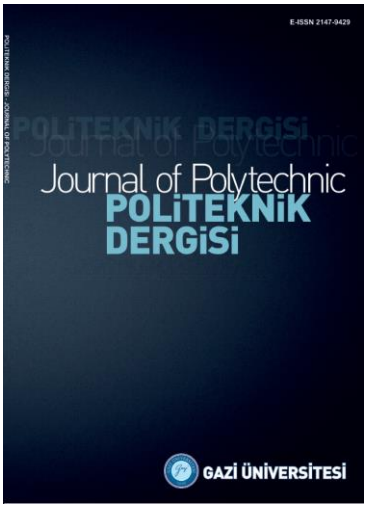

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.gov.tr/politeknik

The corpus based approach to sentiment analysis in modern standard Arabic and Arabic dialects: a literature review

Modern standart Arapça ve Arap lehçelerinde duygu analizine külliyat (korpus) temelli yaklaşım: bir literatür incelemesi

***Yazar(lar) (Author(s))**: Anwar ALNAWAS[1], Nursal ARICI[2]*

ORCID[1]: https://orcid.org/0000-0001-9181-9377

ORCID[2]: https://orcid.org/0000-0002-4505-1341

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): Alnawas A. and Arıcı N., "The corpus based approach to sentiment analysis in modern standard Arabic and Arabic dialects: a literature review", ***Politeknik Dergisi***, 21(2): 461-470, (2018).

Erişim linki (To link to this article): http://dergipark.gov.tr/politeknik/archive

DOI: **10.2339/politeknik.403975**

The Corpus Based Approach to Sentiment Analysis in Modern Standard Arabic and Arabic Dialects: A Literature Review

Review Article / Derleme Makalesi

Anwar ALNAWAS[1,3], Nursal ARICI[2*]

[1]Graduate School of Natural and Applied Sciences, Computer Engineering, Gazi University, Turkey
[2] Department of Computer Engineering, Faculty of Technology, Gazi University, Turkey
[3] Nasiriyah Technical Institute, Southern Technical University, Iraq
(Geliş/Received : 13.10.2016; Kabul/Accepted : 21.02.2017)

ABSTRACT

Sentiment Analysis, is the analysis of ideas, emotions, evaluations, values, attitudes and feelings about products, services, companies, individuals, tasks, events, titles and their characteristics. With the increase in applications on the Internet and social networks, Sentiment Analysis has taken a considerable place in the field of text mining research and has since been used to explore the opinions of users about various products or topics discussed over the Internet. When the literature on Sentiment Analysis is examined, it is seen that the natural language of the Internet information sources that form the basis of the analysis is mostly English. Developments in the fields of Natural Language Processing and Computational Linguistics have contributed positively to Sentiment Analysis studies made from natural languages other than English. The purpose of this study is to examine the literature of Sentiment Analysis conducted in Arabic internet information sources. The literature review includes studies based on the corpus approach, which is made up of Arabic Internet information sources. Studies are being carried out on the works which constitute their own corpora for both Modern Standard Arabic and Arabic dialects and on which sentiment analysis is performed.

Keywords: Sentiment Analysis, corpora, feature extraction, Arabic Language, NLP.

Modern Standart Arapça ve Arap Lehçelerinde Duygu Analizine Külliyat (Korpus) Temelli Yaklaşım: Bir Literatür İncelemesi

ÖZ

Duygu analizi; kişilerin ürünler, servisler, firmalar, bireyler, görevler, olaylar, başlıklar ve bunların özellikleri üzerine fikirleri, duyguları, değerlendirmeleri, değer biçmeleri, tutumları ve hislerini analiz edilmesidir. Internet ve sosyal ağlardaki uygulamaların artmasıyla birlikte, Duygu Analizi (DA), metin madenciliği araştırma alanında dikkate değer bir konuma gelmiş ve o zamandan beri, kullanıcıların Internet üzerinden tartışılan çeşitli ürünler veya konular hakkındaki görüşlerini keşfetmek için kullanılmaktadır. Duygu Analizi üzerine yapılan çalışmalar incelendiğinde analize temel oluşturan Internet bilgi kaynakları doğal dilinin çoğunlukla İngilizce olduğu görülmektedir. Doğal Dil İşleme ve Hesaplamalı Dil Bilim alanlarındaki gelişmeler İngilizce dışındaki doğal dillerden yapılan Duygu Analizi çalışmalarına olumlu katkıları olmuştur. Bu çalışmanın amacı, Arapça içerikli Internet bilgi kaynaklarından gerçekleştirilen Duygu Analizi literaturü incelemektir. Literatür incelemesi, Arapça Internet bilgi kaynaklarından oluşturulan külliyat (corpus) yaklaşımına dayanan çalışmaları kapsamaktadır. Hem Modern Standart Arapça, hem de Arap lehçeleri için kendi külliyatlarını(corpora) oluşturan ve bu metinler üzerinden duygu analizi yapılan çalışmalar incelenmektedir.

Anahtar Kelimeler: Duygu analizi, külliyat, özellik çıkarma, Arapça, doğal dil işleme

1. INTRODUCTION

The Internet has become an important source of user generated content such as text, pictures and emoticon in different sectors including business, health, politics and education. For that user with every login to Social Networks generate content in deferent form like unstructured text, login registration, video streaming and other forms. The generated content from users is gaining considerable attention because of its importance in various sectors like businesses, governmental, political etc. Understanding what people are thinking and their views are essential for decision-making on the subject especially when these comments are combined in one area and on a single topic.

