Arabic Sentiment Analysis approaches: An analytical survey

Gehad S. Kaseb, Mona F. Ahmed

Abstract— Web 2.0 has contributed tremendously towards the rapid growth of web contents. People are motivated to develop a system that can identify and classify opinions which are represented in an electronic text. Because of its valuable return in many fields as e-commerce, politics, tourism, etc., Sentiment Analysis (SA) is one of the most active research areas in Natural Language Processing (NLP). Most research efforts in the area of opinion mining deal with English text and little work is done with Arabic text. This paper provides a summarization of the work done in Arabic SA. The paper also presents some challenges and open issues that need to be addressed and explored in more depth in order to improve this field.

Index Terms— Opinion Mining, Sentiment Analysis, NLP, Classification, Machine learning, Lexicon.

1 Introduction

Nowadays, the web has become a read and write platform where users are no longer consumers of information but producers of it as well. User-generated content written in natural language with unstructured free text is becoming an integral part of the web mainly because of the dramatic increase of social network web sites, video sharing websites, online news, online reviews sites, online forums and blogs. Because of this proliferation of user-generated content, web content mining is gaining considerable attention due to its importance for many businesses, governmental agencies and institutions.

The Arabic language is a collection of different variants where there is only one formal written standard variant in the media and education through the Arab world [33]. This variant is called Modern Standard Arabic (MSA), while others are called Arabic dialects. There is a high degree of difference between MSA and Arabic dialects. One interesting fact is that the MSA is not of any Arab's native languages.

MSA is the official language of the Arab world and it is syntactically, morphologically, and phonologically based on classical Arabic [33]. Classical Arabic is the language of the Qur'an (Islam's Holy Book) while Arabic dialects are true native language forms, they are used in informal daily communication and they are not taught in schools or standardized [33]. In contrast to dialects, MSA is usually written not spoken. Arabic dialects are poorly related to classical Arabic. There are many Arabic dialects and they are different in many aspects. One way for dividing Arab dialects is based on the geographic aspect [33] as follows:

- The most common dialect is Egyptian Arabic, which covers the Nile valley (Egypt and Sudan)
- Levantine Arabic covers the dialects of Syria, Lebanon, Jordan, Palestine and Israel.
- Gulf Arabic includes the dialects of Gulf countries (United Arab Emirates, Saudi Arabia, etc.).
- Maghrebi (North African) Arabic which cover dialects of Algeria, Tunisia, and Morocco.
- Iraqi Arabic covers Iraq and combines elements of Levantine and Gulf dialects.
- Yemenite Arabic.

Each dialect group is completely linguistically homogeneous.

Sentiment Analysis (SA) is the study of people's comments, reviews and opinions about a specific object such as an event, an item, a topic, a news feed, a mobile application, or individuals. Sentiment Classification (SC) approaches can be grouped into three main categories: lexicon-based, Machine Learning (ML) and Hybrid approaches. Lexicon based approaches are unsupervised approaches that depend on external lexica to classify sentiments. ML approaches, are mainly supervised approaches that rely on the existence of labeled training documents/phrases. The main classifiers used in ML are Support Vector Machines (SVM), Naïve Bayesian (NB), Decision Trees (DT), K-Nearest Neighbor (KNN), Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), and Stochastic Gradient Descent (SGD). Hybrid approaches are those that combine lexicon and ML techniques.

According to the literature, only little work has been carried out on Arabic language SA. Arabic language is a semitic language with rich morphology. Classical, Modern Standard Arabic (MSA), and colloquial are the three main variants of Arabic. Arabic is the fifth most widely used language in the world, and is the first language of more than 422 million people. The Arabic language is written from right to left and consists of 28 letters with no upper or lower cases.

The field of Arabic SA has been receiving a lot of attention since its rather shy start a decade ago. Recently, many teams have been making significant contributions to this field. The interested reader is referred to the following surveys [1], [2], [35] and [12] to learn more about the work in this field.

Only a few survey papers can be found in the literature summarizing recent research in the field of Arabic SA. The authors of [12] provide a comprehensive survey of existing lex-

- *Gehad S. Kaseb is currently pursuing master's degree in Computer Engineering at Cairo University, Egypt, PH-0201141213594. E-mail: Gehad.Kaseb@gmail.com*
- *Mona F. Ahmed is Assistant professor in Computer Engineering at Cairo University, Egypt,Email: Mona_Farouk@eng.cu.edu.eg*

icon, machine learning, and hybrid sentiment classification techniques for Arabic language. Thirty-two papers were surveyed.

The authors of [1] and [2] provide surveys of the relatively few references detailing the different methods for building Arabic subjectivity and SA systems from the years 2010-2013 and 2004-2012 respectively, at a time when research on Arabic SA was so limited. The authors of [3] present a survey of Arabic tweets SA referencing only nine methodologies. The authors of [35] provide an extended analytical study in SA.

The remainder of this paper is organized as follows: section 2 explores SA related works. Section 3 presents a discussion with the summarization of the presented work. Section 4 shows Arabic language challenges. Section 5 presents concluding remarks. Section 6 discusses the future plans in this field.

2 Related Work

In order to spot new research advancement in SA of Arabic, an exhaustive search process was performed using the Scopus databases. A total of 133 papers were found since 2003. The number of research on Arabic SA has witnessed an enormous increase in recent years; it reached 39 papers in 2016 up to this date. Figure 1 exhibits this increase.

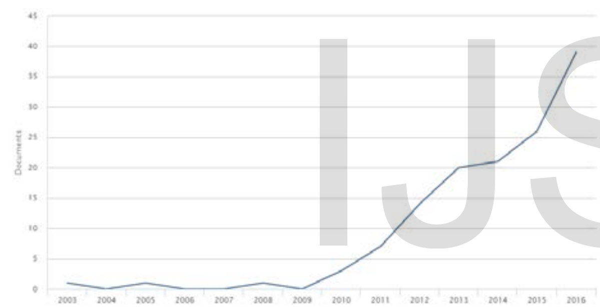


Fig. 1. Research in SA over years [41].

