Soc. Netw. Anal. Min. (2017) 7:54
https://doi.org/10.1007/s13278-017-0474-x

REVIEW ARTICLE

Successes and challenges of Arabic sentiment analysis research: a literature review

Mazen El-Masri[1] · **Nabeela Altrabsheh**[1] · **Hanady Mansour**[1]

Received: 28 May 2017 / Revised: 11 October 2017 / Accepted: 12 October 2017 / Published online: 31 October 2017

Abstract The analysis of sentiment in text has mainly been focused on the English language. The complexity of the Arabic language and its linguistic features that oppose those found in English resulted in the inability to adapt extant research to Arabic contexts limiting advancement in Arabic sentiment analysis. The need for Arabic sentiment analysis research is accentuated by the driving changes in different Arab regions like heavy political movements in some areas and fast growth in others. These changes help shape not just policies and implications of this region but affect the entire world on a global scale. Therefore, it is essential to utilise effective methods of sentiment analysis to analyse Arabic tweets to understand regional and global implications in microblogging mediums such as Twitter. In this paper, we conduct a comprehensive review of Arabic sentiment analysis, present the pros and cons of the different approaches used and highlight the challenges of it. Finally, we outline the relevant gaps in the literature and suggest recommendations for future Arabic sentiment analysis research.

Keywords Arabic sentiment analysis · Arabic · Opinion mining

1 Introduction

In the recent decade, novel microblogging mediums like Twitter, Tumblr, and Instagram became the perfect arena for expression. Participants channel ideas together to convey views of either themselves or to share facts and information to others. This rich and important information and sharing of thoughts create a global network, which allows any contributor the ability to engage in multiple interactions under one umbrella. Recent research has identified it as online word-of-mouth branding (Jansen et al. 2009). Due to the free format of messages and easy accessibility, traditional methods of blogging are fading away and microblogging is standing out as the new and improved version (Pak and Paroubek 2011). Whether it is utilised by economists, marketers, academics, or governmental officials, tools like Twitter and Tumblr offer a thorough development of understanding of a broad set of topics resulting in real-life interaction. Indeed, human, organisational, and governmental activities are somewhat manipulated by an online global social network. This phenomenon provides the research community the power to revolutionise the way research is done.

While abundant data are available on social media, the ability to extract useful information from this colossal pile at our disposal is challenging. Societies, organisations, and governments have a great interest in capturing and aggregating what people think about a particular topic. What the majority think of Donald Trump, Facebook's acquisition of WhatsApp and the new iPhone 8 are important questions in today's world. Sentiment analysis can answer such questions. Researchers and practitioners can use automated text mining and fact-based techniques to determine whether a text is positive or negative. Sentiment analysis makes

This publication was made possible by the NPRP award [NPRP 7-1334-6-039 PR3] from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the author[s].

✉ Nabeela Altrabsheh
nabeela@qu.edu.qa

[1] Qatar University, Doha, Qatar

comprehension of information possible. Research advocates this method as an efficient solution that can provide reliable indicators of public mood. For example, what the majority thinks about a movie can be detected using sentiment analysis techniques. It shows the average public opinion of the movie, as opposed to a haphazard array of reviews without a clear indication of sentiment.

Abundant literature is available on methodologies, challenges, and applications of sentiment analysis. Yet, not much research has been conducted in sentiment analysis of the Arabic language as opposed to English. One determinant for that could be the inappropriateness of replicating English sentiment analysis research in Arabic contexts. Arabic has linguistic features that are not only different but opposing to the English language. The structures and grammar of the Arabic language differ from that English language (Duwairi and El-Orfali 2014). Moreover, Arabic studies must consider a considerable number of different dialects in the Arab world which makes research in this field more complicated. In this research, we provide an overview of the Arabic sentiment analysis literature. Specifically, the aim of this paper is to provide a thorough review of the literature that pertains to:

1. The different preprocessing techniques commonly used in Arabic sentiment analysis;
2. The different features that have been utilised in Arabic sentiment analysis research;
3. The sentiment analysis approach that is the most commonly used in Arabic sentiment analysis; and
4. The challenges that Arabic sentiment analysis researchers face.

This paper is focused on identifying the issues and limitations of Arabic sentiment analysis.

2 Literature review methodology

Web of Science search engine was used as a base for this research. Different keyword terms including ('Arabic' AND 'Sentiment Analysis') and ('Arabic' AND 'Opinion Mining') were searched for. A number of 146 articles were found. The abstracts of these papers were examined. Around 35% of the papers were eliminated and not read further as they were focused on natural language processing. In Fig. 1, we present the articles found using the keywords per year. We found that the articles started in the year 2008 and got more popular in 2012, which is probably due to the increase in social media platforms in the Arab countries in the last years. It could also be due to the unsettling events that occurred in the middle east at that time (i.e. Arab Spring). In the year 2016, the highest spike of articles occurred. This could be due to the increased demand for sentiment analysis research due to the discovery of its importance in the Arab world.

3 Sentiment analysis

Sentiment analysis is an application of natural language processing, computational linguistics, and text analytics techniques to identify and retrieve certain sentiment(s) in text (Mouthami et al. 2013). Its goal is to classify a given text into sentiment polarity, i.e. determining whether the expressed opinion is positive, negative, or neutral (Agarwal et al. 2011).

To conduct sentiment analysis, general techniques can be applied without specifying the domain (Prasad 2010; Agarwal et al. 2011; Wang et al. 2012; Pak and Paroubek 2010; Kumar and Sebastian 2012; Barbosa and Feng 2010; Go et al. 2009a, b). However, previous research suggests

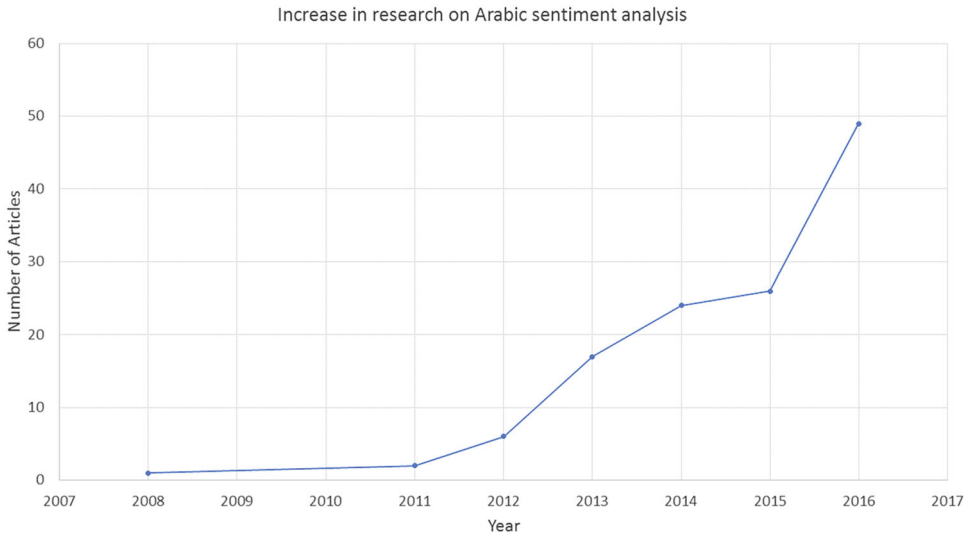

Fig. 1 Increase in research on Arabic sentiment analysis

that sentiment analysis is more effective when applied to specific domains (Vohra and Teraiya 2013; Go et al. 2009a). Word meanings and their sentiments may vary depending on the domain (Yoshida et al. 2011). For example, in Arabic, the word high (عالية) can be negative in the product domain like 'This is a high-cost product' and positive in the places domain like in 'High-quality service' (Al-Kabi et al. 2013).

Sentiment analysis can be applied for different purposes such as political campaigns and riots. However, it is most often applied on consumer reviews from different domains like movies (Asur et al. 2010; Pang et al. 2002; Yessenov and Misailovic 2009), advertisements (Hill et al. 2012), products (Saif et al. 2012), cars (Gamon et al. 2005), smart phones (Chamlertwat et al. 2012), tourism (Claster et al. 2010), and e-learning (Tian et al. 2009; Ortigosa et al. 2014). There have been several studies on sentiment analysis that have used Twitter data. Yet, using such data is challenging due to the typical short length and irregular structure (Saif et al. 2012).

In the past decade, microblogging mediums like Twitter have been intensively used by researchers from various disciplines to study sentiments. Different topics have been explored such as movies, sports, political movements, and mobile applications.

4 Sentiment analysis methodology in Arabic

Sentiment analysis can differ from one language to another. Most of the research on sentiment analysis has been on the English language, and little has been done on the Arabic language. Arabs make around 4.8% of Internet users worldwide (Arabic speaking 2015). The Arabic language can be classified into three categories: classical Arabic, which can be found in religious scripts, modern standard Arabic (MSA), which is the formal written and spoken Arabic and in the media (newspapers, journals, TV, and radio), and informal or dialectical Arabic, which varies from one Arab country to another (Duwairi and El-Orfali 2014; Al-Kabi et al. 2013; Refaee and Rieser 2014). There are six dominant dialects in the Arab world: Egyptian, Moroccan, Levantine, Iraqi, Gulf, and Yemeni (Darwish et al. 2012). Social media sites contain both MSA and informal Arabic (Al-Kabi et al. 2013). Although Arabic is one of the top 10 spoken languages in the world, there is a lack of Arab web content with very few web pages that provide reviews (Shoukry and Rafea 2012). There are major differences between dialects on all levels of linguistic representation: morphology, lexical, phonology, syntax, semantics, and pragmatics (Abdul-Mageed et al. 2014).