For the importance of the generated content; Web Contents Mining (WCM) is being used by organizations to take advantage of this content. Opinion mining and Sentiment Analysis (SA) are the sub research area of WCM appeared to be witnessing increasing research in the Natural Language Processing (NLP) community.

**Sorumlu Yazar (Corresponding Author)*
e-posta : nursal@gazi.edu.tr

SA aims to analyze the contents generated by the user, whether positive or negative feelings about a specific topic [1, 2]. SA is applied at different levels: document, sentence and aspect with different techniques. In general there are two main techniques for SA; lexical and machine learning approaches. Lexical approach, a dictionary is prepared to store the polarity values of each lexicons. For each word of the text present in the dictionary, polarity score calculating by adding to get an 'overall polarity score'[3].

Arabic content has grown on the Internet, especially with the expansion of Social Networks [4]. For example Twitter has about 11.1 million active users for each month in Arabic region until March 2017 [5], and Facebook is the most popular social media platform with 156 million users [6]. In general, textual data available on the Internet are in both; Modern Standard Arabic (MSA) and dialects. The dialects of Arabic are very varied, with five dialects; Gulf (Khaliji), Iraqi, Levantine, Egyptian and Maghrebi. There is a difference between the dialects. Some native dialects cannot communicate with speakers of different dialects, Therefore, MSA is used for communication and understanding [7, 8]. Standard phonological, morphological and lexical does not exist in Arabic dialects like MSA [9]. Therefore, preprocessing is difficult with dialects because it is not based on fixed rules [10].

In general, many studies on English and some other languages are widely available, as opposed to Arabic. The lack of a database of Arabic dialects and studies on the nature of the acoustic-phonetics are the most important reasons for natural language processing [11].

A limited number of reviews dealt with Arabic SA. The study of [12] a comprehensive overview of the latest updates in the field of SA algorithms and applications. On the other hand, this work reviewed the Arabic SA along with the other languages. In [13], reviewed works on Arabic SA and characteristics of the Arabic language. The work also reviewed three of SA approaches; supervised, unsupervised and hybrid.

The proposed paper is a survey on SA of the Arabic content in various social networks, news or commercial especially that created its own corpus. As well as the presentation of the techniques that used to extract sentiment and corpus building with a review of the most important challenges to the SA in Arabic language.

This paper is organized as follows: Section 2 provides the overview of the SA and Arabic language. Section 3 illustrates the works done using corpora for both MSA and dialects. Section 4 discusses the main issues of Sentiment Analysis and Arabic language derived from previous studies. Section 5 concludes the manuscript.

2. BACKGROUND

2.1. Arabic Language

Arabic is an important language, MSA and dialects are spoken by approximately 422 million people [14]. Arabic is one of morphology rich language [8].Arabic consists of 28 letters, including three letters of long vowel and 25 consonants letters. Arabic is written from right to left alignment. The Arabic letter shape changes according to its position from the word, unlike English. Arabic also contains short vowels, or diacritics that change the meaning of the word such as"كَتَبَ"which means "wrote" and "كُتُبْ" which means "books". Arabic is a Semitic language and contains many dialects depending on different regions. MSA used in education, news and official websites.

Arab world describing as an important player in international politics and the global economy. It is therefore a focus of attention for multinationals, interest groups and analysts who want to decipher sentiment on various issues such as oil prices, stock market movement and foreign policy. There is a large amount of Arabic language on the Internet and this quantity needs to analyze the natural language to extract the required sentiments [15].

SA for the Arabic language is complex for several reasons, including characteristics of Arabic language and lack of SA work for the Arabic language. [4] review the most important characteristics of the Arabic language that effect on SA as follow:

-Multiply the root of the Arabic word depending on the context such as (يكتب, كتاب, كتابة).

-Punctuation affects the meaning of words, meaning that two words have the same spelling but different meanings such as (جَد) which means grandfather and (جِد) which means fined.

-Some linking words can make the sentence carry two conflicting feelings such as (لكن).

In additional to these points there are different dialects in each Arabic country or region and this complicates SA because of different texts that carry feelings.

2.2. Sentiments Analysis

SA is a new field of research uses advanced techniques to mining texts. Machine learning techniques, information retrieval and NLP used to handle large amount of unstructured content that generated by Internet users, especially social networking sites [16].

SA can be defined as the process of determining the semantic orientation of the text that holds the opinion. This semantic orientation can be positive, negative or natural.SA deals with the text in several ways. Either deal with a full document as having only one opinion or the document holds more than one opinion. In general, there are three levels of SA; document-level, sentence-level and aspect-based level [17].

•Document-level: at this level, the analysis considers that the entire document has only one opinion.

•Sentence-level: This level deals with each sentence as containing one opinion. The polarization of the whole document depends on the polarization of the sentence.

•Aspect-based level: Sometimes it is called (feature-based sentiment analysis). A single sentence can contain more than one aspect and each has its own special opinion. At this level the desired aspect is determined and then the polarity of this is set.SA process used two main

approaches: Machine Learning Approach and Linguistics Based Approach. Figure 1 show the main approaches.