There are 27 papers for Jordon, 18 for Egypt and 17 for Saudi Arabia and USA. Figure 2 shows paper statistics per country

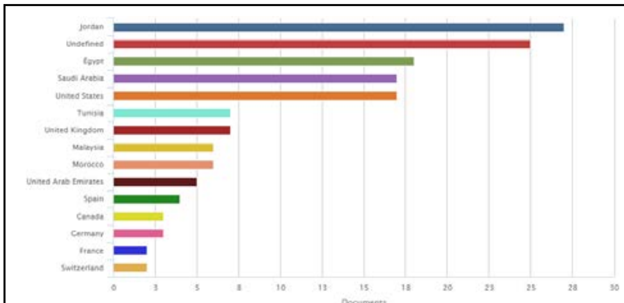


Fig. 2. Research in SA country [41].

Figure 3 shows paper statistics per publishing document type.

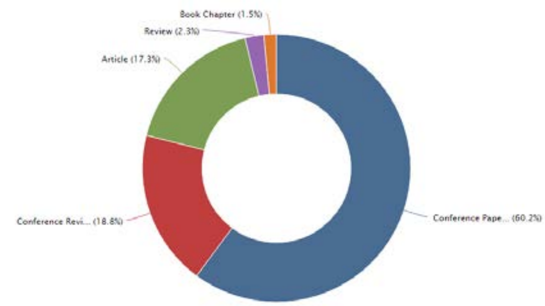


Fig. 1. Research in SA by document type[41].

3 Sentiment Analysis Discussion

It is obvious from the literature that, in contrast to the work in English language, work in Arabic SA is very little, with a lot of potential techniques and approaches still not applied yet. The following tables summarize work - to the best of our knowledge- related to Arabic SA published starting from 2011 with more concentration on much more recent work published in 2016. For simplicity the summary of the analyzed work is formulated in a table focusing on certain important parameters written in the following order:

Work reference.

- The publishing year of this work.
- The sentiment analysis level which can be document or sentence level
- The used Language and also the used dialects
- Whether the article is domain-oriented or not.
- The used polarities; the majority of papers used positive and negative polarities, however, some papers worked with neutral class or objective class and others worked with 5-star rating system.
- The kind of preprocessing and filtering used.
- The type of classification which may be ML, lexicon-based or hybrid.
- The used software if mentioned.
- The context and the dataset size used in experiments.
- The strengths of the applied methodology; it may be the authors opinion of the referenced work or this survey point of view.
- The weaknesses of the applied methodology; it is this survey own point of view.
- The results, the majority of papers use accuracy, however some papers use precision and recall.
- The contributions of this work in the field.
- The suggested future work.

Any empty cells are unknown information for this work.

The author in [5] proposed a hybrid approach that combines three methods.

TABLE 1
SUMMARIZATION OF [5]

Paper	[5]
Publishing year	2011
SA Level	Document level
The Language	Arabic
Domain Oriented	Yes (Multi-domain: Education, Sports, and Politics)
Polarity	Positive & Negative
Preprocessing & Filtering	- Stripping out the HTML tags and non-textual contents - Separating the documents into posts and Converting each post into a single file. - Normalizing some alphabets - Removing some repeated letters - Correcting some of the wrong spelling words - Tokenization - Removing stop words - Applying Arabic light stemmer - Using Term Frequency–Inverse Document Frequency (TF-IDF) weight - Removing some terms with a low frequency of occurrence.
The type of classification	lexicon-based opinion classifier, then Maximum entropy method and finally k-nearest method
Software Used	SentiStrength [24]
The context and the dataset size	Total of 1143 posts which contain 8793 Arabic statements with average of 7.7 statements in each post.
Strengths	The accuracy almost improved from 50% using one method, 60% using two method and 80% using three methods which is a satisfactory performance especially for complex language such as Arabic.
Weaknesses	No Arabic-specific features
Results	Accuracy = 80.29%
Contributions	A combined approach that automatically extracts opinions from Arabic documents that consists of three successive methods; At the beginning, lexicon based method is used to classify as much documents as possible. The resultant classified documents are used as training set for the maximum entropy method which subsequently classifies some other documents. Finally, k-nearest method used the classified documents from lexicon based method and maximum entropy as training set and classifies the rest of the document
Future Work	The experimental results further show that recall and precision of positive documents are better than the negative one. That means further studies should be done for mining of negation of Arabic statements. Also, in the future, it is planned to extend the work to be able to extract features from Arabic opinioned statements

The authors in [8] presented Opinion Corpus for Arabic (OCA) and another one which is the English version of OCA (EVOCA)

TABLE 2
SUMMARIZATION OF [8]

Paper	[8]
Publishing year	2011
SA Level	Sentence Level
The Language	Arabic and English
Domain Oriented	No
Polarity	Positive & Negative
Preprocessing & Filtering	- Tokenization - Removing Arabic stop words - Stemming - Filtering tokens whose length was less than two characters.
The type of classification	SVM, NB using unigrams TF-IDF has been used as a weighting scheme.
Software Used	Rapid Miner [21], PROMT [34] as an online translation
The context and the dataset size	OCA contains 500 reviews in Arabic EVOCA contains 500 reviews in English - 250 positive reviews for each corpus. - 250 negative reviews for each corpus.
Strengths	After accomplishing several experiments using different n-grams models, the obtained results with bi-grams and trigrams were very similar to unigrams. With regard to the machine learning algorithm, it is clear that SVM works better in all cases. In all cases the stemming process gets worse results except when using SVM on the EVOCA corpus. So for the OCA corpus, removing stemmer always improves the results
Weaknesses	Using translation system for building EVOCA; so it's not real reviews
Results	Results with SVM: OCA without stem P=.8699, R=.9480 & F1=.9073 EVOCA with stem P=.9007, R=.8680 & F1=.8840
Contributions	Presenting OCA and EOCA which are freely available for the research
Future Work	

The authors of [16] presented SAMAR, a system for Subjectivity and Sentiment Analysis (SSA) for Arabic social media genres.