Fig. 2 Sentiment analysis methodology

Like English sentiment analysis, there are five main steps to create Arabic sentiment analysis models (see Fig. 2). In Appendix A in Table 5, a summarisation of the Arabic sentiment analysis research is presented. We highlight the methodology steps, pros, and cons for each of the studies. The following subsections explain the steps in the sentiment analysis methodology and literature related to it in more detail.

5 Methodology

There are five main steps to create sentiment analysis models: (1) data gathering (2) preprocessing the data, (3) extracting the features, (4) applying the analysis (lexicon based or machine learning based), and (5) evaluating the results. In Appendix A in Table 5, a summarisation of the Arabic sentiment analysis research is presented. We highlight the methodology steps, pros, and cons for each of the studies. The following subsections explain the sentiment analysis models steps and literature related to it in more detail.

5.1 Data gathering

Opinions can be collected from different sources. Nowadays, the most common source for collecting data is social media websites such as Facebook and Twitter. These are considered multi-domain and contain reviews about different topics. The data could also be collected from domain-specified websites such as Tripadvisor for tourism-related reviews, IMDB for movie reviews, and Amazon for product reviews.

5.2 Preprocessing

Preprocessing is the process of cleaning the data from unwanted elements. It increases the accuracy of the results by reducing errors in the data (Altrabsheh et al. 2014). Not using preprocessing, such as spelling corrections, may lead the system to ignore important words. On the other hand, over using preprocessing techniques may sometimes cause loss of important data. There are general preprocessing techniques and Twitter-related preprocessing techniques. The following are common general preprocessing techniques:

- *Tokenise text*: Breaking a stream of words like a sentence into words, phrases, and characters. Tokenisation was used in Al-Twairesh et al. (2014), Al-Sabbagh and Girju (2012), Abdul-Mageed et al. (2014), and Duwairi et al. (2014).
- *Remove stop words*: Stop words can be words such as 'kan', 'lan', 'fe' كان,لن, في. These words are not helpful to the sentiment. Removal of the stop words will help reduce index space, improve response time, and improve effectiveness. There is not one set of stop words that can be removed. Darwish et al. (2012) constructed a stopword list that combines both 162 MSA stopwords and 90 dialectic stopwords. The dialectic stopwords contained words such as اللي 'Ally' meaning 'that' and مش 'msh' meaning 'not' (Darwish et al. 2012).
- *Remove or identify punctuation*: Removal of punctuation such as full stops and commas as they are irrelevant to the polarity detection (El-Makky et al. 2015). In some cases, punctuation can indicate polarity such as the exclamation mark, which could mean strong polarity strength or the question mark, suggesting confusion (Altrabsheh et al. 2014).
- *Convert text to lower or upper case*: Converting the letters into upper or lower case. If the word is all in capitals, it sometimes suggests strong sentiment/emotion (Prasad 2010; Zhao et al. 2011). In Arabic, there are no upper and lower cases; however, in Arabizi this could be useful.
- *Stemming the word*: Returning the word to the basic form. In non-Arabic languages, a basic stem can be either prefixed or post-fixed to express a grammatical syntax. In the Arabic language, stemming is a challenge as it is difficult to differentiate between root letters and affix letters (Duwairi and El-Orfali 2014). For example, consider the word 'kitab', which means 'book', the alef 'ا' letter here is an affix letter changing the plural of books to a singular form. In Arabic, around 80% of the words are derived from a three-letter root (Duwairi 2007). There are two approaches to stemming: reducing a word to its three-letter root and light stemming, which removes common suffixes and prefixes without reducing a word to its root. Arabic stemmers reduce words to three-letter roots, while some words have four-letter or five-letter roots (Duwairi and El-Orfali 2014).
- *Check the Spelling*: Data from social media and reviews contain many spelling mistakes such as missing letters or extra letters. Spelling can be corrected by removing extra letters such as in Darwish et al. (2012) and Soliman et al. (2014). Rushdi et al. (2011) also corrected spelling mistakes manually.
- *Letter replacement*: The Arabic language has 28 letters. Previous research has replaced forms of letters into another form for the accuracy of the prediction. Some of these letters have variations such as:
  - Alef: Alef includes ا, أ and alef maksoura ى often confused with ya ي;
  - Ta marbota ة is usually confused in writing with the letter ha ه; and
  - Hamza: hamza is interchangeable according to the word and position. It includes ئ, ؤ, آ.

Twitter data require additional preprocessing techniques to be performed. Such data contain emoticons, hashtags, and other chat-related text that can be removed. Some of the common Twitter-specific data preprocessing techniques are: removing hashtags, removing URLs, removing retweets, identifying emoticons, removing user mentions in tweets, removing Twitter special characters, and dialect Arabic language handling. Due to the short length of tweets, some of the words do not occur a lot and are unique. Korayem et al. (2012) added the token 'UNIQUE' where words occur less than four times. There are many studies which replaced dialect words with MSA words such as Duwairi (2015). Indeed, Duwairi (2015) created a mapped lexicon from 100 long Jordanian chats and manually extracted the dialectical words and provided the MSA alternative.

5.3 Features

Features give a more accurate analysis of the sentiments and detailed summarisation of the results (Yoshida et al. 2011). The most common features in sentiment analysis are n-grams (Go et al. 2009a; Agarwal et al. 2011; Wang and Wu 2011) and POS (part of speech) tagging (Pang et al. 2002; Wang et al. 2012). Due to the complexity of the Arabic language, some research explored other linguistic features, which are described in detail below. Additionally, there are other less common features related to lexicons, Twitter, and punctuation.

5.3.1 N-grams

N-grams are sequences of n-items from some text. Letters, syllables, or words are all n-gram items. The most frequently used n-gram items are words, and the most popular n-grams are unigrams (one word), bigrams (two sequent words), and trigrams (three sequent words).

In Arabic sentiment analysis, unigrams have been largely used as features by Saif et al. (2012) and Oraby et al. (2013). Nevertheless, there is no clear answer on which n-grams lead to the best performance (Shoukry and Rafea 2012; Mountassir et al. 2013; Rushdi et al. 2011). Some studies showed that unigrams led to a better performance

than bigrams and trigrams (Saif et al. 2012; Oraby et al. 2013). Shoukry and Rafea (2012) examined the use both unigrams and bigrams as features and found that using bigrams led to no improvement in comparison with the unigrams. Likewise, Mountassir et al. (2013) and Rushdi et al. (2011) examined the combination of unigrams, bigrams and trigrams. They found that trigrams lead to the best performance. Bigrams and trigrams in Arabic could complete the meaning of a statement rather than unigrams alone. For example, the bigram "فاتت الساعة", which means 'One hour passed'. Additionally, the trigram "أليس هذا كافيا" meaning 'Isnt this enough'. It could also show the strength of the emotion expressed; for instance, the bigram "عنيفة اووى", which means 'very violent'. Another example is the bigram "الف مبروك", which is a strong phrase for 'congratulations'.

5.3.2 Stylistic features

Stylistic features include lexical and structural attributes. Some of the stylistic features examined in the context of sentiment analysis are word length distributions, vocabulary richness measures, emoticons, and special character frequencies (Refaee and Rieser 2014; Abbasi et al. 2008). To examine sentiments in Arabic text, Abbasi et al. (2008) used stylistic features like word- and character-level lexical features, word length distributions, special characters, letters, character n-grams, structural features, vocabulary richness measures, digit n-grams, and function words. One example for the stylistic features is the frequency of letters. In English, there are 26 standard letters, while in Arabic there are 28 main letters and 8 modified letters. The authors found that stylistic features lead to a higher performance than word n-grams and POS tags. On the other hand, Refaee and Rieser (2014) used classified emoticons, positive and negative to study sentiments in Arabic text.

5.3.3 Syntactic features

Syntactic features are phrase patterns that help detect sentiments in phrases. These patterns include nouns followed by positive adjective or nouns followed by negative adjective (Fei et al. 2004). The use of syntactic features depends how the sentence is constructed and words are assembled. Refaee and Rieser (2014) used the following syntactic features: n-grams of words and POS tags, lemmas, including Bag of Words (BOW), and Bag of lemmas. On the other hand, Al-Sabbagh and Girju (2012) used transitive vs. intransitive verbs. Abbasi et al. (2008) used the following syntactical features: POS tags, n-grams (for English), word roots (for Arabic), word n-grams, and punctuation. Syntactic features are the most commonly used features in sentiment analysis. One example for the difference in syntactic features is the word roots. In Arabic, the word can have multiple forms all originating from the same root. For instance, the words "سلام, يسلم, سلمت" all come from the root 'سلم'. Abdul-Mageed et al. (2014) used two different configuration settings to extract roots of words which were used as features: (1) Lexeme (LEX) (the set of all forms consisting of the same meaning), surface words are tokenised and the morphotactics at clitic boundaries are handled; (2) Lemma (LEM) (lemma is the exact form chosen to represent the lexeme), words are reduced to their lemma forms, for verbs it is the form of third-person masculine singular perfective, and for nouns it is in the singular default form which is usually masculine.

