ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 1012-1023 (2021)

www.astesj.com

ASTESJ
ISSN: 2415-6698

Text Mining Techniques for Sentiment Analysis of Arabic Dialects: Literature Review

Arwa A. Al Shamsi*, Sherief Abdallah

The British University in Dubai, Faculty of engineering and IT, Dubai, 345015, UAE

ARTICLE INFO

Article history:
Received: 28 November, 2020
Accepted: 06 February, 2021
Online: 16 February, 2021

Keywords:
Natural Language Processing
Arabid Dialects
Sentiment Analysis

ABSTRACT

Social media attracts a lot of users around the world. Many reasons drive people to use social media sites such as expressing opinions and ideas, displaying their diaries and sharing them with others, social communication with family and friends and building new social relationships, learning and sharing knowledge. Written text is one of the most common forms used for communication while using social media sites. People use written texts in different languages, and due to the increased usage of social networking sites around the world, the amount of texts and data resulting from this use is large. These generated data considered as a valuable source of information that attracted business owners, companies, government institutions, and of course, it attracts researchers and data scientists as well. Researchers and data scientists increasingly presented great efforts in investigating and analyzing Arabic Language texts. Most of these efforts targeted the Modern Standard form of Arabic Language. While exploring the social media sites, most of the Arab users tend to use their dialects while utilizing Social Media sites, which results in generating a massive amount of Arabic Dialects texts. The number of researches and analysis of Dialects' form of the Arabic language are limited, however, it is increasing recently. This literature review aims to explore approaches and methods used for Sentiment Analysis of Arabic Dialects text.

1. Introduction

Social Media sites have become very popular in society, the popularity of social media is increasing day by day. Recently, many people prefer to spend their time using various applications in smart devices and using the Internet as well. Perhaps social media may take the majority of this usage. People use social media for various reasons such as online shopping, learning, communication, expressing opinions and ideas, sharing their diaries, and many different reasons. People tend to express their opinions, thoughts, feelings, and comment on the various topics that are posted on social media using their dialects. Dialects are the informal form of the language. Each country of the Arab world has its Dialect, and each dialect has many sub-dialects. In [1], the author stated that the population of the Arab world prefer to utilize their dialects in their daily communication, Arabic dialects increasingly utilized online for communications and in social media, moreover, Arabic dialects utilized in TV shows as well as radio programs. As social media usage increased sharply, the amount of data generated as a result of this usage is increasing as well. In [2], the author stated that due to the great amount of data in the form of Natural Language generated in a daily manner online, there is a great need to process this kind of data. This huge amount of generated data attracted companies' owners, marketers and business owners, government institutes and, scientists and researchers as well.

Recently, the world witnesses the revolution of Artificial Intelligence, Data Science, and Machine Learning. Researchers and data scientists are increasingly interested in studying and analyzing natural language texts. Great efforts and researches increasingly targeted the Modern Standard form of Arabic Language, however, researches that targeted the dialects form of Arabic language are limited. In [1], the author stated that due to the limited tools and software that can be utilized for Dialects, researches that targeted dialects are very limit. Many factors affected this limitation, such as the dialect's complexity. Authors in [3] stated that the Dialectal form of the Arabic language has no standard written format which is considered a challenge for analyzing and processing Arabic Dialects. Authors in [4] added to the Arabic Dialects challenges the Diacritical issue and explained how it may change the meaning of the same word, moreover, authors explained how negation may be considered as a challenge while dealing with Arabic Dialects.

This literature review aims to explore researches that involve constructing resources for Arabic Dialects and investigate approaches and methods used for Sentiment Analysis of Arabic Dialects text, focusing on machine learning approaches and Lexicon-based approaches.

[*] Corresponding Author: Arwa Ahmed Al Shamsi, 20180935@student.buid.ac.ae

https://dx.doi.org/10.25046/aj0601112

2. Literature Review

2.1. Arabic Language Background

Arabic is one of the most popular languages that are spoken by millions of people all around the world. In [5], the author stated that the Arabic language is considered the fifth most common language that is spoken by more than 420 million people all around the world. The Arabic language has its unique features. It consists of 28 letters and it is written from the right side to the left side. In [6], the author mentioned that based upon statistics presented by Wikipedia in 2018; the Arabic language is the official language of 25 nations and 380 million is the approximate number of Arabic speakers. In [7], the author mentioned that the Arabic language is one of the Semitic languages meaning that it is written from the right side to left, moreover, Arabic language letters shape changed according to the position of the letter in the word itself. In [1], the author stated that Arabic language letters are used as well in Malay, Urdu, and Persian languages. The Arabic Language is written from right to left. The Arabic language has three types i.e. Classical Arabic which is the language of the Holy Qur'an, Modern Standard Arabic, and Dialectal Arabic form. The Arabic language attracted researchers due to the increased usage of this language over the internet. In [7], the author stated that Arabic users over social media are increasing year after year according to official statistics, this increase resulted in a massive amount of data generated daily online that are in Arabic language. Hence, there is an increasing need for powerful tools and effective approaches for processing Arabic language texts that are in either the Modern Standard Arabic form or in the Arabic Dialects form.

2.1.1. Modern Standard Arabic

Modern Standard Arabic (MSA) is the standard form of the Arabic language that is used in formal papers, schoolbooks, education, TV news, newspapers, street signs, etc. Modern Standard Arabic has a written standard format while the dialects are not. As mentioned earlier, the Arabic language attracted researchers due to the increased use of this language over the internet. Most of the researches that targeted the Arabic language focused on the Modern Standard Arabic form of the Arabic language. In [8], the author stated that NLP tools and applications are mostly based upon the Modern Standard Arabic form of the Arabic language. Modern Standard Arabic is closer to Classical Arabic compared to Arabic Dialect that is less related to classical Arabic. In [9], the author attracted by NLP for Arabic Language and most of the researches done concentrated on MSA as reported in a systematic review.

2.1.2. Arabic Dialect

Arab World consists of 22 countries. Each of these Arab countries has a special Arabic dialect that their population used for daily conversations and talk. In [1], the author mentioned some of the most common Arabic Dialects such as Levantine Arabic, Egyptian Arabic, Gulf Arabic, North African Arabic, and many other Arabic Dialects are spoken by the Arab population. It stated that the Arab population prefers to use their dialects in their daily communication, Arabic dialects increasingly used online in social media sites, moreover, Arabic dialects appeared and utilized in TV shows as well as radio programs. The main dialect for each country can be divided into more sub-dialects. It stated that Arabic Dialects consists of Arabic and non-Arabic words that exist as a result of many reasons; an example of the reasons: Gulf people traveling to India and Iran, moreover, European traders came to Gulf countries after oil discovery, these reasons resulted in non-Arabic words existence in Gulf dialects. In [10], the author interested in their research in studying Arabic dialects. The authors described how Gulf Arabic dialect is the language of the population of Gulf Cooperation, however, this Gulf Arabic dialect is differing slightly between the population of each of the Gulf Cooperation Countries. In [1], the author mentioned that the limited number of Dialects software and NLP tools resulted in limited works and researches that studied and analyzed Dialects. However, [9] in a systematic review presented some research papers that have studied Arabic Dialects. Authors presented valuable researches done in the field of basic language analysis such as ADAM which is an Analyzer for Dialectal Arabic Morphology of Egyptian and Levantine dialectal language, and CALIMA which is an analyzer for Egyptian dialects morphological. Authors as well presented researches and works that have been concentrated on building resources such as Curras which is a dataset of Palestinian dialects and it consists of 56,000 morphologically annotated tokens, DART which is a dataset of around 25000 Arabic tweets, ArabicWeb16 which is a dataset of 10.8 TB of Arabic dialects, CALYOU which is a dataset of Algerian dialect, NileULex which is an Arabic sentiment lexicon of Egyptian and Levantine dialects in addition to Modern Standard Arabic, and TSAC dataset which is sentiment analysis dataset for Tunisian dialects.