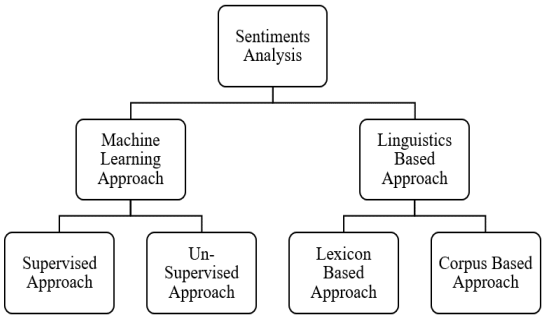

Figure 1: Sentiment classification techniques.

3. CORPORA OF PREVIOUS WORKS

Because of the multitude of dialects in the Arabic language there's no corpus containing all dialects.

Create comprehensive corpus requires great abilities that must be gathered around all dialects in the Arab region. In general, the texts on the Internet are unorganized and contain a lot of noise and repetition in the letters. So it needs to preprocess.

SA of the Arabic language and dialects works, some of them used the old corpora and the other created new corpora. In this section we review the works that created own corpus and how to collect data, source, and method of preprocess.

3.1. Sentiments Analysis of MSA

In recent years, there has been a marked increase in the volume of shared data on daily activities of Internet users. This data contains many information, including user opinions. The large volume of data requests the appearance of a new research field called Sentiment analysis or opinion mining [18].

[19] used the TripAdvisor site to collect 625 review comments. Comments are categorized manually into five categories. Comments are categorized into five categories excellent; very good; middling; weak and horrible. The final data set contains 250 positive comments and 250 negative comments. Tokenization and stemming are applied. They proposed a new mathematical approach to determine polarity of opinion. A linear program was designed to calculate the weights and then calculate the label of each comment. To evaluate the model, used two terms: Root Mean Square Error (RMSE) and Average of Margins (AM).

Some researchers has used part of data set previously collected like in [20], where used large-scale Arabic Book Reviews Dataset (LABR) that collected by [21] as a source for Human Annotated Arabic Dataset (HAAD). HAAD used as annotated corpus for aspect-based sentiment analysis of Arabic text. To classification of the comments that selected out of LABR, they used seven groups of students that studying the course of Natural Language Processing at Jordan University of Science and Technology. Each group contains seven students and was given a task rating of 400 comments. The second phase of auditing work was conducted by native Arabic speaker and holds a Ph.D. degree in computer science. The annotated contains information related to the four tasks: Aspect Term Extraction (T1), Aspect Term Polarity (T2), Aspect Category Identification (T3) and Aspect Category Polarity (T4). Finally, HAAD contains 1513 comments classified in to positive and negative polarity Aspect terms. In order to evaluate T1 and T3, the F1 is measured, where the results were 0.233945 for T1and 0.151815 for T3.to evaluate T2 and T4, the accuracy of approach is measured. The accuracy of T2 and T4 was 0.297064 and 0.425743.

[22] used the TripAdvisor site to compile 625 reviews about hotels. They divided the reviews manually into five categories: excellent, very good, middling, weak and horrible. The normalization process was applied, including the deletion of repeated characters in the word, delete comments that do not represent an opinion. Some comments appear in more than one category, in order to solve this problem, if the comment appears in a category of 80% it is classified into this category. The proposed approach used two of ML algorithms: SVM and K-NN. In first step SVM used to classify in to five classes. To improve the classification process K-NN used in the second step to obtain satisfying results. The hybrid approach showed good results of F-measure up to %97 in average.

[23] used Facebook as a data source. The case study was related to the Israel Gaza conflict 2014. Most of the news on this topic was posted on "breaking news from Gaza" page. As a first stage, about 10,000 posts were collected. After the exclusion of the irrelevant posts, remain 2265 posts. Second stage posts were manually annotated to three classes (positive, negative and neutral) by a group of three member's: graduate student and two senior researchers. BRAT web based annotation tool from [24] also used. AraNLP tool from [25] used to preprocessing text that include: tokenization, stemming, segmentation, part of speech (POS), punctuation and stop word removal N-Gram. For Name entity recognition (NER) feature extraction a web service based tool from [26] was used. The first two tasks T1 and T2 from [20] was applied. The baseline results for T1 was F1=3762% and for T2 was Accuracy= 61.47%.

[27] created corpus using data from Twitter and Facebook. In the beginning 10,000 tweets were collected from Twitter and classified and examined carefully. Some problems have been addressed including repetition in tweets and empty tweets.500 comments were manually collected from Facebook. Many of these comments have been excluded because they were written in Latin characters or contain only emoticons symbols. As a final result the corpus has 2591 comments and tweets (1073 positive, 1518 negative).Tweets classified using the tool in [28]. Either Facebook comments classified manually by the authors of the paper themselves. Using RapidMiner; Tokenize, Stemming, Filter Stop-words, and Generate-n-Grams (Terms) operators were applied to

the data. The researchers presented an SA of the Arabic language from the perspective of Machine Learning (ML).Three supervised methods of classification (Naïve Bayes, SVM and K-Nearest Neighbor classifiers) were built using RapidMiner. To split the data into training and testing sets 10-fold cross validation was used. SVM achieved the best precision it equals to 75.25%.

[29] collected about 350,000 tweet using Twitter's Application Interface (API) by writing PHP script. To facilitate the classification process, a tool was built to help the manual classification of tweets. This tool show and links remover. Three of ML algorithms used to access proposed framework: NB, SVM and K-NN. The experimental of work show a good performance of NB compared to the other.