5.3.4 Part-of-speech features

Part-of-speech (POS) tagging consists of categorising the word into lexical categories. POS tagging, also known as word disambiguation, is the identification and the marking up of words in text according to their nature and their relationships with adjacent and related words in the text. One approach in English POS tagging is to map each word to parts of speech of eight grammatical categories. These categories are the verbs, nouns, pronouns, adverbs, adjectives, prepositions, conjunctions, and interjections. In Arabic, there have been several POS tagging approaches including whole word and segmentation-based tagging which means the tagging of the different segments of the word (Mohamed and Kübler 2010). In the latter approach, POS tags contain information about the morphology of the word. POS tagging has been widely used in Arabic text analysis. For instance, El-Halees (2012) and El-Makky et al. (2015) used POS tagging features such as nouns and proper nouns to examine sentiment in Arabic text. Moreover, tools have been made available to find POS tags in Arabic text. One of the tools is MADA toolkit which has been successful by El-Makky et al. (2015) and Habash et al. (2009) in the context of Arabic sentiment analysis. POS tagging in Arabic is more complex than English POS tagging, as words can be classified into more than one category. For instance, the word 'فهم', which has multiple meanings including 'understanding' and 'them'. This word can be verb, noun, noun–pronoun in a third-person, noun–verb–gerund (a noun subcategory), and verb–perfect (a

verb subcategory) with an active voice (Zeroual et al. 2017).

5.3.5 Semantic features

Semantic features include contextual features which represent the semantic orientation of surrounding text. Semantic features include annotation techniques which add polarity scores to words and phrases. Semantic can measure the overall correlation of a group of entities through different concepts of entities with a given sentiment polarity. Therefore, if an entity has never appeared in the dataset before with the relations of the larger group of entities, the polarity can be detected (e.g. a new iPhone release when the apple product reviews are mostly positive will most likely be positive). Saif et al. (2012) used semantic features to map between entities and their groups. They found that semantic features outperform the unigram and POS tagging features. An example of semantic features in Arabic is the words "السيسي, مرسي" which are Morsi and e El-Sisi related to the Arab spring that happened in Egypt which were more likely to be negative due to the events. A sentence containing Prime Minister(s) Nazif and Shafik can then be mapped to Arab Spring and is more likely to be negative. Al-Sabbagh and Girju (2012) used semantic features including gender, number definiteness agreement between subject and verbs, nouns, and adjectives. Semantic features, unlike syntactic and stylistic features, require human involvement. Generally, researchers and practitioners create dictionaries as semantic feature selection elements. Such semantic features have been found to be very useful for analysing sentiments (Abbasi et al. 2008; Whitelaw et al. 2005).

5.3.6 Lexicon-derived features

There are many other features that could be added to increase model performance. One example is lexicon-derived features such as polarity averages or sums have been commonly used (Zhang et al. 2011; Lu and Tsou 2010). There are many general lexicons for English such as SentiStrength. SentiStrength is a lexicon of 2310 sentiment words and word stems classified into positive and negative scores from 1 to 5 (Thelwall 2013). In the Arabic language, there exist lexicons that were automatically created from Twitter using emoticons and translated English lexicons (Mohammad et al. 2015). Lexicons for the Arabic language would be different than English due to the multiple dialects and the multiple forms of words originating from a single root word. Moreover, it would be larger than English sentiment lexicons.

5.3.7 Other features

There are other features that have been used in Arabic sentiment analysis. One type is Twitter-related features, which could be the number of emoticons or the number of hashtags (Al-Twairesh et al. 2014; Mohammad et al. 2015). Other less common features which could be used in any domain are the number of punctuation signs such as question marks and exclamation marks which could hold some value to the sentiment (Yassine and Hajj 2010). Other features could be related to the domain itself, for instance, in products the time of the review since the purchase of the product (Ibrahim et al. 2015). Additionally, in education, it may be the type of lecture: practical or theoretical or the timing of the lecture: early or late (Altrabsheh et al. 2013). Other features that can be used in Arabic are the dialect type (e.g. Gulf or Egyptian).

1. Unique(Q) and Polarity lexicon (PL) features (Korayem et al. 2012; Abdul-Mageed et al. 2014). The Unique feature replaces low-frequency words with the token 'UNIQUE'. As for the Polarity lexicon (PL) features, they search in a manually created lexicon of 3982 adjectives labelled with positive, negative, and neutral. To determine the subjectivity, Abdul-Mageed et al. (2014) search for adjectives and to extract the polarity, they search for the positive and negative adjectives.
2. The dialectal Arabic features explore adding tags on the Twitter dataset to represent whether the tweet is in MSA or a dialect (Abdul-Mageed et al. 2014).
3. The genre-specific features explore gender (GEN) features corresponding to the set MALE, FEMALE, UNKNOWN. User ID features from the set PERSON, ORGANISATION, and a document ID (DID) feature (Abdul-Mageed et al. 2014).

5.4 Choosing and applying sentiment analysis method

There are three main methods to sentiment analysis. They are the (1) machine learning, (2) lexicon-based, and (3) hybrid or combined methods. These methods are described below.

5.5 Machine learning method

Machine learning method is the most common method used in sentiment analysis. This method uses classifiers to automatically detect the labels of the new data. It can be used only when the dataset is annotated. Many researchers annotate the data manually (Shoukry and Rafea 2012; Al-Subaihin et al. 2011). Some annotate it automatically using

Table 1 Summary of Machine-based sentiment analysis

Paper	Language	Dataset source	Preprocessing	Features	Classifier	Results (accuracy)
Duwairi and El-Orfali (2014)	MSA	Aljazeera website Movie dataset	Stemming and light stemming	N-grams in levels, words, and characters	SVM, K-NN, and NB	NB: 96.6%: Movie dataset and 86%: Politics dataset
Al-Kabi et al. (2013)	Colloquial Arabic and MSA	Social media and news sites	Removed: transliterated Arabic words Arabizi non-alphabet characters, normalised Arabic alphabet	Domain features, sentiment (positive/negative) features	K-NN	90% when $K = 1$
Abdulla et al. (2014)	MSA, Egyptian	Arabic reviews from Yahoo Maktoob website	Tokenisation, stop words removal, weighting techniques, and stemming	Words, term frequency-inverse document frequency (TF-IDF) and feature reduction	SVM and NB	64.1% accuracy with the SVM classifier and 55.9% with the NB classifier
Shoukry and Rafea (2012)	MSA and dialects	Twitter	Removed user-names, pictures, hashtags, URLs, and non-Arabic words	Unigrams and bigrams	SVM and NB	SVM accuracy was 72% for unigrams
Abdul-Mageed et al. (2014)	MSA and dialects	Twitter, chat, forum	Tokenisation	Morphological features, standard features, dialectal Arabic features, and genre-specific features	SVM	85% for the Dardasha dataset using lemmas with extended reduced tag set (ERTS)

online lexicons such as SentiWordNet and Wordnet (Mahyoub et al. 2014). For Arabic sentiment analysis, there are not many resources and lexicons available for annotating the data. Thus, most of the previous research labelled data manually. Lastly, datasets from social media can sometimes be labelled by emoticons (Go et al. 2009a).

There are many machine learning methods that have been used to conduct Arabic sentiment analysis. However, three methods have been consistently showing superior performances. They are support vector machine (SVM), *K*-nearest neighbour (K-NN), and naive Bayes. However, the relevant literature is evident of experiments that use more than one of these machine learning methods. In Table 1, we present a summary of the different studies that have used machine learning method to analyse sentiment.

In order to classify the Arabic reviews into negative or positive classes, Duwairi and El-Orfali (2014) experimented with three classifiers SVM, K-NN, and naive Bayes. Specifically, they investigated seven different models of preprocessing tasks to assess their effect on accuracy. The seven models were:

1. baseline vectors without any processing;
2. stemmed vectors which are stemming words before calculating term frequencies also known as term frequency-inverse document frequency or TF-IDF;
3. vectors including remaining words after applying feature reduction;
4. vectors from word n-grams;
5. vectors from character n-grams;
6. vectors from word n-grams after applying feature correlation reduction; and
7. vectors from character n-grams with feature correlation.

The authors used two datasets: Politics dataset from Aljazeera website (300 reviews: 164 positive reviews and 136 negative reviews) and a publicly available Movie dataset (500 reviews: 250 positive reviews and 250 negative reviews). Their findings show that preprocessing enhances the classifier accuracy. Indeed, most preprocessing techniques overcame the base case. Only when stemming and light stemming was used for the Movie dataset, the performance was less than the base case.

The use of n-grams in levels, words and characters improved the performance of the three classifiers. The naive Bayes classifier reached 96.6% for the Movie dataset when correlated features were used. Character n-grams improved both the SVM and K-NN classifiers accuracies in the Movie dataset which resulted in 89% accuracy. Word n-grams increased the accuracy of K-NN which was 90% for the Movie dataset. As for the Politics dataset, naive Bayes performed the best giving 86% accuracy. Duwairi and El-Orfali (2014) research suggests that exploring features and classifiers is important in sentiment analysis in Arabic.