2.2. Machine. Learning

Machine Learning involves constructing systems and models that can be improved over experience. It stated that Machine Learning involves machines and systems that can program themselves to learn and get the knowledge needed for better performance. The most common machine learning methods are Classification, Clustering, Regression, Deep Learning and Neural Networks, Transfer Learning, Word Embeddings, Natural Language Processing, Dimensionality Reduction, Reinforcement Learning, and Ensemble Methods. Deep Learning and Neural Networks, Word Embeddings, Transfer Learning, and Natural Language Processing methods of Machine Learning will be further explained below.

2.2.1. Deep Learning and Neural Networks

Deep Learning and Neural Networks are considered a revolutionary approach in the Machine Learning domain. In [6], the author explained that Artificial Neural Network utilized for complex problem solving as it functions in a way similar to Neural Network in brains of humans. Deep Neural Networks are known for their accuracy as well as outstanding performance. Deep Learning approaches are used increasingly for NLP tasks. Researchers utilized Deep Learning approaches for Arabic NLP. As an example of researches used Deep Learning for Arabic NLP: [6] used nine Deep Learning models for text categorization. Moreover, they utilized Word Embeddings approaches and evaluated performance and accuracy. Results showed that all of the nine Deep Learning models presented very good performance and high accuracy, moreover, the use of Word embeddings increased the accuracy and improved the performance. Additionally, in [11] the author investigated different Deep Learning models for Arabic Dialects text classification. Authors concentrated on Egyptian, Levantine, and Gulf dialects and reported that for Egyptian-Gulf pair; Bi-Directions LSTM offered better performance than other Deep Learning models, while for other dialects pairs; LSTM presented better performance.

2.2.2. Transfer Learning

Transfer learning involves using tasks or models that have been learned and transfer the learned knowledge along with applying improvements for a new task or model. In [12], the author defined transfer learning as the process of using data from a source domain to solve problems in another domain. The problem aimed to be solved is related to the data from the source domain, but it is different. In [13], the author successfully presented a model in which Transfer Learning can be used effectively in the case of multiple source domains used for solving problems in multiple target domains. In [14], authors used the Transfer Learning method for ANLP. Transfer learning was used as an extension for the word embeddings model. Authors investigated the effectiveness of the extension applied to skipgram model, the extension involved incorporation of lemmas and efficient use of word2vec word embedding model. The authors reported that the extended model presented better performance than word2vec and fastText on the Arabic word similarity task.

2.2.3. Word Embeddings

Word embeddings is an emerging field that involves distributed word representations which mean representing words as vectors in space. Word Embeddings models are either monolingual or bilingual. The most common is the monolingual models. In [15], the author stated that monolingual word embedding models can be utilized for word order and morphology, while bilingual word embedding models can be utilized for machine translation and parallel sentence extraction. The authors explained that bilingual word embedding models are vector representations of two languages, these languages are mapped into the same space. Word Embeddings have been implemented and utilized for NLP purposes. In [16], the author stated that word embeddings involve using semantic features for representing words as vectors, word embeddings utilized in NLP most commonly for classification and sentiment analysis. In [17], the author stated that the most common Word Embeddings methods are Word2Vec and GloVe. In [18], the author mentioned that there are 4 Arabic word Embeddings which are CBOW, GloVe, Skip-gram model, and Arabic part of the Polyglot word embeddings. The authors evaluated 4 Arabic word embeddings models utilizing benchmark and reported that the best performance achieved from the CBOW model, while the least performance was achieved from the Polyglot model of word embeddings. In [5], the author defined Word Embeddings as vectors used to represent words in continuous space to find any relation between them. The authors presented AraVec which is an open-source Word Embeddings project utilized in the ANLP field. In [19], authors enlarged the informative content of the training sentences by efficient adaptations to word embeddings tools which result in improving the accuracy and performance. Authors as well were able to successfully utilize one embedding space to represent disparate dialects.

2.2.4. Natural Language Processing NLP

Language is the way of communication between people. Language helps us to understand the world around us. The languages that are spoken by people all around the world are known as natural languages. Natural Language Processing (NLP) involves the use of computers to understand and deal with natural languages. In [70], the author define NPL as a section of Artificial Intelligence and Computer Science that involves studying the interactions between human natural languages and computers, moreover, NLP involves Natural Language understanding and generation. The authors mentioned that the increased information in natural language form increased the need for understanding and processing this kind of information. In [2], the author agreed that the massive amount of Natural Language form of data generated daily online increased the need for processing this kind of data. The authors identified NLP as the process of automatic analysis, understanding, and presentation of human Natural Languages.

2.2.4.1. Arabic Natural Language Processing ANLP

ANLP is short for Arabic Natural Language Processing and it involves automatic analysis and processing of Arabic Natural Language. As mentioned earlier, the Arabic Language has three main forms; Classical Arabic, Modern Standard Arabic (MSA), and Arabic Dialects (AD). ANLP tools are supposed to have the ability to deal with the three forms of the Arabic language. However, Classical Arabic is rarely targeted by researchers as it represents the Arabic form of the Holy Qur'an. Tools and techniques are mostly utilized for MSA compared to AD. In [10], the author stated that the use of ANLP tools for AD may be hard due to the nature of AD and the differences between MSA and AD i.e. phonological differences and morphological differences.

2.3. Researches and works on Arabic Dialects

Recently, Arabic Dialects AD attracted researchers. The need to analyze, classify and process the Arabic dialects is increasing due to the fact of increasing the content of Dialect texts, especially in Social Media as stated in [10] . The authors stated that efforts done on MSA are big compared to the works on AD which are limited and mostly targeted Egyptian and Saudi Dialects. However, researchers increasingly do researches and studies that targeted AD. Researches conducted on AD involve basic language analysis, building resources, language identification, and Semantic level analysis. One of the most common examples of semantic level analysis is Sentiment Analysis.

2.3.1. Basic Language Analysis

Basic language analysis for Arabic Dialects involves Orthographic Analysis, Morphological Analysis, and Syntactical Analysis. Sections below present a brief description of each of the basic language analysis type.

2.3.1.1. Arabic Dialects Orthographic Analysis

Arabic dialects have no standard orthographic format meaning that the same word can be written in two or more different ways which may release challenges for NLP tools. In [8], the author stated that MSA and AD are phonologically different, AD have no standard orthographic, i.e. there is no standard format for written AD, Arabic Dialects usually written based upon its phonetics which makes it difficult for analyzing and processing AD. Researchers presented efforts in orthographic analysis for the Arabic language. In [20], the author introduced CODA which is a Conventional Orthography for Dialectal Arabic. CODA offered a computational model that can be utilized for AD. In [21], the author presented valuable efforts in providing conventional orthography that can be utilized for Tunisian Arabic. The presented conventional orthography is based upon CODA that was mentioned earlier. In [10], the author introduced Gumar Corpus which is a Gulf dialects corpus that consists of 110 million words. The corpus was annotated, and the authors presented guidelines for standard orthography analysis.

2.3.1.2. Arabic Dialects Morphological Analysis

The Arabic language is recognized as a rich language of Morphology. In [22], the author defined morphology as the science that involves extracting the word's branches from the word's source. In [8], the author explained how the morphology of MSA is different from the morphology of AD even the grammar, as well as stems of words, may differ. The exploration of Arabic dialects morphology attracted researchers early. In [23], the author introduced MAGEAD which is an Arabic Language Morphological analyzer. MAGEAD is considered as an online morphological generator as well as an analyzer. In [24], the author presented an accurate Egyptian dialect morphological analyzer which is an extension for the Egyptian Colloquial Arabic Lexicon. In [25], the author constructed a lexicon for Tunisian dialects and proposed an approach for Tunisian dialects morphological analysis. Researchers presented efforts as well in constructing a corpus that is morphologically annotated. In [26], the author successfully constructed a morphologically annotated Emirati dialects corpus that consists of about 200,000 words.