TABLE 3
SUMMARIZATION OF [16]

Paper	[16]
Publishing year	2012
SA Level	Sentence
The Language	Arabic (MSA and Dialect Arabic (DA))
Domain Oriented	No
Polarity	Positive , Negative and Neutral
Preprocessing & Filtering	- Tokenization (TOK) - Lemmatization (LEM)
The type of classification	A lexicon manually created of 3982 adjectives. Using SVM light [31]. A two-stage classification approach; first OBJ or SBJ; second Positive or Negative Features used: POS tagging (the reduced tagsets RTS and the extended reduced tag set ERTS) , UNIQUE (Q) feature, Gender, User ID, Document ID, Polarity Lexicon (PL)
Software Used	AMIRA [26] for processing of MSA
The context and the dataset size	DARDASHA (DAR): 2798 chat from maktoob TAGREED (TGRD): 3015 Arabic tweets TAHRIR (THR): 3008 MSA Wikipedia Talk Pages MONTADA (MONT): 3097 forum MSA and DA
Strengths	- Tackling a number of research questions - Exploiting data from four different social media genres.
Weaknesses	- Disregarding the neutral class
Results	Best results in this case - DAR dataset: For Objective/Subjective classification accuracy = 95.83 using TOK+ERTS+PL+Q3 For Positive/Negative classification accuracy = 71.28 using LEM+PL+GEN - TGRID dataset: For Objective/Subjective classification accuracy = 72.52 using LEM+ERTS+PL For Positive/Negative classification accuracy = 65.87 using TOK+ERTS+PL+GEN+LV+UID - THR dataset: For Objective/Subjective classification accuracy = 83.33 using TOK+ERTS +PL+Q3 For Positive/Negative classification accuracy = 67.44 using TOK+PL+GEN+UID - MONT dataset: For Objective/Subjective classification accuracy = 84.19 using LEM+RTS+PL+Q3 For Positive/Negative classification accuracy =81.36 using TOK+PL+Q3
Contributions	Arriving at these conclusions: - Linear kernels yield the best performance. - Adding POS information improves accuracy and F score in most cases - RTS outperforms ERTS with TOK, and the opposite with LEM where ERTS outperforms RTS, however, overall TOK+RTS yields the highest performance of 91.49% F score on subjectivity classification for the DAR dataset. - For the TGRD and THR data sets, TOK+ERTS are equal to or outperform the other conditions on subjectivity classification.
Future Work	Carring out a detailed error analysis of SAMAR in an attempt to improve its performance.

The authors in [4] used statistical ML with term frequency to select features, and then applied SVM and NB classifiers.

TABLE 4
SUMMARIZATION OF [4]

Paper Reference	[4]
Publishing year	2012
SA Level	Sentence level
The Language	Arabic (Egyptian dialect)
Domain Oriented	Yes
Polarity	Positive and Negative
Preprocessing & Filtering	- Removing the user-names - Removing the pictures - Removing the hash tags - Removing the URLs - Removing all non-Arabic words. - Removing stop-words
The type of classification	The feature vectors applied to the classifier contained the term frequency, as using statistical machine learning. First the process starts by extracting all the unigrams and bigrams in the corpus that exceed certain threshold. Ignoreing negation for simplicity in experiments
Software Used	WEKA [30]
The context and the dataset size	1000 tweets consisting of - 500 positive and 500 negative.
Strengths	Developing Egyptian stop words list
Weaknesses	- Using small size dataset - Ignoring negation
Results	Get maximum result using SVM, unigrams only, after removing stop-words - Accuracy = 0.726 - Precision = 0.728 - Recall = 0.726 - F-Measure = 0.725
Contributions	Arriving at these conclusions: - SVM produces more accurate results than the NB. - Regarding the n-gram model, bigram model didn't enhance the results using the unigram model.
Future Work	Improving corpus using techniques such as enlarging or fine-grained annotation Adding some stylistics features, in addition to considering adding some semantic features thus creating a hybrid approach that combines both the ML and SO approaches. Re-tweeting gives a misleading boosting to the weight of the terms in the sentence. The problem of opinion spamming or untruthful opinions could affect the accuracy Building a more comprehensive list of all the positive and negative sentiment words for the Egyptian dialect also negations and valence shifters. Neutral sentiment tweets has to be considered as in real world applications neutral tweets cannot be ignored.

The authors of [13] conducted a study to compare the effectiveness of two free online SA tools.

TABLE 5
SUMMARIZATION OF [13]

Paper	[13]
Publishing year	2013
SA Level	Sentence
The Language	Arabic and English
Domain Oriented	No
Polarity	Positive and Negative
Preprocessing & Filtering	- Removing spammed and noisy comments. - Removing duplicate comments and reviews
The type of classification	Building three dictionaries Arabic, English, Emotions. - The English dictionary has 3,392 words/phrases of which 947 positive, 1100 negative, and 1,345 neutral. - The Arabic dictionary has 1,159 words/phrases of which 427 positive, 306 negative, and 426 neutral. - The Emoticons dictionary has 204 Emoticons of which 71 positive, 70 negative, and 99 neutral. Using NB, SVM, and KNN
Software Used	SocialMention [27] and Twendz [28]
The context and the dataset size	4,050 English/Arabic comments and reviews generated by the users of social network sites.
Strengths	The results indicate that SocialMention is more effective than its counterpart (Twendz) to identify the polarity of each entry in the collected dataset. The conducted results showed that the Naïve Bayes algorithm yields the best results for both SocialMention and Twendz tools.
Weaknesses	Using a small dataset compared to dictionaries.
Results	Using Naïve Bayes algorithm, the Twendz tool yields an accuracy of 45.3% while the SocialMention tool yields an accuracy of 66.2%. Using SVM algorithm yields an accuracy of 43.3% for the Twendz tool, and an accuracy of 65.4% for the SocialMention tool Using K-NN algorithm when K = 1 yields an accuracy of 44.4% for the Twendz tool, and an accuracy of 62.5% for the SocialMention tool.
Contributions	Evaluating the two free online SA tools, by finding the classification accuracy of each tool and developing dictionary-based classifier. Building three dictionaries Arabic, English, Emotions
Future Work	Using a larger dataset, besides testing more free online SA tools.

The authors of [36] designed a sentiment analyzer system for Arabic comments.