While some research focused on analysing sentiments in Modern Standard Arabic (MSA), other research aimed at colloquial and dialectical Arabic; the form more likely used in social networks. To this end, Al-Kabi et al. (2013)

developed a specialised sentiment analysis tool for colloquial Arabic and MSA to evaluate Arabic sentiments in social networks. They collected 1080 Arabic reviews from social media and news sites. The reviews contained Egyptian, Iraqi, Jordanian, Lebanese, Saudi, and Syrian dialects. For preprocessing, they removed the transliterated Arabic words like momtaz (i.e. meaning excellent in English) and Arabizi (Arabic chat alphabet). They also removed punctuation and non-alphabet characters and normalised some of the Arabic alphabet. K-NN was used as a classifier, and the evaluation was on a sentence (review)-based level. As a result, the accuracy of polarity detection was around 90% when $K = 1$.

On the other hand, Abdulla et al. (2014) applied sentiment analysis on Arabic reviews and comments from Yahoo Maktoob website. They collected 6921 instances which were manually classified into four topics: arts, politics, science and technology, and social. The data were labelled manually into four labels: positive, negative, neutral, undetermined, and spam. The polarity could not be determined for 4.2% of these reviews. They found the highest domain for ambiguous reviews was the technology domain and the lowest was the political domain. Different Arabic dialogues were used in the reviews: MSA, Egyptian, Gulf, Levantine, and Arabizi, and most of the reviews were written in MSA. Different preprocessing tasks were performed, including tokenisation, stop words removal, weighting techniques, and stemming. Two experiments were performed: one was on unbalanced data and the second was on balanced data. The classifiers chosen for this experiment were SVM and NB. The first experiment was on unbalanced with 3406, 1642, and 1324 for negative, neutral, and positive classes, respectively. They achieved a 64.1% accuracy with the SVM classifier and 55.9% with the NB classifier. In the second experiment, they chose 1000 instances from each class. Surprisingly, it resulted in a decrease in performance for both the SVM and NB classifiers.

Sentence-level Arabic sentiment analysis has also been examined. Indeed, Shoukry and Rafea (2012) performed sentiment analysis on 4000 tweets from different domains. They found that 1000 of these tweets were relevant and held opinion without sarcasm. They used two human annotators and found that 500 of the reviews were positive and 500 were negative. They preprocessed the text by removing user-names, pictures, hashtags, URLs, and non-Arabic words. The authors used unigrams and bigrams as features. Pertaining to the machine learning method, SVM and NB were chosen. They performed two experiments: one that included stop words and another that did not. The authors found that removing the stop words led to very small improvement in the performance, indicating that removing stop words add little value to the sentiment in text. SVM performed better than NB by around 4–6% accuracy, with a rate of 72% for unigrams. As for the features, using bigrams did not enhance the results of the unigram model.

Subjectivity in sentiment analysis was recently explored. Abdul-Mageed et al. (2014) created a supervised system for specialised for Arabic social media (SAMAR). Four feature sets were explored: morphological features, standard features, dialectal Arabic features, and genre-specific features. The authors used SVM as a classifier. Their results showed that adding morphological information either in form of lexemes or lemmas and adding POS tags has a positive effect on subjectivity and sentiment analysis (SSA). In general, adding standard features improved the performance of subjectivity and sentiment classification. The dialectal Arabic features decreased the performance of the classification models. As for the genre-specific features, it improved the subjectivity classification and the user was useful for sentiment classification. The best accuracy they achieved was 85% using lemmas with extended reduced tag set (ERTS).

5.5.1 Deep learning models

Deep learning is a branch of machine learning which aims to model high-level abstractions in data. This is done using model architectures that have complex structures or those composed of multiple nonlinear transformations (Deng and Yu 2014). According to Singhal and Bhattacharyya (2016), deep learning, when given enough data and training time, allows sentiment analysis to analyse data with little restrictions to the specificities of the task or data at hand. Deep learning can save time because human intervention and feature engineering is not needed (Ain et al. 2017). Ain et al. (2017) found in their survey comparing different Arabic sentiment analysis studies that deep learning networks are better than SVM and normal neural networks due to the multiple layers that they have. They also stated that deep learning networks have the capability to provide training in both supervised and unsupervised ways. Only a few researchers have explored deep learning models on Arabic text (see Table 2).

Al Sallab et al. (2015) explored several deep learning models:

- Deep neural network (DNN): DNN applies the back propagation to a conventional neural network with several layers;
- Deep belief networks (DBN): DBN pretrains phases before feeding it into other steps;
- Deep autoencoder (DAE): DAE reduces dimensionality to original models;
- Combined DAE with DBN; and

Table 2 Deep learning studies

Paper	Year	Language	Methods	Results
Al Sallab et al. (2015)	2015	MSA	DNN, DBN, deep autoencoder (DAE), and combined DAE with DBN, RAE	RAE model accuracy of 74.3%
Alayba et al. (2017)	2017	MSA and dialectical collected from Twitter	Naive Bayes, support vector machine, logistic regression, deep and convolutional neural networks	Convolutional neural networks (CNNs) accuracy was 90%

- Recursive autoencoder (RAE): The RAE parses raw sentence words in the best order which then minimises the error of creating the same sentence words in the same order.

Their results show that the DAE model gives better representation of the input sparse vector. The best model was the RAE leading to an accuracy of 74%. Moreover, the RAE model's performance was better that other models by around 9%.

Alayba et al. (2017) applied sentiment analysis on a health dataset using machine learning algorithms (naive Bayes, support vector machine, and logistic regression) alongside deep and convolutional neural networks. The deep neural network accuracy reached to 85%. Convolutional neural networks (CNNs) accuracy was slightly better reaching 90%. The best classifiers found were SVM using linear support vector classification and stochastic gradient descent.

To summarise, deep learning models can be beneficial in Arabic text due to the language's complexity. Although there have been a lot of studies that have applied deep learning models to sentiment text, there do not exist many studies on Arabic (Ain et al. 2017). There needs to be further research to explore these models benefits on Arabic text.

5.6 Lexicon-based method

The lexicon-based method is usually implemented when the data are unlabelled. Lexicons are sentiment dictionaries with the word and its occurring sentiment or sentiment score. Lexicons are used to label the data and to predict the polarity. In the Arabic language, only a few lexicons are available. Some researchers have created lexicons. However, most of these lexicons are not publicly available. There exist lexicons for MSA text, some of which is obtained using the translation of English lexicons (Mohammad et al. 2015). However, MSA lexicons are limiting as most social media users write informally. Therefore, it is important to research different Arabic dialects which have been very limited except for the Egyptian dialect [see Ibrahim et al. (2015), El-Makky et al. (2015), Al-Sabbagh and Girju (2012)]. In our previous research, we have created a new method to combine lexicons [see Altrabsheh et al. (2017)]. In Table 3, we present some examples of the Arabic lexicons available.

One example of research related to the creation of lexicons is Elhawary and Elfeky (2010). The authors created a lexicon consisting of 600 positive words/phrases and more than 900 negative words/phrases and 100 neutral words/phrases. The words and phrases are the most frequently Arabic words/phrases used over the web. In the next subsection, we present more examples of lexicons.

The lexicon-based method is usually known as a weak method in comparison with machine learning method (Ortigosa et al. 2014). To this end, only a few research articles used the lexicon-based method alone to analyse sentiment in the Arabic language. Indeed, most have adopted the lexicon-based approach along with machine learning—that is, the hybrid approach. This approach will be discussed in the next subsection.

Pertaining to studies that solely used the lexicon-based approach, formal and informal Arabic texts that were taken from online reviews and news articles have been examined. Al-Subaihin et al. (2011) proposed the design of a lexicon-based approach to conduct sentiment analysis on informal Arabic text. They used human-based computing to help build a lexicon. Their system consists of two different parts: the first part is a free online computer game that aimed to collect annotations of reviews from online players. The game has two players who are asked to highlight all the words or phrases that have positive and negative meaning. The aim of the game was to build a lexicon with positive and negative words. The second part was the sentiment analyser which classified reviews according to their sentiments using sequence patterns and lexicons created from previous games. They tagged words to POS, NEG, ENT, or NO if it is positive, negative, entity, or a negation, respectively. Moreover, they calculated the overall review polarity according to the score of the negative and positive sentences.

Al-Ayyoub et al. (2015) created a lexicon with 120,000 Arabic terms. They extended on a lexicon created by Abuaiadh (2011). They collected a large number of articles from Arabic news website. They then extracted distinct Arabic stems from them and translated the collected stems.