2.3.1.3. Arabic Dialects Syntactical Analysis

Dialects are different syntactically, the syntax in dialects affected by many factors, the most common factor that affect the syntax of Arabic dialects is the foreign languages. Syntactical analysis for Arabic dialects has been addressed in several research papers. In [27], the author explored the difficulties in Arabic dialects syntactic analysis, the authors proposed an approach for constructing treebank for Tunisian dialects. In [28], the author proposed a method that involves integration between syntactic analysis and morphological tagging for automatic diacritization of the Arabic language. The method is applied through the case and features prediction improvements. In [29], the author presented guidelines used for syntactic annotation for the treebank of Quranic Arabic dependency which is part of Quranic Arabic Corpus. In [30], the author proposed CamelParser which is a syntactic dependency analysis system for the Arabic language. The proposed system can be used for Morphological Disambiguation.

2.3.2. Building Resources

Researchers worldwide have done great efforts on collecting corpus for Modern Standard Arabic MSA, researchers increasingly attracted by Dialectal Arabic, Great efforts as well have been done to collect corpus for Arabic Dialects.

2.3.2.1. Modern Standard Arabic Corpus Resources

Researchers are increasingly attracted by the Arabic language analysis. One of the most important efforts conducted in the field of ANLP is building resources for the Arabic language. Most of the resources that have been built for the Arabic language are in the form of MSA. In [8], the author stated that almost all available Arabic datasets are for MSA form. Below are some of the researches in which great efforts have been conducted to create corpora of MSA. In [6] the author created two corpora of Modern Standard Arabic text i.e. SANAD and NADiA from Arabic news articles and offered the created corpora as open-source for the public to be utilized for further researches. Moreover, in [31] the author constructed a corpus of MSA that is manually annotated on the sentence level. The corpus was collected from newswire documents. In [32], the author presented AWATIF which is a corpus of MSA that is labeled for Sentiment Analysis purposes at the sentence level. In [33], the author presented noticeable efforts for creating a corpus of MSA from online newspapers.

2.3.2.2. Arabic Dialect Corpus Resources

Dialectal Arabic involves all the dialects that the population of the Arab World use. Arabic Dialects can be categorized according to the region and similarity into: (1) Gulf Dialects which include the Arabic Dialects Spoken by Arab Gulf people, (2) Egyptian Dialect, (3) Levantine Dialect which involves dialects spoken by the population of Palestine, Jordan, Syria, and Lebanon, (4) North African Dialect which include dialects spoken by Morocco, Algeria, Libya and Tunis people. In [8], the author stated that social media websites are considered as one of the most precious sources of AD as people tend to express their thoughts and opinions in written forms using their dialects. In [34], the author stated that although the Arabic Language has been used in a wide range online, the available Arabic datasets are still limited. Internet World Stats statistics represented that the Arabic Language is the fourth most common language used across the internet. Recently, researchers tend to present efforts in ANLP and especially in creating an Arabic corpus that can benefit researches in the ANLP domain. In [10], the author constructed corpus for Gulf Dialects that made up of 100 million words collected from 1200 forum novels, and this Gulf Dialects corpus called Gumar Corpus. In [26], utilized Gumar Corpus to collect a corpus of Emirati dialects. The Collected Emirati dialects consist of around 200,000 words of Emirati Dialects. In [8], the author created a Dialectal Arabic Dataset that include Gulf Dialects, Egyptian Dialect, Levantine Dialects, and North African Dialects. In [34], the author presented BRAD 2.0 which is an extension to BRAD 1.0 corpus. BRAD 1.0 is a dataset of Arabic book reviews that can be utilized for Sentiment Analysis as well as Machine Learning. While BRAD 2.0 is a dataset that is much bigger than BRAD 1.0 and it consists of more than 600,000 Arabic book reviews written in both Modern Standard Arabic and Dialectal Arabic. The Arabic dialects in BRAD 2.0 dataset are Gulf, Egyptian, and Levantine. In [35], the author successfully constructed a corpus of MSA and Saudi Dialect from Twitter and manually annotate the constructed corpus, and offered the constructed corpus for the research community. The authors named the generated corpus AraSenTi-tweet corpus, number of tweets collected were 2.2 million tweets while after annotation the remaining tweets are 17,573 tweets. In [36], the author constructed a corpus of Arabic Dialects. The sources for the corpus text are from Twitter, Facebook, and Newspapers comments. The corpus consisted of Gulf, Egyptian, North African, Levantine, and Iraqi dialects. Twitter texts are classified based upon either seed words, or coordinate points. While Comments from Facebook and Newspapers are classified depending on the nationality of the page owner and country of Newspapers respectively. The authors as well presented an online game that is utilized for text annotation. In [37], the author constructed two corpora i.e. News Corpus (NC) and Arts Corpus (AC) both corpora consist of Arabic Dialects texts from Facebook that can be utilized for Sentiment Analysis. From the above, it is clear how researchers are increasingly interested in building resources for Arabic Dialects. In this literature review, the author targeted research papers that are published in the period from 2014 onward. The databases the author utilized are IEEE, Springer, ScienceDirect, ACM, and WorldCat. The keywords mentioned below have been used for collecting the research papers:

- "Arabic Dialects" and "lexicon"
- "Arabic Dialects" and "dataset"
- "Arabic Dialects" and "corpus"

The inclusion criteria for research papers:

- Must involve constructing resources (dataset / corpus / lexicon) for Arabic Dialects.
- Must be for Arabic Dialect texts only.
- Must be published in the period from 2014 onward.
- Research paper published in journal or conference

Table 1: research and studies that involve constructing a dataset for Arabic Dialects texts

Ref	Data size	Platform or Source	Dialect type	Annotated	Features Extracted
[10]	Gumar Corpus: 110 million words	from 1,200 forum novels	Gulf Arabic	Yes, for sub-dialects at document level for the dialect, novel name and writer name for each	
[26]	about 200,000 words	from eight Gumar corpus novels in the Emirati Arabic variety	Emirati Dialect	Yes, manually annotated	
[8]	13,876,504 word	Twitter, comments from online newspapers, and Facebook	Gulf, Iraqi, Egyptian, Levantine, and North African.	Yes, manually using annotation tool	
[34]	BRAD 2.0 : 692586 annotated reviews	Arabic Book Reviews from www.goodreads.com	MSA and DA (Egyptian, Levantine, and Gulf)	Yes	Yes, unigrams and bigrams
[38]	8000 tweets	Twitter	Arabic and mostly Egyptian	Yes, tweets were labelled into positive, negative, neutral and both	
[71]	Arabic Online Commentary Data set AOC : 52.1M words	Three online newspapers Al-Ghad from Jordan, Al-Riyadh from Saudi Arabia, and Al-Youm Al-Sabe' from Egypt	MSA and Arabic Dialects	Yes, at sentence level	
[35]	AraSenTi-Tweet Corpus: 17,573 tweets	Twitter	Saudi Dialect	Yes, manually and lablled into: positive, negative, neutral and mixed	
[36]	200K tweets, 10K online newspaper comments, and 2M comments from Facebook → total words= 13.8M words	Twitter, Facebook, and Online newspaper	Gulf Dialect, Iraqi Dialect, Egyptian Dialect, Levantine Dialect, and North African Dialect.	Yes, using online game	
[37]	2000 posts, (1000 posts News Corpora NC), and (100 post Arts Corpora AC)	(NC) Al Arabiyya News Facebook page posts, and (AC) collected from The Voice Facebook page	Arabic Dialects	Yes, manual annotation	
[39]	Egyptian Dialect Gender Annotated Dataset (EDGAD) : 70000 tweets	Twitter	Eygptian Dialects	Yes, manually	N-Gram Feature Vector (NFV)
[40]	438,931 tweets	Twitter	Arabic Dialects	Yes, automatically	Ngrams feature
[41]	7698 comments	Facebook	Algerian Vernacular Arabic	Yes, manually	
[42]	1800 tweets	Twitter	MSA and Jordanian Dialect	Yes, manually	N-grams TF-IDF TF
[43]	ArSentD-LEV: 4000 tweets	Twitter	Levantine Dialect	Yes, via crowdsourcing	
[44]	Curras: more than 56 K tokens	Facebook, Twitter, Blogs, and Forums	Palestinian Dialect	Yes, manually	
[45]	194 negative comments & 194 positive comments	Three Algerian newspapers	Algerian Dialect	Yes, by two Arabic native speakers	
[46]	8000 messages	Facebook	Algerian dialect (Arabic and Arabizi)	Yes, automatically	
[47]	151,548 tweets	Twitter	MSA and Egyptian Dialect	Yes, manually	
[48]	Around 6 million tweets	Twitter	MSA and Arabic Dialects	Yes, automatically	