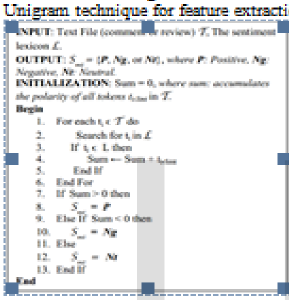
TABLE 6
SUMMARIZATION OF [36]

Paper	[36]
Publishing year	2013
SA Level	Sentence level
The Language	Arabic (Egyptian dialects)
Domain Oriented	Yes
Polarity	Supportive 'y', Attacking 'n', and Neutral 'u'.
Preprocessing & Filtering	- Removing stop words from the comments returning the most important words. - Removing very long comments with number of words -after stop words removal- more than 150 words as about 80% of very long comments in most cases are advertisements for pages on Facebook - Removing the redundant comments; If two comments have similarity value equal to or more than a threshold (0.4), removing the shortest one. - Removing special characters like #, @, !, % and others. - Removing the redundant letters - Segmentation of comments into words, spaces, commas, parenthesis and new line for identifying words .
The type of classification	- Classifiers: Using SVM, NB and DT classifiers - Feature extractors. Three groups of features: 1) Common Words between Post and Comments Features Feature 1: Number of Words in Post Only Feature 2: Number of Words in Comment Only Feature 3: Number of Words Common between Post and Comment 2) All Words in Posts and Comments Features the union of all words in the posts and comments. Each word (feature) takes one of the four values: "C "if the word is not in the post or the comment "M" if the word is in the post only "N" if the word is in the comment only "H" if the word is in both of the post and the comment 3) Negation and Relevance Features Feature 1: Number of Negation Words in Post Feature 2: Number of Negation Words in Comment Feature 3: Relevance with Post
Software Used	WEKA
The context and the dataset size	2400 comments collected from 220 facebook posts 800 neutral, 800 supportive, and 800 attacking.
Strengths	Comparing different machine learning algorithms with different features.
Weaknesses	Counting only five different negation words, whereas there are many more than these
Results	Naive Bayes precision and recall = 59.9%. Decision Tree precision and recall = 69.4% SVM precision and recall = 73.4%.
Contributions	Arriving at these conclusions: - SVM gives the best results - Adding negation words and similarity features for all words in posts and comments gives the best performance.
Future Work	

International Journal of Scientific & Engineering Research, Volume 7, Issue 10, October-2016
ISSN 2229-5518

The authors of [6] compared the lexicon-based and supervised approaches

TABLE 7
SUMMARIZATION OF [6]

Paper	[6]
Publishing year	2013
SA Level	Sentence Level
The Language	Arabic (MSA and the Jordanian dialect)
Domain Oriented	No
Polarity	Positive & Negative
Preprocessing & Filtering	- Correcting misspellings - Removing the repeated letters - Using naive algorithm for repeated letters which simply counts the number of letters in each word. If the number exceeds 5, then it eliminates the repeated letters and looks it up in the MS Word dictionary. - Removing all stop-words. - Normalization process for the letters.
The type of classification	For supervised, SVM, NB, DT, and KNN where K=9. Use lexicon with 3479 words consisting of 1262 positive words and 2217 negative ones. Unigram technique for feature extraction  **INPUT:** Text File (comment or review) $\mathcal{T}$, The sentiment lexicon $\mathcal{L}$. **OUTPUT:** S_{out} = {*P*, *Ng*, or *Nt*}, where *P*: Positive, *Ng*: Negative, *Nt*: Neutral. **INITIALIZATION:** Sum = 0, where sum: accumulates the polarity of all tokens $t_{i,line}$ in $\mathcal{T}$. **Begin** 1. For each $t_i \in \mathcal{T}$ do 2. Search for t_i in $\mathcal{L}$ 3. If $t_i \in$ L then 4. Sum ← Sum + $t_{i,line}$ 5. End If 6. End For 7. If Sum > 0 then 8. S_{out} = *P* 9. Else If Sum < 0 then 10. S_{out} = *Ng* 11. Else 12. S_{out} = *Nt* 13. End If **End** Fig 4. Used Algorithm in [6].
Software Used	MS Word dictionary [22] as a reference for misspelling correction and selecting the first word suggested by it automatically. Khoja stemmer tool [23] as a reference for a list of Arabic stop-words SentiStrength [24] as a reference for 300 seed words translated for lexicon RapidMiner software [21] for supervised experiments
The context and the dataset size	2000 labeled tweets - 1000 positive tweets - 1000 negative tweets
Strengths	In KNN, using K=9 gives the best accuracy than K= 1, 2, 3 and 4 SVM and NB have better accuracy than other classifiers when it comes to text classification. Light-stemming gives better results than other stemming techniques The bigger the lexicon is, the better the results are.
Weaknesses	No neutral class, small lexicon size and no intensive handling for Arabic dialects
Results	- In corpus-based accuracy = 87.2% - In lexicon-based accuracy = 59.6%
Contributions	Building a manually annotated dataset Showing the detailed steps of building the lexicon
Future Work	Enlarging the dataset along with adding a third polarity case (neutral class) Adding strength to the polarity of words that could vary between -5 to +5, which may end up with more accurate outcomes The sarcasm in some tweets always leads to misinterpretation and consequently a wrong polarity classification. Hence, taking into consideration sarcasm absolutely will results in higher accuracy

The authors of [11] developed a framework that determines the polarity of Arabic Tweets.

TABLE 8
SUMMARIZATION OF [11]

Paper	[11]
Publishing year	2014
SA Level	Sentence
The Language	Arabic (MSA/ Jordanian dialect/ Arabizi)
Domain Oriented	No
Polarity	Positive, Negative and Neutral
Preprocessing & Filtering	- Tokenization - Normalization - Filtering Arabic stop-words - Stemming Extensions to Rapidminer: - Emoticons Convertor - Repetitions Remover - Negation Detection - Dialect to MSA convertor - Links Remover - Mentions Remover
The type of classification	Jordanian dialect to MSA parallel dictionary 300 words Negation Dictionary, Arabizi Dictionary SVM, NB, and K-NN
Software Used	Rapidminer [21], MS word [22]
The context and the dataset size	350,000 Arabic tweets 25000+ rated tweets for the training dataset.
Strengths	Building three dictionaries
Weaknesses	The dataset and the dialect dictionary are of small size
Results	SVM accuracy = 71.68% when both stopwords filter and stemming were disabled
Contributions	Handling negations, Arabizi and Arabic dialects.
Future Work	Expanding the dictionaries and solving the memory issue which is inherent in Rapidminer.

The authors in [7] dealt with SA in Arabic reviews from a machine learning perspective.