[21] introduced Large-scale Arabic Book Reviews (LABR) corpus for SA of the Arabic language. About 220000 review have been downloaded from "www.goodreads.com".Approximately 70% of the review were undesirable because they are not written in Arabic or not related to Arabic books. After deleting unwanted contents, the corpus has 63,257 Arabic

Table 1: The sentiments analysis for MSA

Studies	Level	Feature extraction	Data set source	Classifier	Tool	Evaluation	
						Criteria	Result
[19]	Sentences	Low level Light stemming	TripAdvisor website.	SVM	N/A	RMSE	83.5%
						AM	57.6%
[20]	Aspect-based	N-Grams	LABR	T1;T2; T3;T4	BRAT	F1-	23.39% T1
							15.18% T3
						Accuracy	29.7% T2
							42.57% T4
[22]	Sentences	Light-stemming	TripAdvisor website	SVM+ KNN	N/A	Av. F1	97%
[23]	Aspect-based	N-Grams, POS, NER	Facebook	CRF, J48	AraNLP	F1	37.62% T1
							61.47%T2
[27]	Sentences	N-Grams,	Twitter and Facebook	NB,SVM K-NN	RapidMiner	Macro-Precision	66.2%NB 75.2%SVM 70.9%KNN
[29]	Sentences	N-Grams,	Twitter	NB,SVM K-NN	RapidMiner	Accuracy	75.4%NB
							71.6%SVM
							51.5% K-NN
[21]	Sentences	N-Grams,	book readers social network www.goodreads.com	MNB, NB, SVM	Scikit-learn	F1	42.6% MNB
							21.1 NB
							41.0% SVM
[31]	Sentences	Did not extracted	PATB, WTP, WF	SIMP, LG	Manual annotation	Kappa (k)	82.0% ATB 79% WTP 79.3% WF
[33]	Sentences	N-Grams,	Movies webpages	NB, SVM	RapidMiner	Accuracy	90.6% SVM 89% NB

to the user each tweet individually and the possibility to choose the category (positive, negative, neutral and not applicable).The tweets filtering process was based on specific criteria like: each tweet contains at least 100 characters, does not contain more than 4 Hashtag, free of mentions and links and not duplicate or retweets. In the last stage, more than 25,000 tweets were rated. They developed RapidMiner extension to match the task of work. This extension include: Emoticons convertor, repetitions remover, negation detection, and dialect (Jordanian dialect) to MSA convertor, Arabizi convertor

reviews. Reviews submitted by users with a rating of 1-5. The researchers assumed reviews those with ratings 4 or 5 as positive, those with ratings 1 or 2 as negative and reviews with rating 3 are considered neutral and not included in the polarity classification. To rating classification two settings used: a balanced and unbalanced split of reviews number. Different features: unigrams, bigrams, and trigrams with/without Term Frequency–Inverse Document Frequency (TF-IDF) weighting also used in experiments phase. Scikit-learn used to implement experiments phase with Python [30].

Multinomial Naive Bayes (MNB), NB (for binary counts), and SVM used as classifiers. With unbalanced setting SVM does much better, while in the balanced setting MNB is slightly better than SVM.

[31] show how to create AWATIF, which represents a multi-genre corpus of MSA for SA.The dataset that used were taken from three sources; Penn Arabic TreeBank (PATB)[32], About 5342 sentences are taken classified from Wikipedia Talk Pages (WTP) and 2532 of interrelated conversations taken from 7 Web forum (WF). To label AWATIF, Identify two types of labeling guidelines; simple (SIMP) and linguistically-motivated and genre nuanced (LG). For SIMP, to classify the sentence to one of three categories (positive, negative, or natural), they prepared two examples for each of these species to help annotators. For LG, expose annotators to a linguistics background and explain the nuances of the genre to which each data set belongs. With those two types of guidelines the annotators can classify sentences under three conditions; GH-LG, GH-SIMP and AMT-SIMP. GH-LG, Here they use the expertise of students specializing in linguistics (Referred as Gold Human (GH)) to classifying sentences using the LG guidelines. GH-SIMP, Where the GH team works under SIMP conditions. AMT-SIMP, using Amazon Mechanical Turk (AMT) as crowd-sourced with SIMP conditions.

[33] collected data from movies web pages in order to create Opinion Corpus for Arabic (OCA). OCA contains 500 reviews; 250 positive and 250 negative. The process of OCA creating included the collection of reviews from several Arabic blog sites and web pages using a simple bash script for crawling. Manually deleted HTML tags and special characters, and corrected spelling mistakes. Tokenizing, removing Arabic stop words, and stemming and filtering those tokens whose length was less than two characters also applied using RapidMiner. To evaluate OCA many of the experiments have been accomplished. They used cross-validation, Unigram, Bigrams, Trigrams and TF-IDF, to compare the performance of two of machine learning algorithms: SVM and NB. Their results were promising. Table 1 summarizes the sentiments analysis for MSA which are described above.

3.2. Sentiments Analysis of Arabic Dialects.

Arabic dialects vary according to criteria such as geography and social class. The following list is only part of many that covers the main Arabic dialects [34]:

•Gulf Arabic (GLF) includes: of Saudi Arabia Kuwait, Qatar, Bahrain, Oman, and United Arab Emirates dialects.

•Iraqi Arabic (IRQ) includes Iraq dialect..

•Sham Arabic (SHM) includes: the of Lebanon, Syria, Jordan, and Palestine dialects.

•Egyptian Arabic (EGY) includes: Egypt and Sudan dialects.

•Maghrebi includes: covers the of Morocco, Algeria, Tunisia, Mauritania and Libya dialects.