Table 1 illustrates some of the research and studies that involve constructing a dataset for Arabic Dialects texts. Table 2 below illustrates some of the research and studies that involve constructing a Lexicon for Arabic Dialects texts.

Table 2: research and studies that involve constructing a Lexicon for Arabic Dialects texts

	Data size	Platform or Source	Dialect type	Method
[38]	8000 tweets	Twitter	Arabic and Egyptian	lexicon was created by the human annotators from tweets
[7]	1527 idioms and 7358 Words	From NileULex [51]	MSA and Egyptian Dialect	Extend NileULex which is an Arabic sentiment lexicon
[51]	NileULex: 5953 unique terms	Perivoulsy available lexicons	MSA and Egyptian Dialect	Automatically and manually update to lexicons that are developed earlier
[37]	2817 lexemes	(NC) Al Arabiyya News Facebook page posts, and (AC) collected from The Voice Facebook page	Arabic Dialects	Lexemes were extracted from the posts
[50]	words lexicon, idioms lexicon, emoticon lexicon and special intensified word lexicon	Constructed from different datasets based on manual annotation	Arabic Dialects	Lexicons were built by tool dynamically and expanded incrementally over time.
[41]	Keywords lexicon (L1): 2380 negative polarity and 713 positive polarity	Arabic and Egyptian Lexicon	Algerian Vernacular Arabic	Lexicons built by translating words from Arabic and Egyptian Lexicons
	Negation words lexicon (L2)	MSA dictionary		
	Intensification words Lexicon (L3)	MSA dictionary		
[52]	2000 topics (1000 tweets and 1000 comments)	Twitter and Arabic microblogs	MSA and Egyptian Dialect	Basic lexicon is manually collected and annotated, then synonym set and antonym set are used for automatic expansion of the lexicon
[53]	25086 words	Dialect Lexicon	Algerian dialect	Construct a dialect lexicon then merge two lexicons (a dialect and a sentiment lexicon)
[54]	AIPSeLEX: 3632 idioms/ proverbs	Websites and books	MSA and Egyptian Dialect	Collected and annotated manually at sentence level

2.3.2.1. Arabic Dialect lexicon

Great researches and studies have been conducted in the field of creating lexicon for the English Language texts that can be used for the NLP domain while a limited number of research papers considered creating lexicon for Arabic Language either in its Modern Standard Arabic form or Arabic Dialects form. In [37], the author defined lexicon as a set of lexemes utilized for text classification. In [49], the author created a lexicon of MSA form of Arabic Language. The created lexicon used for text classification and the accuracy was high and reached around 97% of classification accuracy. In [37], the author successfully developed a lexicon that can be utilized for Sentiment Analysis. In [50], the author utilized 5 datasets for lexicon construction. All the utilized datasets are constructed from Twitter i.e. consisted of tweets that are annotated. the generated lexicon is dynamic as it is updated automatically to include new words. Table 2 below illustrates some of the research and studies that involve constructing a Lexicon for Arabic Dialects texts.

2.3.3. Language Identification: Arabic Dialect Identification

Language Identification involves the automatic identification of the language from speech or text. Researchers are increasingly interested in exploring approaches for dialects identification. Arabic Dialect Identification involves dialect automatic identification either dialectal text identification or dialectal speech identification. Some of the researches and studies in the domain of dialect identification are mentioned below. In [55] authors identified and classified Arabic Dialects text of 25 cities of the Arab world. Results were promising as the accuracy of the developed system was 67.9% for sentences of about 7 words length and 90% accuracy in the case of utilizing 16 words. Additionally, in [71] the author utilized an annotated dataset of online newspaper contents to train classifiers for the identification of Arabic dialects. The proposed system determines whether the given sentence is in Modern Standard Arabic form or Gulf, Levantine, Egyptian, Iraqi, Maghrebi dialects forms.

2.3.4. Semantic-level Analysis

The semantic-level analysis involves Machine Translation and Sentiment Analysis. In this literature survey, the author concentrated on Sentiment Analysis for Arabic dialects.

2.3.4.1. Sentiment Analysis Literature

One of the most common implementations that involve the use of NLP is Sentiment Analysis (SA). SA involves classifying text to describe whether its expressions are positive or negative. In some cases, the text is classified into positive, negative, or neutral. In [38], the author mentioned that Sentiment Analysis involves the text classification based upon its polarity or emotion. In [17], the author stated that people recently tend to express their thoughts, ideas, and opinions about products, services, etc. on websites, blogs, social media, and many other channels through the web. This massive content generated by users all over the world attracted NLP researchers. In [34], the author mentioned how Sentiment Analysis is important for investigating public attitudes toward product or services, Sentiment Analysis as well can be used for exploring wider public opinions. In [56], the author agreed that online websites and applications recently considered as a valuable source of opinions that can benefit business owners, services providers as well as customers who aim to explore public reviews about different products or facilities, etc. In Arab world, Arab people usually tend to use dialect language in their daily life rather than MSA form. Moreover, Arab people express their ideas and opinions as well thoughts through the web most commonly using their dialectal form of language which results in generating a massive amount of dialectal Arabic texts that are considered a challenge for ANLP researchers.

2.3.4.1.1. Sentiment Analysis Approaches

There are different approaches used for Sentiment Analysis; Lexicon-based approach for Sentiment Analysis in which lexicon is utilized, Machine learning approach for Sentiment Analysis, or in some cases, researchers utilized an approach that is a combination of both Lexicon-based approach and Machine Learning approach. In [7], the author mentioned that the Sentiment Analysis approaches are the Lexicon-based approach, machine learning approach, and hybrid approach which is a mix of both approaches. In this literature review, the author targeted research papers that are published in the period from 2014 onward. The databases the author utilized are IEEE, Springer, ScienceDirect, ACM, and WorldCat. The keywords mentioned below have been used for collecting the research papers:

- “Arabic Dialects” and “Sentiment Analysis”
- “Arabic Dialects” and “Sentiment Analysis” and “approach”

The inclusion criteria for research papers:

- Must involve Sentiment Analysis experiment study
- Must be for Arabic Dialect texts only.
- Must be published in the period from 2014 onward.
- Research paper that is published in journal or conference.