TABLE 9
SUMMARIZATION OF [7]

Paper	[7]
Publishing year	2014
SA Level	Sentence
The Language	Arabic
Domain Oriented	No
Polarity	Positive and Negative
Preprocessing & Filtering	- Tokenization - Stemming (Arabic) - Filtering Stop-words (Arabic) - Generating n-Grams (Terms) operators n=2
The type of classification	Using NB, SVM and KNN (at K=10) classifiers
Software Used	RapidMiner [21], crowdsourcing tool for annotating dataset
The context and the dataset size	2591 Tweet/Comment - 1073 Positive and 1518 Negative
Strengths	Utilizing crowdsourcing to label the used dataset
Weaknesses	Using small dataset
Results	SVM achieve the best precision = 75.25. KNN achieve the best recall = 69.04
Contributions	Arriving at this conclusion: SVM gives the highest precision while KNN gives the highest Recall.
Future Work	Increasing the datasets. Semi-supervised learning techniques could be used in Arabic SA text as these techniques have been applied successfully to other languages.

The authors in [37] built a collected dataset of Arabic twitter corpora for SSA

TABLE 10
SUMMARIZATION OF [37]

Paper	[37]
Publishing year	2014
SA Level	Sentence Level
The Language	Arabic
Domain Oriented	No
Polarity	Positive, Negative and Neutral
Preprocessing & Filtering	- Normalization - Removing usernames and digits. - Eliminating Latin characters (i.e. URLs, emails).
The type of classification	Overview of annotated feature sets Morphological Feature sets: Diacritic, Aspect, Gender, Mood, Person, Part of speech, State, Voice, having morphological analysis. Syntactic Feature sets: n-grams of words, POS, lexemes including Bag of Words (BOW), Bag of lexemes. Semantic Feature sets: Having positive lexicon, Having negative lexicon, Having neutral lexicon, Having negator. Stylistic Feature sets: Having positive emoticon, Having negative emoticon. Social Signals Feature sets: Having consent, Having dazzle, Having laugh, Having regret, Having sigh
Software Used	
The context and the dataset size	Development dataset: This dataset contains 7,503 random multi-dialectal Arabic tweets. Test data: This dataset contains a total of 1,365 instances.
Strengths	Manually labling dataset which increase the accuracy
Weaknesses	No contextual features to detect sarcasm. Significant drop in performance for sentiment analysis which was found out by error analysis to be due to the noise introduced by the emoticon-based labels.
Results	The overall observed agreement is 91.74% and resulting weighted Kappa reached k= 0.816, which indicates reliable annotations
Contributions	Presenting a gold-standard annotated corpus to support SSA of Arabic twitter feeds, the first publicly available corpus for this task which is released via the ELRA repository.
Future Work	

The authors in [9] proposed the Arabic social sentiment analysis dataset (ASTD)

TABLE 11
SUMMARIZATION OF [9]

Paper	[9]
Publishing year	2015
SA Level	Sentence
The Language	Arabic (MSA and Egyptian)
Domain Oriented	No
Polarity	Objective, subjective positive, subjective negative and subjective neutral.
Preprocessing & Filtering	- Filtering out the non-Arabic tweets - Removing HTML content
The type of classification	MNB, BNB, SVM, SGD and KNN
Software Used	
The context and the dataset size	10,006 Arabic tweets.
Strengths	SVM is the best classifier so it is a reliable choice.
Weaknesses	The small number of subjective tweets it contains.
Results	Accuracy = 0.691 F1 score = 0.626 for SVM trigram unbalanced TF-IDF
Contributions	Presenting the properties and the statistics of the dataset ASTD
Future Work	Increasing the size of the dataset. Discussing the issue of unbalanced dataset and text classification. Extending the generated method either automatically or manually.

The author of [14] investigated SVM, NB, KNN and DT on Arabic Twitter corpus.

TABLE 12
SUMMARIZATION OF [14]

Paper	[14]
Publishing year	2015
SA Level	Sentence Level
The Language	Arabic
Domain Oriented	No
Polarity	Positive, Negative and Neutral.
Preprocessing & Filtering	- Removing all the user identifiers. - Removing Arabic function words - Removing pictures. - Removing Arabic stop words - Removing non-Arabic words - Removing digits and punctuation marks. - Normalization some Arabic letters such as Hamza, Alef and Tatweel.
The type of classification	DT, NB, KNN and SVM
Software Used	Weka [30]
The context and the dataset size	3700 Arabic tweets - 1579 Positive - 1374 Negative - 747 Neutral
Strengths	SVM performed better than all learning methods on three classes but NB excelled SVM on Negative class.
Weaknesses	Neutral class has unacceptable results, this shows that the Neutral class is extremely correlated with other classes.
Results	Precision = 0.727, F1 score = 0.722
Contributions	Arriving at this conclusion: SVM learning method outperformed the KNN, NB and Decision tree learning methods with regard to Recall, Precision and F1 measures.
Future Work	Developing a new stemming method that stems Arabic words because the writers do not write the message based on standard Arabic language that works on traditional stemmer.

The authors of [15] explored Distant Supervision (DS) based on emoticons.

TABLE 13
SUMMARIZATION OF [15]

Paper	[15]
Publishing year	2015
SA Level	Sentence
The Language	Arabic
Domain Oriented	No
Polarity	Positive and Negative.
Preprocessing & Filtering	- Normalization of digits and non-Arabic characters - Removing user-names and links
The type of classification	DS, SVM
Software Used	
The context and the dataset size	Gold-Standard Dataset Emoticon-Based Queries training Dataset (120,747 data instances) Lexicon-Based annotation training Dataset (18,105 positive, 10,039 negative instances)
Strengths	DS approaches to SSA for Arabic Twitter feeds show significantly higher performance in accuracy and F-score than a fully supervised approach. SVM is the best performing scheme.
Weaknesses	Using DS but not providing improvement over the supervised techniques in the task of Sentiment Analysis
Results	Accuracy = 95.19%
Contributions	The first to explore distant supervision (DS) approaches for automatic SSA classification for Arabic social networks.
Future Work	

The authors of [10] presented a SA system for MSA and Egyptian dialect.