Table 3 Summary of Lexicons

Lexicon	Author	Dialect/ MSA	Source	Annotation	Size
AraSentiLexicon	Al-Twairesh et al. (2014)	Dialect	Tweets	Positive and negative emoticons and keywords	225,329
Arabic Emoticon Lexicon	Mohammad et al. (2015)	Dialect	Tweets	A set of 23 emoticons such as :) and :(	43,308
Arabic Hashtag Lexicon	Mohammad et al. (2015)	Dialect	Tweets	A set of 230 Arabic words that were manually selected for being highly positive or highly negative	22,006
Arabic Hashtag Lexicon (dialectal)	Mohammad et al. (2015)	Dialect	Tweets	A set of 483 dialectal Arabic words compiled by Refaee and Rieser (2014) from tweets	20,127
ArabicSentimentLexicon	Mahyoub et al. (2014)	MSA	Arabic WordNet	Positive and negative seed lists, the Arabic WordNet database, and a special sentiment orientation flags	15,110
Harvard Lexicon	Stone et al. (1966)	MSA	Harvard	Positive and negative keywords	1662
Colloquial Arabic Tweets	El-Makky et al. (2015)	Dialect	Tweets	Manually annotate	8867
Slang Lexicon	ElSahar and El-Beltagy (2014)	Dialect	Tweets	Manually annotate	378
Arabic translation of Bing Liu's Lexicon	Mohammed et al.	MSA	Bing Liu	Google translate	6789
Arabic translation of MPQA Subjectivity Lexicon	Mohammad et al. (2015)	MSA	MPQA	Google translate	7619
Arabic translation of NRC Emoticon Lexicon	Mohammad et al. (2015)	MSA	NRC Emoticon Lexicon	Google translate	26,740
Arabic translation of NRC Emotion Lexicon	Mohammad et al. (2015)	MSA	NRC Emotion Lexicon	Google translate	141,904
Arabic translation of NRC Hashtag Sentiment Lexicon	Mohammad et al. (2015)	MSA	NRC Hashtag Sentiment Lexicon	Google translate	32,582
Unweighted Opinion lexicon	El-Beltagy and Ali (2013)	Dialect	Tweets	Manually annotate	4391

The stems are searched for in lexicons from websites containing sentiment for the English language. The testing data included 300 tweets for each positive, negative, and neutral class. They preprocessed text by removing repetition of vowels, fixing spelling mistakes, and fixing mistakes caused by sound similarities. For this study, they did two experiments. The first experiment was lexicon based using the lexicon they created. The second experiment was a keyword-based approach where the keywords are simply the most frequent words in the tweet. They found that the lexicon-based approach led to the highest performance with an accuracy of 87%.

5.7 Hybrid techniques

The hybrid or combination approach uses both lexicon- and machine learning-based methods. This approach is more dominant in the relevant literature and is usually known to have a higher performance that lexicon-based method and machine learning method alone. The lexicon scores are typically used as features to input in the classifier. Pertaining to lexicons, the research explored word-level and sentence-level modern standard Arabic (MSA), dialectical or informal Arabic (DA), and MSA and DA combined. As for the machine learning classifiers, support vector machines and naive Bayes were the dominating methods. Yet, other approaches like K-nearest neighbour and entropy were also used. Table 4 illustrates a summary of previous research using the hybrid approach.

Some research aimed at determining the most appropriate approach, Lexicon based or hybrid, to analyse Arabic sentiment [e.g. El-Halees (2012)]. Specifically, El-Halees (2012) contrasted the accuracy rates of three methods: The lexicon-based only, combined lexicon-based and machine learning-based maximum entropy, and combined lexicon-based and machine learning-based K-nearest neighbour. Firstly, lexicon-based method was used to classify the documents. For the lexicon, they used

Table 4 Summary of hybrid-based sentiment analysis

Paper	Language	Dataset source	Preprocessing	Features	Classifier	Results (accuracy)
El-Halees (2012)	Domains: education, politics, and sports	MSA	Removed HTML tags, repeated letters and stop words, normalising Arabic letters, tokenisation, and Arabic light stemmer	Word	SVM, k-nearest method and maximum entropy and lexicon-based method	Lexicon-based method in combination with the maximum and the k-nearest method was the highest leading to an accuracy of 80%
Soliman et al. (2014)	Aljazera, BBCarabic, Alyoum Alsabe, Alarabia, Constitution Facebook Page, and People's Opinion Facebook page	MSA and dialects	Removed stop words, stemming and data autocorrection	Word	SVM	Second experiment which led to an accuracy of 87%
Abbasi et al. (2008)	Movie reviews and posts in hate/extremist group forums	MSA	None	Stylistic and syntactic features	SVM	Using EWGA for feature selection in conjunction with stylistic and syntactic features led to an accuracy of 95%
El-Makky et al. (2015)	Twitter	Dialects	They removed punctuation, numbers, special characters, and repetitions and replaced some characters such as Alef	Normalised word feature, stem level features, Tweet-specific features, language-independent features, semantic orientation feature	SVM	They achieved a accuracy of 84% with the hybrid approach
Duwairi (2015)	Twitter	Dialects	Tokenisation, removed stop words (except negation), and converting emoticons to their corresponding words	Words	SVM and NB	SVM F-score was 87% with dialect lexicon and 84% without. NB Fscore was 88% with dialect lexicon and 84% without it
Ibrahim et al. (2015)	Twitter and microblogs	MSA and dialects	They replace known idioms and proverbs with text masks	Standard features, sentence-level features, linguistic features, and syntactic features for conflicting phrases	SVM	The SVM accuracy for all the datasets combined before expansion led to a 94% accuracy and after expansion 95%

SentiStrength software and translated words from English to Arabic and an online dictionary. Then, the classified documents were used as a training set for maximum entropy method which classified other documents. Finally, the k-nearest method used classified documents from both the lexicon-based method and maximum entropy as a training set to classify the rest of the documents. They collected documents related to opinions in Arabic from three domains: education, politics, and sports. The total amount of data was 635 positive files and 508 negative files. This included 4375 positive statements and 4118 negative statements. For preprocessing, they used several techniques which included removing HTML tags, repeated letters and stop words, normalising Arabic letters, tokenisation, and Arabic light stemmer. They obtained the vectors using TF-IDF (term frequency-inverse document frequency). Consequently, their results showed that the accuracy of the lexicon-based alone method was 50% and the lexicon-based method in combination with the maximum entropy was 61%. The lexicon-based method in combination with the maximum and the k-nearest method led to the highest accuracy at 80%.

Another attempt to examine the most adequate hybrid approach to Arabic sentiment analysis combined lexicon-

based with machine learning-based SVM approaches. More specifically, Soliman et al. (2014) conducted experiments with a Slang Sentimental Words and Idioms Lexicon (SSWIL) that they constructed. The data used were Arabic slang comments from different sites like: Aljazeera, BBCarabic, Alyoum Alsabe, Alarabia, Constitution Facebook Page, and People's Opinion Facebook page. The authors first preprocessed by removing stop words, stemming, and data autocorrection. Subsequently, they conducted three experiments. The first experiment classified comments using the SVM machine learning-based approach only. Alternatively, the second experiment used the lexicon-based approach only with the lexicon they constructed SSWIL in order to classify the comments. The third experiment combined the two approaches, that is, classifying comments using the SSWIL lexicon and then applying the SVM machine learning-based approach. The highest classification performance out of the three experiments was the third experiment, the combined approach, which led to an accuracy of 87%.

Attempts to combine multiple approaches to analyse sentiments and for multiple languages have also been made. Indeed, Abbasi et al. (2008) investigated sentiment analysis on web forum opinions in multiple languages Arabic and English. Two machine learning-based approaches combined with rich feature extractions were examined separately. The machine learning approaches were the entropy weighted genetic algorithm (EWGA) and SVM with fivefold cross-validation and bootstrapping methods to randomly select the data. The former approach incorporated the information gain (IG) heuristic with a genetic algorithm (GA) to improve feature selection performance. Pertaining to the features explored, stylistic and syntactic features were used. The syntactic features included POS tags, n-grams (for English), word roots (for Arabic), word n-grams, and punctuation. On the other hand, stylistic features included word- and character-level lexical features, word length distributions, special characters, letters, character n-grams, structural features, vocabulary richness measures, digit n-grams, and function words. To extract features, they determine the word roots by comparing them against a root dictionary. Abbasi et al. (2008) also used feature reduction and selection to improve classification and accuracy of the models and provide a greater insight into important class attributes. As a dataset, the authors investigated movie reviews and posts in hate/extremist group forums. Specifically, two datasets were used: (1) A movie review dataset consisting of 2000 reviews (1000 positive and 1000 negative) taken from the IMDb movie review archives and (2) messages from two major extremist forums (one USA and one Middle Eastern) collected as part of the Dark Web project containing 1000 instances each (500 positive and 500 negative). The results show that the performance is achieved when using stylistic and syntactic features together combined with EWGA machine-learning approach. Indeed, this hybrid approach led to an accuracy of 95%.

Sentiment analysis research also explores dialectical Arabic text. Some research merged multiple lexicons, modern standard Arabic and dialectical, with a machine learning-based approach (El-Makky et al. 2015). More specifically, El-Makky et al. (2015) built a new Arabic lexicon by merging two MSA lexicons and two Egyptian Arabic lexicons. They used SVM as a classifier and adapted one of the state-of-the-art Semantic Orientation (SO) algorithms. The authors also applied feature selection using the information gain measure. Three datasets from Arabic Egyptian tweets were used. For preprocessing, they removed punctuation, numbers, special characters, and repetitions and replaced some characters such by Alef. Their results show a significant increase in the performance of the models and an accuracy rate of 84% was achieved with the hybrid approach.

