Automatic lexicon construction for arabic sentiment analysis", 2014 International Conference on Future Internet of Things and Cloud, (2014) August, pp. 547-552.
- [25] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining", *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, (2006), pp. 417-422.
- [26] M. Thelwall, K. Buckley and G. Paltoglou, "Sentiment strength detection for the social web", *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, (2012), pp. 163-173.
- [27] H. S. Ibrahim, S. M. Abdou and M. Gheith, "Idioms-proverbs lexicon for modern standard arabic and colloquial sentiment analysis", *International Journal of Computer Applications*, vol. 118, no. 11, (2015), pp. 26-31.
- [28] A. Shoukry and A. Rafea, "A hybrid approach for sentiment classification of egyptian dialect tweets", *Proceedings of the 2015 First International Conference on Arabic Computational Linguistics (ACLing), ACLING '15*, (Washington, DC, USA), IEEE Computer Society, (2015), pp. 78-85.
- [29] K. Khalifa and N. Omar, "A hybrid method using lexicon-based approach and naive bayes classifier for arabic opinion question answering", *Journal of Computer Science*, vol. 10, no. 10, (2014), pp. 1961-1968.
- [30] R. Mihalcea, C. Banea and J. M. Wiebe, "Learning multilingual subjective language via cross-lingual projections", (2007).