Table 3 below presents a comparative summary between the different approaches that are used so far in recent researches and studies for Sentiment Analysis of Arabic Dialects.

Table 3: Approaches that are used so far in recent researches and studies for Sentiment Analysis of Arabic Dialects

Ref	Data size	Platform	Dialect type	Features	Approach	Performance			
						Accuracy	**Precision**	**Recall**	**F-Measure**
[38]	8000 tweets	Twitter	Arabic and mostly Egyptian	unigrams, bigrams and trigrams	Lexicon-based□ approach	0.665	0.485	0.845	N/A
					machine learning approach: Naive Bayes	0.7	0.47	0.35	N/A
[34]	BRAD 2.0 : 692586 annotated reviews	Arabic Book Reviews from www.goodreads.com	MSA and DA (Egyptian, Levantine, and Gulf)	unigrams and bigrams	Naive Bayes (NB)	87.14	N/A	N/A	N/A
					Decision Tree (DT)	83.80	N/A	N/A	N/A
					Random Forest (RF)	86.18	N/A	N/A	N/A
					XGBoost	88.71	N/A	N/A	N/A
					Support Vector Machines (SVM)	90.68	N/A	N/A	N/A
					Convolutional Neural Networks (CNN)	89.61	N/A	N/A	N/A
					Recurrent Neural Networks (RNN)	90.05	N/A	N/A	N/A
[7]	2067 tweets	Twitter	MSA and Egyptian Dialect	Ngrams	Lexicon + Look-Up stemmer	82.58	N/A	N/A	N/A
					Lexicon + Look-Up stemmer + idioms list	90.8	N/A	N/A	N/A
					Lexicon + Look-Up stemmer + idioms list + SVM	96	N/A	N/A	N/A
[50]	No of tweets: 1-ASTD: 9,174 2-MASTD: 1850 3-ArSAS: 19,762 4-GS: 4,191 5-Syrian Tweets	Twitter	Arabic Dialects	Feature Vector	Hybrid system apply Machine Learning approaches and Lexicon based approaches Unbalanced 3-Class Results	73.67 (RNN)	N/A	N/A	68.57 (L2R2)
					Balanced 3-Class Results	66.83 (L2R2)	N/A	N/A	66.55 (L2R2)
					Unbalanced 2-Class Results	83.73 (L2R2)	N/A	N/A	82.03 (L2R2)
					Balanced 2-Class Results	79.87 (SVM)	N/A	N/A	79.86 (SVM)

	Corpus: 2000 6- ArTwitter: 2000 tweets					Lexicon Update Results	85 (RNN)	N/A	N/A	84.95 (RNN)
[40]	75,774 positive tweets and 75,774 negative tweets	Twitter	Arabic Dialects	Ngrams	ML classifiers	NB	95.91	96.15	95.90	95.90
						BNB	98.00	98.00	98.00	98.00
						MNB	98.19	98.21	98.18	98.18
						ME	94.18	94.44	99.31	94.17
						Ada-Boost	73.76	78.95	73.75	72.53
						SVM	99.31	99.32	99.31	99.31
						LR	98.96	98.98	98.97	98.97
						SGD	99.11	99.12	99.11	99.11
						RR	99.96	99.96	99.96	99.96
						PA	99.96	99.96	99.96	99.96
[17]	-Health services dataset: 2026 tweets	Twitter	Arabic Dialects	TF, TF-IDF, POS, Lex, Auto-Lex	-Word2Vec model used to build an Automatic Arabic Lexicon that used with different Machine Learning methods - Word2Vec used apart from of the lexicon in Convolutional Neural Networks in order to expand the vocabularies		0.85 to 0.92	N/A	N/A	N/A
	-subset of the dataset: 1732 tweets						0.87 to 0.95	N/A	N/A	N/A
[41]	7698 comments	Facebook	Algerian Vernacular Arabic	Code-Switched	Lexicon Based Approach		79.13%	N/A	N/A	N/A
				French encoded in Arabic letters						
				combination of the two first features						
				words written in a form that most Algerians generally used						
[57]	7287 comments	Tunisian dialect-TSAC corpus	Algerian Dialect	Bag of Words (BOW), word2vec and doc2vec	Best result on different corpus are presented in tis table	Shallow Learning	N/A	94%	86%	78%
	10254 comments	Morocco dialect-ElecMorocco2016								
	200 words	Magrebi dialect-Northafrica corpus				Deep Learning	N/A	84%	83%	83%
[42]	1800 tweets	Twitter	MSA and Jordanian Dialect	N-grams TF-IDF TF	NB		83.61	79.38	93.11	84.73
					SVM		88.72	92.10	84.89	88.27
[58]	(ASTD): 10006 tweets	Twitter	Arabic Dialect	-	narrow convolutional neural network (NCNN): classification at sentence level		N/A	N/A	N/A	75.90
[45]	SANA: 194 negative comments & 194 positive comments	Three Algerian newspapers	Algerian Dialect	N-grams TF-IDF TF TO BTO	SANA dataset	SVM	72.16	N/A	N/A	N/A
						NB	75.00	N/A	N/A	N/A
						KNN	66.49	N/A	N/A	N/A
	OCA: 250 positive reviews and 250 negative reviews				OCA dataset	SVM	82.80	N/A	N/A	N/A
						NB	89,80	N/A	N/A	N/A
						KNN	88.40	N/A	N/A	N/A
[59]	22550 tweets	Twitter	MSA and Jordanian Dialect	-	Dialectal words translated	SVM	N/A	0.878	0.868	0.867
						NB	N/A	0.905	0.997	0.876

					into MSA form using Dialects lexicon then classifiers were used					
[60]	3550 tweets	Twitter	Jordanian Dialect	N-grams	SVM		0.84	0.85	0.84	0.84
					NB		0.58	0.71	0.58	0.54
[61]	17,573 tweets	Twitter	MSA and Saudi Dialect	Semantic, Stylistic, and Tweet-specific features	Lexicon-based approach (calculates TweetScore) and corpus-based approach (SVM along with features)		N/A	N/A	N/A	69.9
[62]	Five datasets	Twitter and Facebook	MSA and Arabic Dialects	N-gram Embeddings	Embeddings were composed and learned using unordered composition function and a shallow neural model.		88.2	87.4	88.4	87.8
[63]	3 datasets	Twitter and Facebook	MSA nd Tunisian Dialect	N-grams TF	Lexicon-Based Approach		81.9	83.2	83.4	81.9
					Supervised approach	SVM	94.0	93.9	93.8	93.9
						NB	82.7	83.9	82.5	82.5
[47]	151,548 tweets	Twitter	MSA and Egyptian Dialect	TF-IDF	NB		95.98	96.22	95.98	95.97
					AdaBoost		72.83	77.27	72.82	71.66
					SVM		98.94	98.95	98.94	98.94
					ME		94.22	94.48	94.22	94.21
					RR		99.90	99.90	99.90	99.90

* only best classification results in term of accuracy, recall, Precision, and F-measure are included in this table

2.3.4.1.2. Applications of Sentiment Analysis

Companies, Government authorities, institutions as well show great interest in Sentiment Analysis. In [64], the author explained in detail some of the most common applications of Sentiment Analysis. In the field of business, Sentiment Analysis can be utilized for consumer reviews analysis. Such implementations of Sentiment Analysis witnessed in Google Product search and Amazon websites. Moreover, Business owners and companies value the information retrieved from Sentiment Analysis as it would positively affect their production and help them apply required improvements. On the other hand, in the business field, Sentiment Analysis can be utilized for advertising and commerce online as well as for brand reputation. While in the political field, Sentiment Analysis can be used for monitoring public opinions about government practices and services provided. Sentiment Analysis can be utilized as well in the finance field to monitor financial situations and avoid financial risks. These are some of the applications in which Sentiment Analysis can be effectively used.