Other research experimented with dialectical Arabic words and modern standard Arabic (MSA) words [see Duwairi (2015)]. Duwairi (2015) converted the former to the latter to find out which of the approach is more useful. The classifiers they used were naive Bayes and support vector machine. To convert dialectical words to MSA, the authors used a dialect lexicon that contains dialectical words with their corresponding MSA words. The text was preprocessed using tokenisation, removing stop words (except negation), and converting emoticons to their corresponding words. Twitter API was used to collect 22550 dialectical tweets (positive: 8529, negative: 7021, and neutral: 7000). The text was annotated using the crowdsourcing tool. Their findings reveal that replacing dialectical words with the MSA words improves the overall performance. Precision in the positive class was improved when the dialect lexicon was used with both classifiers. Additionally, precision in the negative class was improved when the dialect lexicon was used with the NB classifier. The dialect lexicon did not have any effect on the neutral class. Moreover, the recall of the negative class was noticeably improved with the dialect lexicon across both classifiers. SVM F-score was 87% with dialect lexicon and 84% without. NB F-score was 88% with dialect lexicon and 84% without it.

Lastly, some research on dialectical Arabic proposed a feature-based sentence-level approach for Arabic sentiment analysis [see Ibrahim et al. (2015)]. Ibrahim et al. (2015) created a corpus containing 2000 Arabic sentiment statements including 1000 MSA and dialect tweets and 1000 microblogs (hotel reservation comments, product reviews, TV programme and movie comments). The data consisted of both Modern Standard Arabic and Egyptian dialectal

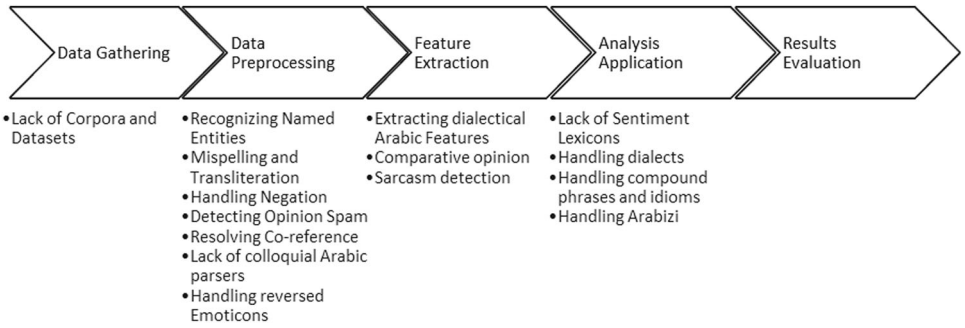

Fig. 3 Limitations of Arabic sentiment analysis

Arabic. The authors built two lexicons: Arabic sentiment words lexicon and Arabic sentiment idioms phrase lexicon and replaced known idioms and proverbs with text masks. One example to this is the idiom 'crocodile tears', which is usually negative (NG) and therefore is replaced by (NG Phrase). Their polarity lexicon, which was manually collected and annotated, consisted of 5244 words (2003 positive (PO), 2829 negative (NG), 412 neutral (NU)). To predict the sentiment of a word they searched for the synonym of the words in a previous lexicon. Different features were used in this research: standard features, sentence-level features, linguistic features, and syntactic features for conflicting phrases. Support vector machine (SVM) was used as a classifier. The data were divided into 80% for training and 10% for developing and 10% for testing. The authors expanded their lexicon with an expansion technique, which adds new sentiment words to the polarity lexicon automatically. Their results show that the DA lexicon expansion increased the performance 1–4%. The SVM accuracy for all the datasets combined before expansion led to a 94% accuracy and after expansion 95%.

6 Challenges in sentiment analysis of Arabic text

In the previous chapter, we reviewed and presented the three approaches for Arabic sentiment analysis that were found in the relevant literature. From our analysis, we identified the challenges and limitations that researchers highlight. While there are many challenges incurred when conducting sentiment analysis research, we limit this section to those challenges that are specific to the Arabic language. Accordingly, we organise those challenges under the corresponding sentiment analysis phase (see Fig. 3).

6.1 Data gathering

Data gathering is challenging for the Arabic language, due to the limited sources available to collect it (Al-Twairesh et al. 2014). Data are usually collected from social media websites or online forums (Al-Sabbagh and Girju 2012). Additionally, the data collected could be from different dialects or in Arabizi, meaning that the amount of data that will be used to train sentiment analysis models will be small, which could lead to inaccurate results or lower accuracies.

6.1.1 Lack of corpora and datasets

The Arabic language is the fifth languages spoken worldwide (Duwairi 2015); however, there is a lack of Arabic reviews and web content. Compared to the English language, there are not many datasets available to apply sentiment analysis. This makes performance comparison between languages difficult, as sentiment analysis accuracy depends on the amount of data. Refaee and Rieser (2014) presented a corpus of annotated tweets to support sentiment analysis of Arabic Twitter feeds. They manually annotate a random subset of 8868 examples of the collected tweets with polarity and mark them with neutral, mixed, positive, and negative. When annotators were not sure what to label the tweet, they labelled it with other/uncertain. Two Arabic native speakers were recruited for annotation. The annotators labelled the data stating the reason when neutral or uncertain. Some of the tweets contained sarcasm which was challenging for even human annotators.

Al-Sabbagh and Girju (2012) built YADAC, a multigenre dialectal Arabic (DA) corpus from microblogs (i.e. Twitter), blogs/forums, and online knowledge market services. This research focused on the Egyptian dialect only. Base phrase chunking was done and the following features were incorporated: semantic features: gender, number definiteness agreement between subject and verbs, nouns, and adjectives; morphological features: subject and object clitics; syntactic feature: transitive vs. intransitive verbs; lexical features: using function words (i.e. prepositions, conjunctions, interjections, and relative pronouns) as anchors; metalinguistic features: punctuation markers.

Zaidan and Callison-Burch (2011) built a multi-dialectal Arabic corpus with crowdsourcing. The data contained 1.4 M comments, corresponding to 52.1 M words. The data

were annotated to dialect and MSA labels by 455 annotators.

6.2 Data preprocessing

6.2.1 Recognising named entities

The detection of Arabic names in a sentence is important because the classifier may confuse Arabic names which are derived from Arabic adjectives (e.g. 'جميلة') for sentiments (Al-Twairesh et al. 2014). Unlike English where the name is usually cued with a capital letter, there is no method available to detect Arabic names. It is also important when wanting to find the opinion holder.

6.2.2 Misspelling and transliteration in microblogs

Microblogs such as Twitter often contain misspelled words and transliterated ones such as the word 'فالنتاين', which refers to the word 'valentine' spelt in Arabic letters (Refaee and Rieser 2014). Additionally, name mentions and hashtags are used in Twitter, which adds noise to the dataset.

6.2.3 Handling negation

Negation in dialects is different to MSA as it can be expressed in many ways. The negating words can have other meanings making detecting negation hard and error prone (Darwish et al. 2012), One example is the word 'مش'- 'msh' meaning 'not'.

6.2.4 Detecting opinion spam

Opinion spam is when an opinion is untruthful or fake with the aim of misguiding the reader. This is usually done in businesses to promote or demote and product. Little research has been conducted in Arabic opinion spam (Hammad 2013; Wahsheh et al. 2013).

6.2.5 Resolving co-references

Co-reference resolution is a challenging problem in sentiment analysis (Al-Twairesh et al. 2014). It is unclear what the sentiment is referring to if there exist multiple opinions. For example, the sentence: "عجبتني الصور ما عدا المناظر بس التصوير مو واضح"which means I liked the photos apart from the views, but the photographs are not clear. The sentence here has a mixture of positive and negative opinions. It is dialectical and has no sentence structure; therefore, it is difficult to know what the polarity is referring to.

6.2.6 Lack of colloquial Arabic parsers

Parsing Arabic text is difficult due to its morphological complexity with high inflectional and derivational nature (El-Makky et al. 2015). In MSA, there are publicly available parsers such as Stanford Arabic Parser. On the other hand, for colloquial Arabic, there is no standardisation, and it differs from MSA phonologically, morphologically, and lexically, which makes it a complex task to build morphological analysers and POS taggers (El-Makky et al. 2015).

6.2.7 Handling reverse emoticons

Analysing Emoticons is challenging in the Arabic text due to the nature of Arabic being from right to left. Therefore, the emoticons are often mistakenly interchanged, leading to contradictory sentiments within the tweet. This makes labelling unsupervised data using the emoticons challenging such as in the research conducted by Go et al. (2009a). Altrabsheh et al. (2015) found that sometimes people use sarcasm in their tweets along with opposite emoticon. Also, using the wrong emoticons could change the whole meaning of the sentence.

6.3 Feature extraction

6.3.1 Extracting dialectal Arabic features

Informal (dialectal) Arabic is challenging to analyse as it is generally known to be non-structured and difficult to standardise (Al-Ayyoub et al. 2015; Al-Kabi et al. 2013; Refaee and Rieser 2014). It is different from MSA phonologically, morphologically, and syntactically, which make morphological analysers and POS taggers very challenging (El-Makky et al. 2015; Al-Twairesh et al. 2014). Negation and stop words can differ in dialects from MSA. Additionally, concepts do not have the same lexical choices in different DAs, which makes building lexicons for multiple dialects very challenging (Al-Twairesh et al. 2014). Some users mix between DA and MSA in one text like when using poetry and Qur'anic verses.

6.3.2 Sarcasm detection

Sarcasm is a general issue in sentiment analysis due to its effect in misleading the classification. Detecting sarcasm is not an easy process and limited research has been conducted on this issue. Sarcasm has not been explored in the Arabic language (Al-Twairesh et al. 2014).