•Yemeni dialect.

[35] create two dialects corpora: one for news; and another for art. The news corpus contains 1000 posts collected from “Al Arabiyya” news Facebook page. The corpus of art also contains 1000 posts collected from “The Voice” page on Facebook. They used three stages of preprocessing for posts; removing redundancies, time stamps and Likes. In order to classify the posts manually, four experts in the Arabic language and dialects were employed. Five rules applied for classification; Negative, Positive, Dual, Spam and Neutral. The corpus was validated, the result of performance was ranging between 73% and 96%.

[36] create a corpus of Mubasher products review using a small program developed in C # and Twitter’s API. Over a period of 57 days, 2051 tweets were collected from Mubasher company Twitter's website in Saudi Arabia wrote in MSA and local Saudi Arabian dialects. Two experts rated tweets manually into three classis (positive, negative and neutral), where labeled positive tweets as “1”, negative as “-1” and neutral as “0”. After deleted unreverent tweets, the corpus remains 1331 tweets. In the normalization phase, a set of signs, codes and Arabic words taken to different forms and English words were replaced standard Arabic equivalents using RapidMiner. Tokenization, Removal stop word, Light stem and Filter token by length. NB and SVM are applied with two feature selection schemes TF-IDF and BTO (Binary-Term Occurrence) create the word vector. The analysis showed good results with SVM.

[37] classified 3700 tweets manually to create their own corpus. During the classification process, only 1550 tweets appeared related to the specific topic. Users’ names, pictures, hashtags, URLs, emoticons, emoticons, and all non-Arabic words deleted from Tweets. To overcome spelling mistakes and to standardize word-writing formulas in tweets, the normalization process was applied. Extract feature was based on unigrams, bigrams, and trigrams. Two classifiers were used to test corpus: NB and SVM using Weka Suite Software. Researchers noted that the best results were achieved for both the NB and SVM classifiers in the unigram language model.

[38] created corpus to predict sentiment in the commercial sector. The dataset are manually collected from several web pages like: reviewzat1, jawal1232 and jumia3. Those data were reviews about five products: Camera, notebook PC, mobile phone, tablet and television. Three experts were hired to classify 250 reviews manually, 125 positive and 125 negative were obtained. Also in this work, they have developed a small (symbol to word) converter to convert emoticon and symbol to the corresponding words that match it. Stop words, special characters, non-Arabic words, and numbers were deleted in normalization phase. Unigram, bigrams, trigrams used to extract features. Three of ML algorithms were used: SVM, NB and K-Nearest Neighbor (KNN). They realized that the use of SVM and NB algorithms detecting the polarity of opinions, gives higher performance than the use of the KNN algorithm.

Table 2: The sentiments analysis for dialects Arabic.

Studies	Level*	Features extraction	Dialect	Data set source	Classifier	Tool	Evaluation	
							Criteria	Result
[35]	Sent	Did not extracted	Multi dialects	Facebook	Manual tagging	Manual tagging	Accuracy	73-96
[36]	Sent	TF-IDF and BTO	Saudi Arabian	Twitter	NB, SVM	RapidMiner	Accuracy	83.6% NB
								79.6% SVM
							Precision	78.2% NB
								98.0% SVM
[37]	Sent	N-Grams	Multi dialects	Twitter	NB, SVM	Weka	Accuracy	84.2% NB
								84.13% SVM
							F1	83% NB
								84% SVM
[38]	Sent	N-Grams	Multi dialect	reviewzat1, jawal1232, jumia3 web sites	NB, SVM	Weka	Precision	946% NB
								948% SVM
							F1	939% NB
								948% SVM
[39]	Sent	N-Grams	Egyptian Dialect	Twitter	ML+SO	Weka	Accuracy	80.9%
							F1	80.6%
[40]	Sent	POS N-Grams	Egyptian dialectal	tweets, product review, hotel reservation comments and TV program comments	Manual annotation	N/A	Kappa	97%
[41]	Sent	Unigram, Bigram	Jordanian dialect	Twitter	SVM; NB; D-Tree; and KNN	RapidMiner	Accuracy	87.2% SVM
								81.3% NB
								51.45% D-T
								50% KNN
[42]	Sent	Unigram, Bigram	Egyptian dialect	Twitter	NB; SVM	Weka	Accuracy	65% NB
								72% SVM
							F1	65% NB
								72% SVM

* Sent= Sentence

[39] used Twitter to create corpus for the Egyptian dialect. Retrieved 20,000 tweets using Twitter's API. Among the retrieved tweets, they used 4800 tweets (1600 positive 1600 negative and 1600 neutral) which every tweet contained only one opinion. To classify tweets benefited from two raters. In the initial processing the images, non-Arabic words, hash tags and URLs were deleted. To apply preprocessing, the tweets must pass through the tokenization "is a process of cutting the document units into small pieces called tokens". Stemming and stop-word removal were applied. Stemming is the process of returning the word to its root form; and stop-word removal is remove words that have little meanings such as"في" (in), " على " (on), " من " (of), etc.. Feature selection was based three different N-Grams models. The hybrid classifier used a new feature set combining the ML (NB and SVM) and the semantic orientation (SO) features. The combination done by adding the SO score as feature in the feature vector built using the ML approach; each sentiment word found is multiplied by the inverse of its SO weight. The results obtained by proposed hybrid approach showed better performance than using either ML or SO approaches.