2.3.4.1.3. Sentiment Analysis of Arabic Dialect

Huge works for Sentiment Analysis have been conducted and targeted the English language, moreover, researches in the field of Sentiment Analysis for the Arabic Language are increasing as well. In [56], the author stated that limited works and researches have been conducted for Arabic Sentiment Analysis due to many reasons such as the morphological complexity nature of the Arabic Language, the requirement for pre-processing, feature representation, spam opinion elimination and handling the negation in Arabic language. The authors explained how the Arabic language has complex morphological nature such as words with different meanings that may have the same root. Sentiment Analysis for Arabic dialects has been addressed by several research papers. Below are some of the researches and studies that investigated Sentiment Analysis for Arabic Dialects. In [7], the author perform automatic extraction of opinions over social media that are written in MSA and Egyptian Dialects, Authors analyzed Sentiment automatically into either positive or negative. In [50], the author successfully generated hybrid system that can be utilized for Sentiment Analysis for Arabic language. The developed system offered high accuracy and great performance as lexicon was generated from five datasets and it intelligently allows for an automatic update to include new words. In [65], the author utilized OCA freely available corpus and generated ARMD corpus, both are for movie reviews analysis. The authors utilized both supervised and unsupervised approaches for Sentiment Analysis, after that, the authors combined both approaches. The authors reported that the hybrid approach in which supervised and unsupervised methods are used offered the best results in terms of precision, recall, and F-measure. In [22], the author stated that the most common classifiers for Sentiment Analysis of Arabic language are Support Vector Machine and Naïve Bayes. Authors found that the hybrid approach for Sentiment Analysis presented the best results in terms of preciseness both at the document level and sentence level. In [1], the author proposed a rule-based stemmer that can be utilized for gulf dialects. The performance of the offered stemmer is better than other algorithms. The offered stemmer as well showed acceptable accuracy. In [38], the author presented valuable efforts in creating a web-based tool that can be utilized for Sentiment Analysis of Arabic text. The presented web-based tool was developed using the R language and it showed good performance in term of Accuracy. To perform Sentiment analysis for Arabic dialects, some important steps should be taken into consideration such as Pre-Processing and Feature Extraction.

2.3.4.1.3.1.Pre-Processing

Pre-processing is a critical step; sometimes it is referred to as normalization and it involves transforming the word into its standard form. In [38], the author define the pre-processing step as the process of cleaning data to reduce errors and improve Sentiment Analysis performance. In [34], the author mentioned that the pre-processing step for the dataset would allow classifiers to efficiently learned the dataset. In [16], the author stated that pre-processing is essential for Arabic Natural Language Processing implementations such as sentiment analysis and summarization tasks. Authors explained that pre-processing for Dialectal Arabic involves the following steps: Tokenization, Remove Diacritics, remove non-Arabic words and letters, Remove Punctuations, replace Arabic Letters (أ،إ،/آ), (ة), (ىئ،ي) and (ؤ) with (ا), (ه), (ي) and (و) respectively. [56] as well described the pre-processing steps and mentioned that it involves Tokenization, non-Arabic words removal, Normalization, stop words removal, and light stemming. In [7], the author stated that the steps of pre-processing involve Tokenization which means text splitting into separated words, Normalization which involve return letters into the same form, all stop words are removed, and finally words stemming.

2.3.4.1.3.2. Feature Extraction

Sentiment Analysis involves text classification. The classification of texts requires the selection and extraction of text features. Features are the classifier's input. In [39], the author stated that feature selection is the process of extracting features that would affect the classification process. In [38], the author explained how features can be utilized for analyzing raw data. In [66], the author stated that features include part of Speech, frequency, opinion words, and negation. In [38], the author stated that the most common features utilized are N-grams which is frequency (terms presence) features, and the most commonly utilized type of N-grams is unigram followed by bigram and trigram. In [34], the author stated that bigram features consider two words, and these words most commonly come together. Authors mentioned as well that bigram tokens can be effectively utilized for negation detection for either MSA or AD as well. In [55], the author effectively extracted words n-grams and characters n-grams and utilized them as features for AD identification.

2.3.4.1.4. Machine Learning Approaches for Sentiment Analysis of Arabic Dialects

Machine Learning techniques have been widely used for Sentiment Analysis purposes for Many Languages. Machine Learning techniques as well have been used for Sentiment Analysis for the Arabic Language. In [22], the author stated that machine learning techniques can be used in sentiment analysis for Arabic text and SVM presented good performance when used for the sentiment analysis of Arabic texts. In [40], the author utilized Machine learning approaches for Sentiment Analysis for Arabic Dialects. The authors utilized different classifiers for Sentiment Classification of a labeled dataset and reported that PA and RR classifiers presented the best results in terms of accuracy, recall, F-measures, and precision. However, in [6] the author stated that the utilization of Deep Learning approaches recently for NLP tasks presented better performance and results. Table 3 above presents a comparative summary between the different approaches that are used so far in recent researches and studies for Sentiment Analysis of Arabic Dialects.

2.3.4.1.5. Sentiment Analysis for Arabic Dialects Challenges

In [22], the authors mentioned several challenges encountered while working with ANLP such as its complexity. Moreover, fewer works and researches have been done in the field of the Arabic Language compared to English language. In [10], the author presented how Dialectal Arabic does not have a standard orthographic written form. In [3], the author described how dialectal Arabic has no standard written form which results in a lack of NLP tools for Arabic Dialects. In [1]. the author mentioned challenges while working with Arabic dialects; Arabic dialects have no standard written format, moreover Arabic Dialects have complicated morphological forms. In [7], the author agreed that the Arabic language has complex nature; for MSA each word has a root and the task of finding the root for words is not easy and may reduce the accuracy, moreover, the Dialectal Arabic represents the language of different regions, meaning that each Dialect has its collection of words and this would add further challenges to Dialectal Arabic processing and analyzing tasks. In [4], the author mentioned several challenges while dealing with the Arabic language, first, Diacritical may change the meaning of the same word, second, the negation in Arabic may be challenging compared to English language in which negation is presented mostly using the prefix, moreover, the use of dialectal Arabic may present spelling errors since there is no standard written form for Arabic dialects.

3. Conclusion

Social media attracts people all around the world. Due to the increased utilization of Social Media, a massive amount of written text is generated daily and considered as a valuable source of information that attracted business owners, companies, government institutions, and of course, it attracts researchers and data scientists as well. Natural Language Processing NLP is an important field of science that involves studying and analyzing Natural language texts. Increasing efforts were presented in investigating and analyzing the Modern Standard form of Arabic Language as well as the Arabic Dialects. This literature review aims to explore researches that involve constructing resources for Arabic Dialects and investigate approaches and methods used for Sentiment Analysis of Arabic Dialects text, focusing on machine learning approaches and Lexicon-based approaches.

4. References

[1] B. Abuata, A. Al-Omari, "A rule-based stemmer for Arabic Gulf dialect," Journal of King Saud University - Computer and Information Sciences, **27**(2), 104–112, 2015, doi:10.1016/j.jksuci.2014.04.003.

[2] P.D. Kilmer, "Review Article: Review Article," Journalism: Theory, Practice & Criticism, **11**(3), 369–373, 2010, doi:10.1177/1461444810365020.

[3] F. Mallek, B. Belainine, F. Sadat, "Arabic Social Media Analysis and Translation," Procedia Computer Science, **117**, 298–303, 2017, doi:10.1016/j.procs.2017.10.121.