6.4 Analysis application

6.4.1 Lack of sentiment lexicons

Although many researchers have created sentiment lexicons for the Arabic language, it has not been made publicly available (Al-Twairesh et al. 2014). SANA, a large-scale multi-genre multi-dialect Arabic sentiment lexicon, was built; however, it only covered two dialects (i.e. Egyptian and Levantine) and has not been applied to SSA tasks (Abdul-Mageed et al. 2014).

6.4.2 Handling compound phrases and idioms

Compound phrases and idioms are commonly used in Arabic text in social media. These phrases could vary from one dialect to another and may have a reversed meaning in different regions of the Arab world. Phrases and words are subject to usage trends, with new phrases evolving every day (Al-Twairesh et al. 2014). One example is the phrase "افتكرناه موسى طلع فرعون", which means 'we thought he was Moussa but he was the Pharaoh' reflecting negative polarity (Ibrahim et al. 2015). Ibrahim et al. (2015) proposed a method to address idioms and phrases by finding their polarity using polarised synonyms from MSA lexicons. However, there are many idioms available in the Arabic language and it is impossible to cover the polarity for all of them.

6.4.3 Use of Arabizi

Arabizi is a new trend in social media where the person uses Latin characters to represent Arabic words. Additionally, some Arabic users tend to switch between Arabic and English, making it difficult to detect if a word is written is Arabizi or English (Al-Twairesh et al. 2014). The issue of Arabizi has not dealt with in the literature yet.

6.5 Results of Arabic sentiment analysis models challenges

From our literature review, we did not find any challenges in the results of Arabic sentiment analysis found yet. The results are dependent on the classification, and the general challenge is to achieve results above 70% accuracy. To the most of the Arabic sentiment analysis, research is known to achieve such a rate.

7 Discussion

Arabic sentiment analysis research has increased noticeably in the last decade. In this paper, we covered the Arabic sentiment analysis literature, highlighting challenges. We also outlined the differences between Arabic and English sentiment analysis.

After examining the sentiment analysis approaches used in Arabic, we found several observations. Firstly, we found that the lexicon method alone may not be the best method to analyse Arabic sentiment, due to the numerous words from different dialects and the reality of fitting all of the words in a lexicon. Moreover, most of the research in Arabic has been using machine learning techniques or the combination method. The literature review suggested that there is no best approach for sentiment analysis; however, the hybrid approach of sentiment analysis gave a higher performance than other approaches in many studies. Adel Assiri and Aldossari (2015) presented an Arabic survey review on several studies and found that researchers in Arabic sentiment analysis should use more diverse techniques and approaches. They also suggested that using deep learning techniques could be beneficial in analysing Arabic sentiment. Ain et al. (2017) also presented a general review on deep learning techniques. They found that deep learning can be beneficial for analysing sentiment more than normal methods. As opposed to our review, their review was not specific to Arabic language. Indeed, we found only two researches had evaluated deep learning in Arabic text and only one of the studies had compared it with basic machine learning (Al Sallab et al. 2015; Alayba et al. 2017). Alayba et al. (2017) found that the SVM performed 1% better than deep learning methods. Hence, there is need of exploring deep learning performance further on Arabic sentiment and to compare its performance with basic machine learning techniques.

For the hybrid approach, lexicons are needed as part of the analysis. These could be general or domain specific. Most of the research conducted in Arabic sentiment analysis has been in general without specifying the domain. Sentiment analysis is domain dependent and words could differ in sentiment from one domain to another. Although this approach has been found give a good performance, it is more accurate to perform domain based sentiment analysis. The hybrid method could be beneficial in Arabic sentiment analysis due to the multiple dialects. Some words may infer polarity in some dialect and not exist in another one. By creating a large lexicon that covers all the dialects, we could increase the performance of the classification. As for the machine learning techniques, most of the researchers in Arabic sentiment analysis use SVM and NB as classifiers. Other classifiers such as ME and K-NN have not been

explored much. These classifiers may lead to a better performance than SVM and NB and hence would be useful to explore in future research.

Domain-dependent sentiment analysis requires larger amounts of data. However, one main challenge in Arabic sentiment analysis is the lack of resources and corpora. This could be due to the lack of webpages available for users to provide Arabic reviews. The amount of Arabic reviews and datasets are considerably small and limited in comparison with the number of Arabic users online. The lack of Arabic corpora makes it more challenging to compare between English and Arabic sentiment analysis results, due to the large number of corpora in English language. It will be very advantageousness if researchers could share lexicons and knowledge to work towards building more accurate sentiment analysis models and analysis.

It is still unclear whether preprocessing is beneficial for Arabic sentiment analysis as there is no study which compares the preprocessing individual performance apart from Duwairi and El-Orfali (2014). We found most of studies applied the preprocessing step. The most common preprocessing step was stemming which could be due to the many variations of Arabic prefixes and suffixes. In general, the Arabic language is complex and parsing it is difficult due to its morphological complexity with high inflectional and derivational nature (El-Makky et al. 2015). There also exist many compound phrases and idioms which reflect a certain polarity. Detecting these idioms will lead to a better sentiment analysis performance. Moreover, due to the non-existence of name cues, it is difficult to recognise names and the classifier may confuse Arabic names which are derived from adjectives with sentiment. Another issue related to DA text is negation. Negation can be expressed in many ways in DA text, and negating words can have various meanings making detecting negation challenging and error prone.

Different features used in Arabic sentiment analysis were discussed. While many have used n-grams and POS tagging, some have used linguistic features which are useful in Arabic text due to the complexity. Some argue that n-grams and POS tagging led to the best performance, but recent research shows by including other features, the performance of the analysis is increased. We found that syntactic and semantic relations could be useful to the analysis of Arabic text. Studying linguistic features could be beneficial and could increase the performance of polarity detection. The best suitable feature depends on many variables such as different dialects and text contents. This can be explored further in future work.

We summarised the common Arabic sentiment analysis challenges, many of which have not been addressed. This opens opportunity for future researchers to overcome and find solutions for these challenges and to make sentiment analysis in Arabic more accurate. The challenges in Arabic sentiment analysis are considerably more than other languages due to the complexity of the language. Additionally, due to the dialectics, their is no set of grammar rules and sentence structure. Also, each dialect needs to be understood on its own in regard to word meanings, phrases, and sentence structure. One of the challenges which has not been studied thoroughly in the existing literature is Arabizi. According to our knowledge, only one research studied Arabizi in sentiment analysis (Duwairi et al. 2016). Hence, the need to research this further.

8 Conclusion

This research explored sentiment analysis in the Arabic language. Different challenges in Arabic and applying sentiment analysis to Arabic were highlighted. A detailed review of previous literature related to sentiment analysis in Arabic was presented.

Some of the main highlights of the literature was that sentiment analysis is challenging for the Arabic language due to the complex language and multiple dialects. We identified the main challenges of Arabic sentiment analysis in each of the process parts. We found that the main challenge in the data collection part is a lack of corpora and lexicons. There is a need to create lexicons that can be used for all Arabic dialects.

For the preprocessing part, we identified several challenges including misspellings, negation and recognising named entities. As for the feature challenges, we found that extracting features from informal Arabic is challenging. Challenges in the analysis also include how accurate the model will predict the sentiment with different dialects and sarcasm. Arabizi is also another challenge and has not been explored in the literature before.

For future work, we will explore the creation of a multi-dialect Arabic lexicon. This can be used to mine tweets from different domains and trends in Twitter. We will also explore different linguistic features such as semantical, syntactical, and stylistic features which can be adapted in sentiment analysis models.

Acknowledgements This publication was made possible by the NPRP award [NPRP 7-1334-6-039 PR3] from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the author[s].

Appendix: Literature summarisation

See Table 5.

Table 5 Literature summarisation

Paper	Dataset source	Dataset size	Language	Approach	Preprocessing	Features	Classifier	Results (Accuracy)	Pros	Cons
Duwairi and El-Orfali (2014)	Politics dataset from Aljazeera website and Movie dataset which is publicly available	Politics dataset (300 reviews: 164 positive reviews and 136 negative reviews) and Movie dataset (500 reviews: 250 positive reviews and 250 negative reviews)	MSA	Supervised	Stemming and light stemming	N-grams in levels, words, and characters	SVM, K-NN, and NB	NB: 96.6% for the Movie dataset and 86% accuracy for the Politics dataset	Explored different features	No Neutral class
Al-Kabi et al. (2013)	Social media and news sites	1080 Arabic reviews	Colloquial Arabic and MSA	Supervised	Removed the transliterated Arabic words like momtaz (i.e. meaning excellent in English) and Arabizi (Arabic chat alphabet). They removed punctuation and non-alphabet characters, and normalised some of the Arabic alphabet	Domain features, sentiment (positive/ negative) features	K-NN	90% when $K = 1$	Different dialects	Small dataset, only one classifier
Abdulla et al. (2014)	Arabic reviews from Yahoo Maktoob website	3406, 1642, and 1324 for negative, neutral and positive classes	MSA, Egyptian	Supervised	Tokenisation, stop words removal, weighting techniques, and stemming	Words, term frequency-inverse document frequency (TF-IDF) and feature reduction	SVM and NB	64.1% accuracy with the SVM classifier and 55.9% with the NB classifier	Different dialects	Low results
Shoukry and Rafea (2012)	Twitter	4000 tweets from different domains and used 1000 (500 of the reviews were positive and 500 were negative)	MSA and dialects	Supervised	Removed user-names, pictures, hashtags, URLs and non-Arabic words	Unigrams and bigrams	SVM and NB	SVM accuracy was 72% for unigrams	Different dialects	Small dataset