[40] created MIKA corpus consisting of 4000 subjects of different types of data (tweets, product reviews, hotel reservation comments and TV program comments) written in MSA and the Egyptian dialect. Normalization phase also applied. With the assistance of three Arabic specialists, the data were classified into three categories; positive (PO), negative (NG) and neutral (NU) sentiment. During the classification process, give a power of sentiment to each topic ranging from 1 to 10 for the positive, and (-1 to -10) for the negative, and 0 for the natural.

Several studies have been conducted about SA for English and two main approaches have been used: corpus-based and lexicon-based for this purpose. [41] address both approaches to Arabic starts by create corpus using Twitter data where the corpus contained 2000 tweets (1000 positive and 1000 negative). The tweets written in MSA and the Jordanian dialect for different subjects. The tweets were manually classified by three experts. Repeated letters and stop-words are removed. Also normalization for the letters is done by their tool. They used four supervised methods of classification using RapidMiner; SVM, NB, D-Tree, and KNN. For testing and validation, they applied the 5-fold cross validation. Different stemming techniques tested (root-stemming, light-stemming and no stemming) on each classifier. The results show best accuracy with apply light-stemming for SVM by 87.2%.

[42] use Twitter data to create corpus. Retrieved approximately 4000 tweets, after extracting tweets containing only one opinion 1000 tweets were collected; 500 positive and 500 negative. With the help of three experts, tweets were classified as positive and negative. Normalization also applied by deleting the user name, pictures, hash tags, the URLs and all non-Arabic words. They used Weka Suite Software to extract feature and apply classifier. Unigrams and bigrams used extract feature, NB and SVM used as classifier. The results of the study showed that SVM more accurate than NB, because NB depends on the possibilities therefore is suitable for large training dataset. Table 2 summarizes the sentiments analysis of Arabic dialects that reviewed.

4. DISCUSSION

Social media has become one of the most important of online activities, and began to attract attention in the Arabic-speaking countries. Recently, the Arabic language has begun to raise interest. A lot of the researches have dealt with the Arabic language within the field of NLP with current concerns and SA.The most important issues facing the analysis of feelings for the Arabic language is the lack of annotated data set. This is due to several reasons: Arabic has many dialects; morphological complexity of the Arabic language;

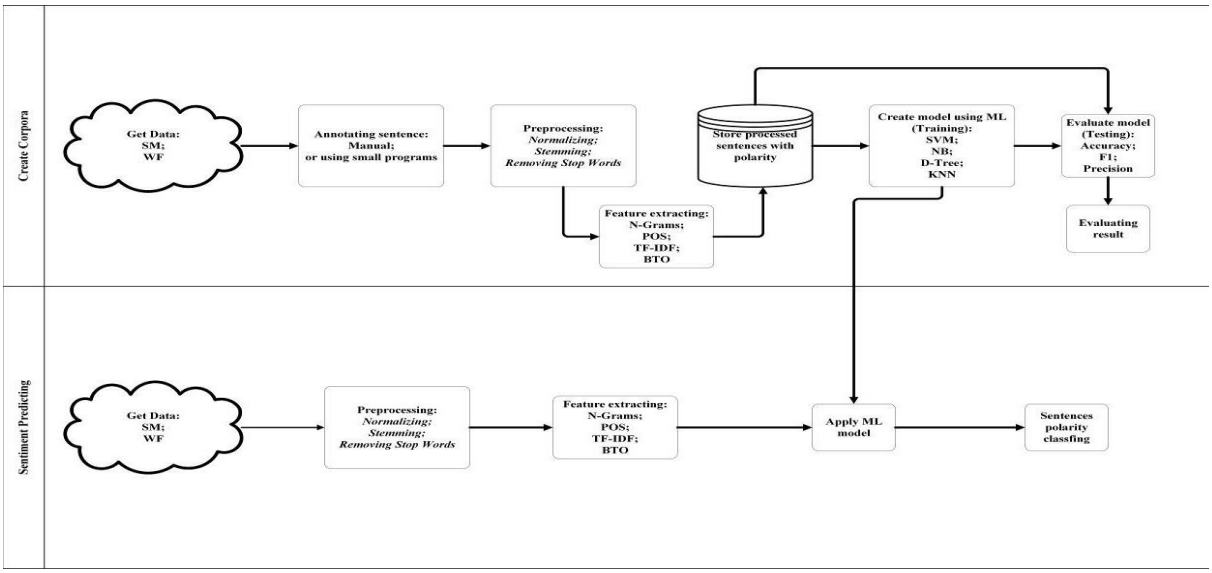

Figure 2: The general framework for corpus based sentiments analysis

Arabic dialects differs from MSA phonologically, morphologically, and lexically.

The dialects included in the previous studies are the Saudi dialect and the Egyptian dialect with a clear lack of other dialects. Previous work showed that there were only two studies [20], [14] that reviewed SA at aspect-level. While, other studies focused on SA at the sentence level. From previous studies, we can summarize the basic steps needed to Arabic SA. Figure 2 show main steps in general framework for SA.

The general framework for corpus based sentiments analysis consists of two parts; the first part represents the establishment of the corpus, model and testing Model. Corpus creation is done by collecting textual data from sources such as Facebook or Twitter, then classify these data manually into the appropriate categories such as negative and positive. Classified text data is not normalized (that means writing one word or one letter in several forms or writing extra characters in the word or repeating letters within the word) Therefore, the process of normalization is applied in preprocessing to unify these different forms.