[4] L. Almuqren, A.I. Cristea, "Framework for sentiment analysis of Arabic text," HT 2016 - Proceedings of the 27th ACM Conference on Hypertext and Social Media, 315–317, 2016, doi:10.1145/2914586.2914610.

[5] A.B. Soliman, K. Eissa, S.R. El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP," Procedia Computer Science, **117**, 256–265, 2017, doi:10.1016/j.procs.2017.10.117.

[6] A. Elnagar, R. Al-Debsi, O. Einea, "Arabic text classification using deep learning models," Information Processing and Management, **57**(1), 2020, doi:10.1016/j.ipm.2019.102121.

[7] H. H. Mustafa, A. Mohamed, D. S. Elzanfaly, "An Enhanced Approach for Arabic Sentiment Analysis," International Journal of Artificial Intelligence & Applications, **8**(5), 1–14, 2017, doi:10.5121/ijaia.2017.8501.

[8] A. Alshutayri, E. Atwell, "A social media corpus of Arabic dialect text," Computer-Mediated Communication and Social Media Corpora. Clermont-Ferrand: Presses Universitaires Blaise Pascal, 1–23, 2019.

[9] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, D. Nouvel, "Arabic natural language processing: An overview," Journal of King Saud University - Computer and Information Sciences, (xxxx), 2019, doi:10.1016/j.jksuci.2019.02.006.

[10] S. Khalifa, N. Habash, D. Abdulrahim, S. Hassan, "A large scale corpus of Gulf Arabic," Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, 4282–4289, 2016.

[11] L. Lulu, A. Elnagar, "Automatic Arabic Dialect Classification Using Deep Learning Models," Procedia Computer Science, **142**, 262–269, 2018, doi:10.1016/j.procs.2018.10.489.

[12] S.J. Pan, J.T. Kwok, Q. Yang, "Transfer learning via dimensionality reduction," Proceedings of the National Conference on Artificial

Intelligence, **2**, 677–682, 2008.

[13] Y. Yoshida, T. Hirao, T. Iwata, M. Nagata, Y. Matsumoto, "Transfer learning for multiple-domain sentiment analysis - Identifying domain dependent/independent word polarity," Proceedings of the National Conference on Artificial Intelligence, **2**, 1286–1291, 2011.

[14] P. Shapiro, K. Duh, "Morphological Word Embeddings for Arabic Neural Machine Translation in Low-Resource Settings," 1–11, 2018, doi:10.18653/v1/w18-1201.

[15] A. Erdmann, N. Zalmout, N. Habash, "Addressing noise in multidialectal word embeddings," ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), **2**, 558–565, 2018, doi:10.18653/v1/p18-2089.

[16] E.H. Almansor, A. Al-Ani, "Translating dialectal Arabic as low resource language using word embedding," International Conference Recent Advances in Natural Language Processing, RANLP, **2017-Septe**, 52–57, 2017, doi:10.26615/978-954-452-049-6-008.

[17] A.M. Alayba, V. Palade, M. England, R. Iqbal, "Improving Sentiment Analysis in Arabic Using Word Representation," 2nd IEEE International Workshop on Arabic and Derived Script Analysis and Recognition, ASAR 2018, 13–18, 2018, doi:10.1109/ASAR.2018.8480191.

[18] M. Elrazzaz, S. Elbassuoni, C. Helwe, K. Shaban, "Methodical evaluation of Arabic word embeddings," ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), **2**, 454–458, 2017, doi:10.18653/v1/P17-2072.

[19] A. Erdmann, N. Zalmout, N. Habash, "Addressing noise in multidialectal word embeddings," ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), **2**, 558–565, 2018, doi:10.18653/v1/p18-2089.

[20] N. Habash, M. Diab, O. Rambow, "Conventional orthography for dialectal Arabic," Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, 711–718, 2012.

[21] I. Zribi, R. Boujelbane, A. Masmoudi, M. Ellouze, L. Belguith, N. Habash, "A conventional orthography for tunisian Arabic," Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, 2355–2361, 2014.

[22] A. Alsayat, N. Elmitwally, "A comprehensive study for Arabic Sentiment Analysis (Challenges and Applications)," Egyptian Informatics Journal, (xxxx), 4–9, 2019, doi:10.1016/j.eij.2019.06.001.

[23] N. Habash, O. Rambow, "M Agead : C #@," **M**(July), 681–688, 2006.

[24] N. Habash, R. Eskander, A. Hawwari, "A Morphological Analyzer for Egyptian Arabic," Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology SIGMORPHON2012, 1–9, 2012.

[25] I. Zribi, M.E. Khemakhem, L.H. Belguith, "Morphological Analysis of Tunisian Dialect," International Joint Conference on Natural Language Processing, (October), 992–996, 2013.

[26] S. Khalifa, N. Habash, F. Eryani, O. Obeid, D. Abdulrahim, M. Al Kaabi, "A morphologically annotated corpus of Emirati Arabic," LREC 2018 - 11th International Conference on Language Resources and Evaluation, 3839–3846, 2019.

[27] A. Mekki, I. Zribi, M. Ellouze, L.H. Belguith, "Syntactic analysis of the Tunisian Arabic," CEUR Workshop Proceedings, **1988**, 2017.

[28] A. Shahrour, S. Khalifa, N. Habash, "Improving Arabic diacritization through syntactic analysis," Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, (September), 1309–1315, 2015, doi:10.18653/v1/d15-1152.

[29] K. Dukes, E. Atwell, A.B.M. Sharaf, "Syntactic annotation guidelines for the quranic Arabic dependency treebank," Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, 1822–1827, 2010.

[30] A. Shahrour, S. Khalifa, D. Taji, N. Habash, "CamelParser: A system for Arabic syntactic analysis and morphological disambiguation," COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: System Demonstrations, 228–232, 2016.

[31] M. Abdul-Mageed, M.T. Diab, "Subjectivity and sentiment annotation of modern standard Arabic newswire," ACL HLT 2011 - LAW 2011: 5th Linguistic Annotation Workshop, Proceedings, (3), 110–118, 2011.

[32] M. Abdul-Mageed, M. Diab, "AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis," Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, 3907–3914, 2012.

[33] A. Abdelali, J. Cowie, H. Soliman, "Building A Modern Standard Arabic Corpus," Workshop on Computational Modeling of Lexical Acquisition. The Split Meeting. Croatia, 25th to 28th of July, 2005.

[34] A. Elnagar, L. Lulu, O. Einea, "An Annotated Huge Dataset for Standard and Colloquial Arabic Reviews for Subjective Sentiment Analysis," Procedia Computer Science, **142**, 182–189, 2018, doi:10.1016/j.procs.2018.10.474.

[35] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, Y. Al-Ohali, "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets," Procedia Computer Science, **117**, 63–72, 2017, doi:10.1016/j.procs.2017.10.094.

[36] O. Newspapers, E. Twitter, O. Newspapers, I. Proceedings, A. Alshutayri, E. Atwell, "This is a repository copy of Creating an Arabic Dialect Text Corpus by Exploring Twitter , Version : Accepted Version Creating an Arabic Dialect Text Corpus by Exploring Twitter , Facebook , and Online Newspapers," 2018.

[37] M. Itani, C. Roast, S. Al-Khayatt, "Developing Resources for Sentiment Analysis of Informal Arabic Text in Social Media," Procedia Computer Science, **117**, 129–136, 2017, doi:10.1016/j.procs.2017.10.101.

[38] M. El-Masri, N. Altrabsheh, H. Mansour, A. Ramsay, "A web-based tool for Arabic sentiment analysis," Procedia Computer Science, **117**, 38–45, 2017, doi:10.1016/j.procs.2017.10.092.