TABLE 14
SUMMARIZATION OF [10]

Paper	[10]
Publishing year	2015
SA Level	Sentence
The Language	Arabic (MSA and Egyptian dialect)
Domain Oriented	No
Polarity	Positive and Negative.
Preprocessing & Filtering	- Removing foreign characters, symbols, numbers, etc. - Removing stop-words - Normalization (unified Arabic characters and removing all diacritics)
The type of classification	A semi-supervised approach for SA using a high coverage Arabic sentiment words lexicon (400 adjectives collected manually from different websites) which is automatically increased, and Arabic sentiment idioms/saying phrase lexicon (12785 wisdoms and idioms are collected from websites and books) to improve the classification process. Using SVM classifier with linguistically and syntactically influenced features.
Software Used	
The context and the dataset size	2000 Arabic sentiment statements - 1000 MSA tweets and Arabic dialect tweets - 1000 microblogs (hotel reservation comments, product reviews, TV program& movie comments)
Strengths	Exploiting idioms and saying lexicon with a high coverage polarity lexicon has the largest impact on classification accuracy. Also, the automatic expansion of the polarity lexicon yields a great effect on sentiment classification.
Weaknesses	A dataset of small size
Results	Accuracies over 95% in some cases.
Contributions	Employing a number of novels and rich feature sets to handle the valence shifters (negation, intensifiers), questions and supplication terms and to improve the classification performance. Building two lexicons; Arabic sentiment words lexicon and Arabicsentiment idioms/saying phrase lexicon.
Future Work	

The authors of [18] compared between SVM and NB for political Arabic twitter data.

The authors of [20] used Twitter in trading strategy with Mubasher products [25].

TABLE 15
SUMMARIZATION OF [18]

Paper	[18]
Publishing year	2016
SA Level	Sentence
The Language	Arabic
Domain Oriented	Yes
Polarity	Positive and Negative.
Preprocessing & Filtering	- Tokenization - Normalization - Stemming (light) - Removing non-Arabic characters - Removing Stop words
The type of classification	SVM and NB Using of TF-IDF
Software Used	WEKA [30]
The context and the dataset size	18278 tweets - 11910 positive and 6368 negative
Strengths	Building a political corpus
Weaknesses	No real addition
Results	SVM get P =.862 and R =.884 and F =.871 NB get P =.925 and R =.921 and F =.922
Contributions	Arriving at this conclusion: The results shows that the NB method is of the highest accuracy and the lowest error rate.
Future Work	Comparing the results obtained from these used classifiers with other classifiers. Comparing the results obtained from light stemmer which is used and Khoja stemmer [23]. Comparing the results obtained from unigram, bigram and trigram.

TABLE 16
SUMMARIZATION OF [20]

Paper	[20]
Publishing year	2016
SA Level	Sentence
The Language	Arabic
Domain Oriented	
Polarity	Positive, Negative and Neutral
Preprocessing & Filtering	- Replacing some words, such as company codes, percentage sign (%). - Normalization. - Tokenization - Removing stop word - Stemming Light - Filtering token by length 3.
The type of classification	NB and SVMs TF-IDF and Binary-Term Occurrence.
Software Used	Mubasher software [25], Rapidminer [21], Twitter Data Grabber [38]
The context and the dataset size	1331 Total - 378 Positive, 755 Negative, 198 Neutral
Strengths	The best accuracy was achieved by SVM without N-Gram feature .On the other hand, the best accuracy was completed by Naive-Bayes when the N-Gram feature is involved.
Weaknesses	Dataset of small size
Results	Accuracy = 89.68% for SVM using TF-IDF
Contributions	Classifiying Arabic sentiments toward Mubasher products through different algorithms
Future Work	Extracting technical features aspects of Mubasher products such as Human Computer Interaction (HCI).

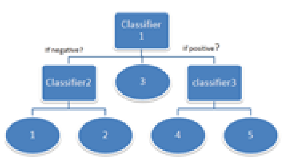
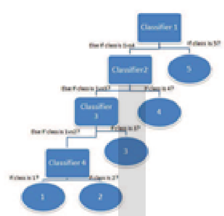
The authors of [19] used supervised learning for tweets written in Arabizi.

TABLE 17
SUMMARIZATION OF [19]

Paper	[19]
Publishing year	2016
SA Level	Sentence level
The Language	Arabic (Arabizi)
Domain Oriented	No
Polarity	Positive, Negative and Neutral
Preprocessing & Filtering	- Tokenizing tweets into words - Mapping every emoticon into its corresponding word - Converting tweets written in Arabizi into tweets written in Arabic using the built-in rule-based converter. - Removing stop-words. - Calculating the weight of every token using the Binary Model.
The type of classification	NB and SVM
Software Used	Crowdsourcing was used to label the dataset.
The context and the dataset size	Arabizi dataset consists of 3206 tweets. - 1803 positive, 831 negative, 572 neutral
Strengths	SVM accuracies are higher than NB accuracies. Removal of stop-words and mapping emoticons to their corresponding words did not greatly improve the accuracies for Arabizi data. Eliminating neutral tweets at early stage in the classification improves precision for both NB and SVM. However, Recall values fluctuated, sometimes they got improved; on other times they did not improve
Weaknesses	Dataset of small size
Results	SVM (before removing Neutral class) - with filter Recall for positive class = .831 for negative class =.377 - without filter Recall for positive class = .814 for negative class =.386 SVM (after removing Neutral class) - with filter Recall for positive class = .869 for negative class =.367 - without filter Recall for positive class = .862 for negative class =.404
Contributions	The first to provide this treatment for Arabizi.
Future Work	

The authors of [17] showed the use of divide-and-conquer hierarchical structure of classifiers

TABLE 18
SUMMARIZATION OF [17]

Paper	[17]
Publishing year	2016
SA Level	Sentence
The Language	Arabic (MSA and colloquial Arabic)
Domain Oriented	No
Polarity	1,2,3,4,5 rating classes
Preprocessing & Filtering	- Bag-Of-Word using StringToWordsVector - Tokenization using WordTokenizer - Removing stop words - No stemmers
The type of classification	SVM, NB, KNN and Decision Tree (DT). Construct different hierarchical classifier trees.  Fig. 5. The 2-level hierarchical classifier.  Fig. 6. The 4-level hierarchical classifier.
Software Used	Weka [30]
The context and the dataset size	LABR dataset: consists of 63,257 book reviews which have a rating (1 to 5) - Class 1 contains 2,939 reviews. - Class 2 contains 5,285 reviews. - Class 3 contains 12,201 reviews. - Class 4 contains 19,054 reviews. - Class 5 contains 23,778 reviews.
Strengths	Hierarchical classifiers give significant improvements (of more than 50% in certain cases) over flat classifiers.
Weaknesses	No stemmers were used during preprocessing. Using LABR dataset only for the experiments.
Results	Accuracies SVM (Flat= 45.7%, 2-level= 45.2%, 4-level =47.4%) DT (Flat= 40.2%, 2-level= 43.9% , 4-level= 47.6%) NB (Flat= 38.2% , 2-level= 39.9% 4-level= 48.9%) KNN (Flat= 38.6% , 2-level= 46.2%, 4-level= 57.8%)
Contributions	Showing how the use of this divide-and-conquer hierarchical structure of classifiers can generate better results than the use of existing flat classifiers for the Multi-Way Sentiment Analysis (MWSA) problem (5 star scoring).
Future Work	

The authors of [39] built a Saudi twitter corpus for SA.