The preprocessing also contains the delete stop-words process because these words do not effect on the classification process. Most of words contains prefix and suffix and when delete the prefix and suffix we get a large set of single-root words that improve model.

Using a feature set with low representation of text instead of full-size text will perform the required task, therefore input data must be converted to set of features and carefully extract the features. There are many approaches to extract features like POS, TF-IDF and N-Gram.

To create SA classification model based on created corpus, one of the classification algorithms is used or by using a hybrid method of more than one algorithm.

The created model must be tested to determine its validity. They are many criteria's to validate the model such as F1, accuracy and precision.

The second part of the general framework consists of collecting the text data from sources, applying preprocessing and extracting the features as mentioned in the first part.

5. CONCLUSION

In this work, we reviewed the works related to the Sentiments Analysis of the Arabic language for both dialects and Modern Stander Arabic. We provided a detailed analysis of the methods used and the results obtained. In the evaluation, it is found that the most popular methods used to evaluate are NB AND SVM. In general, the dialects included in the previous studies are the Saudi dialect and the Egyptian dialect with a clear lack of other dialects. Also we found that there are environments that offer support for the Arabic language such as RapidMiner and Weka. The nature of Arabic was one of the most important challenges facing Sentiments Analysis with clear lack of datasets availability.

REFERENCES

[1] Aliane A., Aliane H., Ziane M., and Bensaou N., "A Genetic Algorithm Feature Selection Based Approach for Arabic Sentiment Classification", ***2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)***, Agadir, Morocco, 1-6, (2016).

[2] Ravi K. and Ravi V., "A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications", ***Knowledge-Based Systems***, 89: 14-46, (2015)

[3] Bhadane C., Dalal H., and Doshi H., "Sentiment Analysis: Measuring Opinions", ***Procedia Computer Science***, 45: 808-814, (2015)

[4] Alhumoud S. O., Altuwaijri M. I., Albuhairi T. M., and Alohaideb W. M., "Survey on Arabic Sentiment Analysis in Twitter", ***International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*** 9: 364-368, (2015)

[5] Internet: WEEDOO, *Twitter Arab World – Statistics Feb 2017*, 2017, Available: https://weedoo.tech/twitter-arab-world-statistics-feb-2017/, Accessed: 29 July 2017

[6] Internet: WEEDOO, *Facebook Arab World – Statistics Feb 2017*, 2017, Available: https://weedoo.tech/facebook-arab-world-statistics-feb-2017/, Accessed: 29 July 2017

[7] 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)

[8] 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).

[9] Biskri I., Berrakem F.-Z., and Jebali A. "The Applicative Combinatory Categorial Analysis of Arabic", ***In: Proceedings of the 3rd International Conference on Arabic Computational Linguistics, ACLing 2017***, Dubai, United Arab Emirates, 199-207, (2017).

[10] Abuata B. and Al-Omari A., "A Rule-Based Stemmer for Arabic Gulf Dialect", ***Journal of King Saud University-Computer and Information Sciences***, 27: 104-112, (2015)

[11] Alshutayri A. and Atwell E., "Exploring Twitter as a Source of an Arabic Dialect Corpus", ***International Journal of Computational Linguistics (IJCL)***, 8: 37-44, (2017)

[12] Medhat W., Hassan A., and Korashy H., "Sentiment Analysis Algorithms and Applications: A Survey", ***Ain Shams Engineering Journal***, 5: 1093-1113, (2014)

[13] Boudad N., Faizi R., Thami R. O. H., and Chiheb R., "Sentiment Analysis in Arabic: A Review of the Literature", ***Ain Shams Engineering Journal***, (2017)

[14] Internet: UNESCO, *Unesco World Arabic Language Day*, 2012, Available: http://www.unesco.org/new/en/unesco/events/prizes-and-celebrations/celebrations/international-days/world-arabic-language-day, Accessed: 23 March 2017

[15] Al-Kabi M. N., Abdulla N. A., and Al-Ayyoub M. "An Analytical Study of Arabic Sentiments: Maktoob Case Study", ***In: Proceedings of the 2013 8th International***

Conference for Internet Technology and Secured Transactions (ICITST), London, UK, 89-94, (2013).

[16] Sharma A. and Dey S. "A Comparative Study of Feature Selection and Machine Learning Techniques for Sentiment Analysis", ***In: Proceedings of the Proceedings of the 2012 ACM research in applied computation symposium***, San Antonio, Texas,USA, 1-7, (2012).

[17] Awwad H. and Alpkocak A. "Performance Comparison of Different Lexicons for Sentiment Analysis in Arabic", ***In: Proceedings of the 2016 Third European Network Intelligence Conference (ENIC)***, Wrocław, Poland, 127-133, (2016).

[18] Ibrahim M. A. and Salim N., "Opinion Analysis for Twitter and Arabic Tweets: A Systematic Literature Review", ***Journal of Theoretical & Applied Information Technology***, 56: (2013)

[19] Cherif W., Madani A., and Kissi M. "A New Modeling Approach for Arabic Opinion Mining Recognition", ***In: Proceedings of the 2015 Intelligent Systems and Computer Vision (ISCV)***, Fez, Morocco, 1-6, (2015).

[20] Al-Smadi M., Qawasmeh O., Talafha B., and Quwaider M. "Human Annotated Arabic Dataset of Book Reviews for Aspect Based Sentiment Analysis", ***In: Proceedings of the 2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud)***, Rome, Italy, 726-730, (2015).