[39] S. Hussein, M. Farouk, E.S. Hemayed, "Gender identification of egyptian dialect in twitter," Egyptian Informatics Journal, **20**(2), 109–116, 2019, doi:10.1016/j.eij.2018.12.002.

[40] D. Gamal, M. Alfonse, E.S.M. El-Horbaty, A.B.M. Salem, "Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features," Procedia Computer Science, **154**, 332–340, 2018, doi:10.1016/j.procs.2019.06.048.

[41] M. Mataoui, O. Zelmati, M. Boumechache, "A Proposed Lexicon-Based Sentiment Analysis Approach for the Vernacular Algerian Arabic," Research in Computing Science, **110**(1), 55–70, 2016, doi:10.13053/rcs-110-1-5.

[42] K.M. Alomari, H.M. Elsherif, K. Shaalan, Arabic tweets sentimental analysis using machine learning, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **10350 LNCS**, 602–610, 2017, doi:10.1007/978-3-319-60042-0_66.

[43] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj, K.B. Shaban, "ArSentD-LEV: A Multi-Topic Corpus for Target-based Sentiment Analysis in Arabic Levantine Tweets," (1), 2019.

[44] M. Jarrar, N. Habash, F. Alrimawi, D. Akra, N. Zalmout, "Curras: an annotated corpus for the Palestinian Arabic dialect," Language Resources and Evaluation, **51**(3), 745–775, 2017, doi:10.1007/s10579-016-9370-7.

[45] H. Rahab, A. Zitouni, M. Djoudi, "SANA: Sentiment analysis on newspapers comments in Algeria," Journal of King Saud University - Computer and Information Sciences, (xxxx), 2019, doi:10.1016/j.jksuci.2019.04.012.

[46] I. Guellil, A. Adeel, F. Azouaou, A. Hussain, "SentiALG: Automated Corpus Annotation for Algerian Sentiment Analysis," Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **10989 LNAI**(Ml), 557–567, 2018, doi:10.1007/978-3-030-00563-4_54.

[47] D. Gamal, M. Alfonse, E.-S. M.El-Horbaty, A.-B. M.Salem, "Twitter Benchmark Dataset for Arabic Sentiment Analysis," International Journal of Modern Education and Computer Science, **11**(1), 33–38, 2019, doi:10.5815/ijmecs.2019.01.04.

[48] H. Abdellaoui, M. Zrigui, "Using tweets and emojis to build TEAD: An arabic dataset for sentiment analysis," Computacion y Sistemas, **22**(3), 777–786, 2018, doi:10.13053/CyS-22-3-3031.

[49] F.H.H. Mahyoub, M.A. Siddiqui, M.Y. Dahab, "Building an Arabic Sentiment Lexicon Using Semi-supervised Learning," Journal of King Saud University - Computer and Information Sciences, **26**(4), 417–424, 2014, doi:10.1016/j.jksuci.2014.06.003.

[50] K. Elshakankery, M.F. Ahmed, "HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis," Egyptian Informatics Journal, **20**(3), 163–171, 2019, doi:10.1016/j.eij.2019.03.002.

[51] S.R. El-Beltagy, "NileULex: A phrase and word level sentiment lexicon for Egyptian and modern standard Arabic," Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, 2900–2905, 2016.

[52] H.S. Ibrahim, S.M. Abdou, M. Gheith, "Automatic expandable large-scale sentiment lexicon of modern standard Arabic and colloquial," Proceedings - 1st International Conference on Arabic Computational Linguistics: Advances in Arabic Computational Linguistics, ACLing 2015, (July 2016), 94–99, 2016, doi:10.1109/ACLing.2015.20.

[53] I. Guellil, F. Azouaou, "Bilingual Lexicon for Algerian Arabic Dialect Treatment in Social Media," WiNLP, 1–4, 2017.

[54] H. S.Ibrahim, S. M. Abdou, M. Gheith, "Idioms-Proverbs Lexicon for Modern Standard Arabic and Colloquial Sentiment Analysis," International Journal of Computer Applications, **118**(11), 26–31, 2015, doi:10.5120/20790-3435.

[55] M. Salameh, H. Bouamor, N. Habash, "Fine-Grained Arabic Dialect Identification," Processdings of the 27th International Conference on Computational Linguistics Santa Fe, New Mexico, USA, 1332–1344, 2018.

[56] R.M.K. Saeed, S. Rady, T.F. Gharib, "An ensemble approach for spam detection in Arabic opinion texts," Journal of King Saud University -

Computer and Information Sciences, (xxxx), 2019, doi:10.1016/j.jksuci.2019.10.002.

[57] I. Guellil, M. Mendoza, F. Azouaou, "Arabic dialect sentiment analysis with ZERO effort. Case study: Algerian dialect," Inteligencia Artificial, **23**(65), 124–135, 2020, doi:10.4114/intartif.vol23iss65pp124-135.

[58] M. Alali, N. Mohd Sharef, M.A. Azmi Murad, H. Hamdan, N.A. Husin, "Narrow Convolutional Neural Network for Arabic Dialects Polarity Classification," IEEE Access, **7**, 96272–96283, 2019, doi:10.1109/ACCESS.2019.2929208.

[59] R.M. Duwairi, "Sentiment analysis for dialectical Arabic," 2015 6th International Conference on Information and Communication Systems, ICICS 2015, (April), 166–170, 2015, doi:10.1109/IACS.2015.7103221.

[60] J.O. Atoum, M. Nouman, "Sentiment analysis of Arabic Jordanian dialect tweets," International Journal of Advanced Computer Science and Applications, **10**(2), 256–262, 2019, doi:10.14569/ijacsa.2019.0100234.

[61] N. Al-Twairesh, H. Al-Khalifa, A. Alsalman, Y. Al-Ohali, "Sentiment Analysis of Arabic Tweets: Feature Engineering and A Hybrid Approach," 2018.

[62] H. Mulki, H. Haddad, M. Gridach, I. Babaoğlu, "Syntax-Ignorant N-gram Embeddings for Sentiment Analysis of Arabic Dialects," 30–39, 2019, doi:10.18653/v1/w19-4604.

[63] H. Mulki, H. Haddad, C.B. Ali, I. Babaoglu, "Tunisian dialect sentiment analysis: A Natural Language Processing-based Approach," Computacion y Sistemas, **22**(4), 1223–1232, 2018, doi:10.13053/CyS-22-4-3009.

[64] A. D'Andrea, F. Ferri, P. Grifoni, T. Guzzo, "Approaches, Tools and Applications for Sentiment Analysis Implementation," International Journal of Computer Applications, **125**(3), 26–33, 2015, doi:10.5120/ijca2015905866.

[65] B. Brahimi, M. Touahria, A. Tari, "Improving sentiment analysis in Arabic: A combined approach," Journal of King Saud University - Computer and Information Sciences, (xxxx), 2019, doi:10.1016/j.jksuci.2019.07.011.

[66] W. Medhat, A. Hassan, H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, **5**(4), 1093–1113, 2014, doi:10.1016/j.asej.2014.04.011.

[69] Biermann A.W. (1986) Fundamental mechanisms in machine learning and inductive inference. In: Bibel W., Jorrand P. (eds) Fundamentals of Artificial Intelligence. Lecture Notes in Computer Science, **232**, Springer, Berlin, Heidelberg

[70] Abhimanyu Chopra, Abhinav Prashar, Chandresh Sain, 2013. Natural Language Processing. INTERNATIONAL JOURNAL OF TECHNOLOGY ENHANCEMENTS AND EMERGING ENGINEERING RESEARCH, VOL 1, ISSUE 4, ISSN 2347-4289

[71] Omar F. Zaidan and Chris Callison-Burch, 'Arabic Dialect Identification', Computational Linguistics. **40**(1), 171-202, 2014