TABLE 19
SUMMARIZATION OF [39]

Paper	[39]
Publishing year	2016
SA Level	Sentence
The Language	Arabic (Saudi dialect)
Domain Oriented	No
Polarity	Positive, Negative and Neutral.
Preprocessing & Filtering	- Removing duplicate tweets before and after the cleaning process - Removing links, hashtags and non Arabic words

The type of classification	begin
	Prepare set of three application accounts on twitter to be used as a way of authenticating
	while *Authenticated by the account that has available rate limit* do
	for *each hashtag w in Hashtags* do
	if *current account rate limit is reached* then
	break
	end
	Use twitter api to search for term (hashtag)
	Save the collected tweets for the current hashtag
	end
	end
	end
	Fig. 1 Pseudo Code for collecting tweets usiing twitter API used in this work

Software Used	Ruby-on-Rails (RoR) Twitter API to collect data.
The context and the dataset size	4700 tweets - First annotator: 1830 positive, 1991 negative and 904 neutral - Second annotator: 2100 positve, 2016 negative and 584 neutral
Strengths	Building a free public corpus
Weaknesses	Dataset of small size
Results	Saudi dialect corpus with (Kappa = 0.807)
Contributions	Presenting an annotated publicly available corpus that applied SA to Twitter content.
Future Work	Extending this corpus Generating a large scale lexicon for Saudi dialect Building a comprehensive SA system for Saudi dialect using big data technique.

The authors of [40] built a mutli-dialects Arabic corpus.

TABLE 20
SUMMARIZATION OF [40]

Paper	[5]
Publishing year	2011
SA Level	Document level
The Language	Arabic
Domain Oriented	Yes (Multi-domain: Education, Sports, and Politics)
Polarity	Positive & Negative
Preprocessing & Filtering	- Stripping out the HTML tags and non-textual contents - Separating the documents into posts and Converting each post into a single file. - Normalizing some alphabets - Removing some repeated letters - Correcting some of the wrong spelling words - Tokenization - Removing stop words - Applying Arabic light stemmer - Using Term Frequency–Inverse Document Frequency (TF-IDF) weight - Removing some terms with a low frequency of occurrence.
The type of classification	lexicon-based opinion classifier, then Maximum entropy method and finally k-nearest method
Software Used	SentiStrength [24]
The context and the dataset size	Total of 1143 posts which contain 8793 Arabic statements with average of 7.7 statements in each post.
Strengths	The accuracy almost improved from 50% using one method, 60% using two method and 80% using three methods which is a satisfactory performance especially for complex language such as Arabic.
Weaknesses	No Arabic-specific features
Results	Accuracy = 80.29%
Contributions	A combined approach that automatically extracts opinions from Arabic documents that consists of three successive methods; At the beginning, lexicon based method is used to classify as much documents as possible. The resultant classified documents are used as training set for the maximum entropy method which subsequently classifies some other documents. Finally, k-nearest method used the classified documents from lexicon based method and maximum entropy as training set and classifies the rest of the document
Future Work	The experimental results further show that recall and precision of positive documents are better than the negative one. That means further studies should be done for mining of negation of Arabic statements. Also, in the future, it is planned to extend the work to be able to extract features from Arabic opinioned statements

4 Arabic Language Challenges

Arabic language is complex to analyze because of the special properties it has. The following points will explain the challenges in Arabic language:

(1) The limited work in this area when compared to other languages especially English. Arabic online resources are increasing nowadays, but they are still comparatively very small.

(2) Morphological complexities and dialectal varieties of the Arabic language which require advanced pre-processing and lexicon-building steps. Every country/part of a county has its own version or dialect of Arabic. That means there are different dialects of Arabic text available online that could hold different meaning [3].

(3) The limitation of customized tools for Arabic SA may not be easy to come by. Available tools may be limited in current functionality or may not be freely available.

(4) Multiple words prefixes, suffixes, affixes, and diacritical forms add high-order dimensionality for words, where the same three-letter root can generate different words in each case.

(5) Arabic grammar is highly complex. Different types of sentence structures can exist in Arabic: verbal, where the sentence starts with a verb phrase, and nominal, where the sentence starts with a noun phrase. Additionally the language allows for different variants within each type of sentence [32].

(6) The presence of negation words can cause a sentence to have two opposite sentiments at the same time or toggle the polarity.

5 Conclusion

From the study, comparison and analysis of the different proposed methodologies for SA it was observed that SVM yield the best performance in case of SA. The problem with NB is that it is based on probabilities, thus it is more suitable for inputs with high dimensionality. Because of the principal advantages of SVM, it was applied successfully in several sentiment analysis tasks. These principal advantages can be summarized as follows: They are robust in high dimensional spaces. All features are considered relevant. They are robust when there is a sparse set of samples. Finally, most text categorization problems are linearly separable [29].

This work highlighted important papers in Arabic Sentiment analysis. It also revealed many challenges and open areas that need to be addressed and investigated to enhance this field.

6 Future Work

We can summarize the future work in Arabic SA as follows:
- Building a big corpus and putting it freely public.
- Using the big corpus to compare methodologies.
- Building larger lexicons and dictionaries
- More investigation for Arabic negation.
- Building appropriate Stemmer for Arabic dialects.
- More research in semantic analysis as the same word can have many meanings in different contexts.

References

[1] Fouzi Harrag. Estimating the Sentiment of Arabic Social Media Contents: A Survey. 5th International Conference on Arabic Language Processing, 2014.

[2] Korayem, M., Crandall, D., and Abdul -Mageed, M. 2012. Subjectivity and Sentiment Analysis of Arabic: A Survey. Advanced Machine Learning Technologies and Applications , Communications in Computer and Information Science . 322, 128- 139.

[3] Sarah O. Alhumoud, Mawaheb I. Altuwaijri, Tarfa M. Albuhairi, Wejdan M. Alohaideb. Survey on Arabic Sentiment Analysis in Twitter. International Journal of Social, Behavioral, Educational, Economic and Management Engineering Vol:9, No:1, 2015.