[21] Aly M. A. and Atiya A. F. "Labr: A Large Scale Arabic Book Reviews Dataset", ***In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics***, Sofia, Bulgaria, 494-498, (2013).

[22] Cherif W., Madani A., and Kissi M., "Towards an Efficient Opinion Measurement in Arabic Comments", ***Procedia Computer Science***, 73: 122-129, (2015)

[23] AL-Smadi M., Al-Ayyoub M., Al-Sarhan H., and Jararweh Y. "Using Aspect-Based Sentiment Analysis to Evaluate Arabic News Affect on Readers", ***In: Proceedings of the 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)***, Limassol, Cyprus, 436-441, (2015).

[24] Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., and Tsujii J. i. "Brat: A Web-Based Tool for Nlp-Assisted Text Annotation", ***In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics***, Avignon, France, 102-107, (2012).

[25] Althobaiti M., Kruschwitz U., and Poesio M. "Aranlp: A Java-Based Library for the Processing of Arabic Text", ***In: Proceedings of the Ninth International Conference on Language Resources and Evaluation***, Reykjavik, Iceland, (2014).

[26] Al-Rfou R., Kulkarni V., Perozzi B., and Skiena S. "Polyglot-Ner: Massive Multilingual Named Entity Recognition", ***In: Proceedings of the Proceedings of the 2015 SIAM International Conference on Data Mining***, British Columbia, Canada, 586-594, (2015).

[27] Duwairi R. and Qarqaz I. "Arabic Sentiment Analysis Using Supervised Classification", ***In: Proceedings of the 2014 International Conference on Future Internet of Things and Cloud (FiCloud)***, Barcelona, Spain 579-583, (2014).

[28] Duwairi R., Marji R., Shaban N., and Ershaidat S., "Sentiment Analysis", **B.S. thesis**, Jordan University of Science and Technology, (2012).

[29] Duwairi R., Marji R., Sha'ban N., and Rushaidat S. "Sentiment Analysis in Arabic Tweets", ***In: Proceedings of the 2014 5th international conference on Information and communication systems (ICICS)***, Irbid, Jordan, 1-6, (2014).

[30] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., *et al.*, "Scikit-Learn: Machine Learning in Python", ***Journal of Machine Learning Research***, 12: 2825-2830, (2011)

[31] Abdul-Mageed M. and Diab M. T. "Awatif: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis", ***In: Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)***, Istanbul, Turkey, 3907-3914, (2012).

[32] Maamouri M., Bies A., Buckwalter T., and Mekki W. "The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus", ***In: Proceedings of the NEMLAR conference on Arabic language resources and tools***, Cairo, Egypt, 466-467, (2004).

[33] Rushdi-Saleh M., Martín-Valdivia M. T., Ureña-López L. A., and Perea-Ortega J. M., "Oca: Opinion Corpus for Arabic", ***Journal of the Association for Information Science and Technology***, 62: 2045-2054, (2011)

[34] Biadsy F., Hirschberg J., and Habash N. "Spoken Arabic Dialect Identification Using Phonotactic Modeling", ***In: Proceedings of the Proceedings of the eacl 2009 workshop on computational approaches to semitic languages***, Athens, Greece, 53-61, (2009).

[35] Itani M., Roast C., and Al-Khayatt S. "Corpora for Sentiment Analysis of Arabic Text in Social Media", ***In: Proceedings of the 2017 8th International Conference on Information and Communication Systems (ICICS)***, Irbid, Jordan, 64-69, (2017).

[36] Al-Rubaiee H., Qiu R., and Li D. "Identifying Mubasher Software Products through Sentiment Analysis of Arabic Tweets", ***In: Proceedings of the 2016 International Conference on Industrial Informatics and Computer Systems (CIICS)***, Sharjah, United Arab Emirates, 1-6, (2016).

[37] Hathlian N. F. B. and Hafezs A. M. "Sentiment-Subjective Analysis Framework for Arabic Social Media Posts", ***In: Proceedings of the Saudi International Conference on Information Technology (Big Data Analysis) (KACSTIT)***, Riyadh, Saudi Arabia, 1-6, (2016).

[38] Sghaier M. A. and Zrigui M. "Sentiment Analysis for Arabic E-Commerce Websites", ***In: Proceedings of the International Conference on Engineering & MIS (ICEMIS)***, Agadir, Morocco, 1-7, (2016).

[39] Shoukry A. and Rafea A. "A Hybrid Approach for Sentiment Classification of Egyptian Dialect Tweets", ***In: Proceedings of the 2015 First International Conference on Arabic Computational Linguistics (ACLing)***, Cairo, Egypt, 78-85, (2015).

[40] Ibrahim H. S., Abdou S. M., and Gheith M. "Mika: A Tagged Corpus for Modern Standard Arabic and Colloquial Sentiment Analysis", ***In: Proceedings of the 2015 IEEE 2nd International Conference on Recent***

Trends in Information Systems (ReTIS), , Kolkata, India, 353-358, (2015).

[41] Abdulla N. A., Ahmed N. A., Shehab M. A., and Al-Ayyoub M. "Arabic Sentiment Analysis: Lexicon-Based and Corpus-Based", ***In: Proceedings of the 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)***, Amman, Jordan, 1-6, (2013).

[42] Shoukry A. and Rafea A. "Sentence-Level Arabic Sentiment Analysis", ***In: Proceedings of the 2012 International Conference on Collaboration Technologies and Systems (CTS)***, Denver, CO, USA, 546-550, (2012).