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

[5] El-Halees, A.: Arabic Opinion Mining Using Combined Classification Approach. In: Proceedings of the International Arab Conference on Information Technology, ACIT (2011)

[6] Nawaf A. Abdulla, Nizar A. Ahmed, Mohammed A. Shehab and Mahmoud Al-Ayyoub. Sentiment Analysis: Lexicon-based and Corpus-based. 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) 978-1-4799-2303-8/13.

[7] Rehab M. Duwairi and Islam Qarqaz, Arabic Sentiment Analysis using Supervised Classification, The 1st International Workshop on Social Networks Analysis, Management and Security (SNAMS - 2014), August 2014, Barcelona, Spain.

[8] Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. (2011). Bilingual experiments with an arabic-english corpus for opinion mining.

[9] Mahmoud Nabil, Mohamed Aly and Amir F. Atiya. ASTD: Arabic Sentiment Tweets Dataset. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2515–2519, Lisbon, Portugal, 17-21 September 2015.

[10] Hossam S. Ibrahim, Sherif M. Abdou and Mervat Gheith.SENTIMENT ANALYSIS FOR ODERN STANDARD ARABIC AND COLLOQUIAL. International Journal on Natural Language Computing (IJNLC) Vol. 4, No.2,April 2015.

[11] R. M. Duwairi, Raed Marji, Narmeen Sha'ban, Sally Rushaidat. Sentiment Analysis in Arabic Tweets. 2014 5th International Conference on Information and Communication Systems (ICICS) 978-1-4799-3023-4/14

[12] Biltawi M., Etaiwi W., Tedmori S., Hudaib A., and Awajan A., "Sentiment Classification Techniques For Arabic Language: A Survey," in Proceedings of the 7th International Conference on Information and Communication Systems, Jordan, 2016.

[13] M. Al-Kabi, N. M. Al-Qudah, I. Alsmadi, M. Dabour, and H. Wahsheh, "Arabic/english sentiment analysis: An empirical study," in The fourth International Conference on Information and Communication Systems (ICICS 2013), 2013.

[14] Hadi, Wa'el. "Classification of Arabic Social Media Data." Advances in Computational Sciences and Technology 8.1 (2015): 29-34.

[15] E. Refaee and V. Rieser, "No Bad Feelings : Distant Supervision Helps Subjectivity but not Sentiment Analysis of Arabic Twitter Feeds," Proceedings of the 8th Saudi Scientific Conference, London, UK, 2015, pp. 2–18.

[16] Abdul-Mageed, Muhammad, Sandra Kübler, and Mona Diab. "Samar: A system for subjectivity and sentiment analysis of arabic social media." Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis. Association for Computational Linguistics, 2012.

[17] M. Al-Ayyoub, A. Nuseir, G. Kanaan, and R. Al-Shalabi, "Hierarchical classifiers for multi-way sentiment analysis of arabic reviews," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 2, pp. 531–539, 2016.

[18] Elghazaly, Tarek, Amal Mahmoud, and Hesham A. Hefny. "Political Sentiment Analysis Using Twitter Data." Proceedings of the International Conference on Internet of things and Cloud Computing. ACM, 2016.

[19] Duwairi, Rehab M., et al. "Sentiment analysis for Arabizi text." 2016 7th International Conference on Information and Communication Systems (ICICS). IEEE, 2016.

[20] Al-Rubaiee, Hamed, Renxi Qiu, and Dayou Li. "Identifying Mubasher software products through sentiment analysis of Arabic tweets." 2016 International Conference on Industrial Informatics and Computer Systems (CIICS). IEEE, 2016.

[21] The rapid-I website. [Online]. Available: http://www.rapid-i.com/

[22] Microsoft Office. (2013) The Office website. [Online]. Available: http://office.microsoft.com/en-GB/?CTT=97

[23] Shereen Khoja. (1999) The Pacific University homepage. [Online]. Available: http://zeus.cs.pacificu.edu/shereen/research.htm

[24] The SentiStrength website. [Online]. Available: http://sentistrength.wlv.ac.uk/SentStrength_Data/

[25] The Mubasher website. [Online]. Available: http://www.mubasher.net/index.html

[26] Diab, Mona. "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking." 2nd International Conference on Arabic Language Resources and Tools. 2009.

[27] The SocialMention website. [Online]. Available: http://socialmention.com

[28] The Twendz website. [Online]. Available: http://twendz.waggeneredstrom.com/

[29] M. Rushdi-Saleh, M. Teresa, L. Martín-Valdivia, A. Ureña-López, and J. M. Perea-Ortega, "OCA: Opinion corpus for Arabic". Journal of the American Society for Information Science and Technology, 62: 2045–2054.doi:10.1002/asi.21598.2011.

[30] The WEKA website. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka

[31] Thorsten Joachims. 2008. Svmlight: Support vector machine. http://svmlight.joachims.org/ , Cornell University, 2008.

[32] Farra, Noura, et al. "Sentence-level and document-level sentiment mining for arabic texts." 2010 IEEE International Conference on Data Mining Workshops. IEEE, 2010.

[33] N. Habash. Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1):1–187, 2010.

[34] The PROMT website. [Online]. Available: http://translation2.paralink.com

[35] Abdulla, Nawaf A., Mahmoud Al-Ayyoub, and Mohammed Naji Al-Kabi. "An extended analytical study of arabic sentiments." International Journal of Big Data Intelligence 1 1.1-2 (2014): 103-113.

[36] Hamouda, Alaa El-Dine Ali, and Fatma El-zahraa El-taher. "Sentiment analyzer for arabic comments system." Int. J. Adv. Comput. Sci. Appl 4.3 (2013).

[37] Refaee, Eshrag, and Verena Rieser. "An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis." LREC. 2014.

[38] The Twitter Data Grabber website. [Online]. Available: https://dev.twitter.com/rest/public

[39] Assiri, Adel, Ahmed Emam, and Hmood Al-Dossari. "Saudi Twitter Corpus for Sentiment Analysis." World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering 10.2: 242-245.

[40] Al-Kabi, Mohammed N., et al. "A Prototype for a Standard Arabic Sentiment Analysis Corpus." International Arab Journal of Information Technology (IAJIT) 13 (2016).

[41] The Scopus website. [Online]. Available: https://www.scopus.com/