Table 5 continued

Paper	Dataset source	Dataset size	Language	Approach	Preprocessing	Features	Classifier	Results (Accuracy)	Pros	Cons
Abdul-Mageed et al. (2014)	Twitter, Chat, Forum	14934 tweets and sentences(4366 positive, 4288 negative, 4966 Objective)	MSA and dialects	Supervised	Tokenisation	Morphological features, standard features, dialectal Arabic features, and genre-specific features	SVM	85% for the Dardasha dataset using lemmas with extended reduced tag set (ERTS)	Different features	80% training, 10% for development, and 10% for testing
Abdulla et al. (2014)	Twitter Yahoo!-Maktoob	Twitter 1000 positive 1000 negative, Yahoo!-Maktoob 1000 positive 1000 negative	MSA and dialects	Unsupervised	Tokenisation, removed repeated letters, normalisation, light stemming, removed stop words	Word	Lexicon-based	Twitter 70.05% Yahoo!-Maktoob 63.75%	Different dialects	Low results
Al-Ayyoub et al. (2015)	Twitter	300 positive, 300 negative, 300 neutral	MSA	Unsupervised	Removed repetition of vowels, fixing spelling mistakes, fixing mistakes caused by sound similarities	Word	Lexicon-based	Accuracy of 87%	Created largest lexicon	Small dataset
El-Halees et al. (2011)	Domains: education, politics and sports	4375 positive statements and 4118 negative statements	MSA	Hybrid	Removed HTML tags, repeated letters and stop words, normalising Arabic letters, tokenisation, and Arabic light stemmer	Word	SVM, k-nearest method and maximum entropy and lexicon-based method	Lexicon-based method in combination with the maximum and the k-nearest method was the highest leading to an accuracy of 80%	Different domains and a new technique using multiple machine learning techniques	This method may be time consuming

Table 5 continued

Paper	Dataset source	Dataset size	Language	Approach	Preprocessing	Features	Classifier	Results (Accuracy)	Pros	Cons
Soliman et al. (2014)	Aljazera, bbcarabic, Alyoum Alsabe, Alarabia, Constitution Facebook Page, and People's Opinion Facebook page	1355 random comments are taken	MSA and dialects	Hybrid	Removed stop words, stemming and data auto correction	word	SVM	second experiment which led to an accuracy of 87%	New Slang word list	Little data and classes not specified.
Abbasi et al. (2008)	Movie reviews and posts in hate/ extremist group forums	A movie review dataset consisting of 2000 reviews (1000 positive and 1000 negative) taken from the IMDb movie review archives and messages from two major extremist forums (one U.S. and one Middle Eastern) collected as part of the Dark Web project containing 1000 instances each (500 positive and 500 negative)	MSA	Hybrid	None	Stylistic and syntactic features	SVM	Using EWGA for feature selection in conjunction with stylistic and syntactic features led to an accuracy of 95%	Useful features	Small dataset
El-Makky et al. (2015)	Twitter	7800 tweets	Dialects	Hybrid	They removed punctuation, numbers, special characters, and repetitions and replaced some characters such as Alef	Normalised word feature, stem level features, Tweet-specific features, Language-independent features, semantic orientation feature	SVM	They achieved a accuracy of 84% with the hybrid approach	Large variety of features	They did not specify how much instances for each polarity

Table 5 continued

Paper	Dataset source	Dataset size	Language	Approach	Preprocessing	Features	Classifier	Results (Accuracy)	Pros	Cons
Duwairi (2015)	Twitter	22550 tweets (positive :8529, negative: 7021 and neutral: 7000)	Dialects	Hybrid	Tokenisation, removed stop words (except negation), and converting emoticons to their corresponding words	Words	SVM and NB	SVM F-score was 87% with dialect lexicon and 84% without. NB F-score was 88% with dialect lexicon and 84% without it	Large dataset	They did not report accuracy
Ibrahim et al. (2015)	Twitter and microblogs	1000 MSA and dialect tweets and 1000 microblogs	MSA and dialects	Hybrid	They replace known idioms and proverbs with text masks	Standard features, sentence-level features, linguistic features, and syntactic features for conflicting phrases	SVM	The SVM accuracy for all the datasets combined before expansion led to a 94% accuracy and after expansion 95%	Different features	They used 80% of the data for training and 10% for developing and 10% for testing. They used a variety of features

References

Abbasi A, Chen H, Salem A (2008) Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums. ACM Trans Inf Syst (TOIS) 26(3):12

Abdulla NA, Al-Ayyoub M, Al-Kabi MN (2014) Int J Big Data Intell 1(1–2):103

Abdulla NA, Ahmed NA, Shehab MA, Al-Ayyoub M, Al-Kabi MN, Al-rifai S (2014) Towards improving the lexicon-based approach for Arabic sentiment analysis. Int J Inf Technol Web Eng 9(3):55

Abdul-Mageed M, Diab M, Kübler S (2014) SAMAR: subjectivity and sentiment analysis for Arabic social media. Comput Speech Lang 28(1):20

Abuaiadh D (2011) Dataset for Arabic document classification. http://diab.edublogs.org/dataset-for-arabic-document-classification

Adel Assiri AE, Aldossari H (2015) Int J Adv Comput Sci Appl. https://doi.org/10.14569/ijacsa.2015.061211

Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) In: Proceedings of the workshop on languages in social media. Association for Computational Linguistics, Stroudsburg, pp 30–38

Ain QT, Ali M, Riaz A, Noureen A, Kamran M, Hayat B, Rehman A (2017) Sentiment analysis using deep learning techniques: a review. Int J Adv Comput Sci Appl 8(6):424

Al Sallab AA, Baly R, Badaro G, Hajj H, El Hajj W, Shaban KB (2015) In: ANLP workshop, vol 9

Alayba AM, Palade V, England M, Iqbal (2017) arXiv preprint arXiv:1702.03197

Al-Ayyoub M, Essa SB, Alsmadi I (2015) Lexicon-based sentiment analysis of Arabic tweets. Int J Soc Netw Min 2(2):101

Al-Kabi M, Gigieh A, Alsmadi I, Wahsheh H, Haidar M (2013) In: The fourth international conference on information and communication systems (ICICS 2013), pp 23–25

Al-Sabbagh R, Girju R (2012) In: Proceedings of the eight international conference on language resources and evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA), pp 2882–2889

Al-Subaihin AA, Al-Khalifa HS, Al-Salman AS (2011) In: Proceedings of the 13th international conference on information integration and web-based applications and services. ACM, pp 543–546

Altrabsheh N, Gaber M, Cocea M (2013) SA-E: sentiment analysis for education. Int Conf Intell Decis Technol 255:353

Altrabsheh N, Cocea M, Fallahkhair S (2014) In: Adaptive and intelligent systems. Springer, New York, pp 40–49

Altrabsheh N, Cocea M, Fallahkhair S (2015) Tenth European conference on technology enhanced learning

Altrabsheh N, El-Masri M, Mansour H (2017) In: American conference information systems

Al-Twairesh N, Al-Khalifa H, Al-Salman A (2014) In: 2014 IEEE/ACS 11th international conference on computer systems and applications (AICCSA), IEEE, pp 148–155

Arabic speaking internet users and population statistics (2015). http://www.internetworldstats.com/stats19.htm

Asur S, Huberman B et al. (2010) In: Web intelligence and intelligent agent technology, IEEE, vol. 1, pp 492–499

Barbosa L, Feng J (2010) Robust sentiment detection on Twitter from biased and noisy data. Int Conf Comput Linguist 23:36

Chamlertwat W, Bhattarakosol P, Rungkasiri T, Haruechaiyasak C (2012) Discovering consumer insight from Twitter via sentiment analysis. J Univers Comput Sci 18(8):973

Claster W, Cooper M, Sallis P (2010) In: Computational intelligence, modelling and simulation second international conference (CIMSiM), pp 89–94. https://doi.org/10.1109/CIMSiM.2010.98

Darwish K, Magdy W, Mourad A (2012) In: Proceedings of the 21st ACM international conference on information and knowledge management. ACM, pp 2427–2430

Deng L, Yu D et al (2014) Foundations and trends®. Sig Process 7(3–4):197

Duwairi RM, Alfaqeh M, Wardat M, Alrabadi A (2016) In: 7th International conference on information and communication systems (ICICS), IEEE, pp 127–132

Duwairi RM (2015) In: Information and communication systems (ICICS), 2015 6th international conference on IEEE, pp 166–170

Duwairi RM (2007) Arabic text categorization. Int Arab J Inf Technol 4(2):125

Duwairi R, El-Orfali M (2014) A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. J Inf Sci 40(4):501

Duwairi R, Marji R, Sha'ban N, Rushaidat S (2014) In: Information and communication systems (ICICS), 2014 5th international conference on IEEE, pp 1–6

El-Beltagy SR, Ali A (2013) In: Innovations in information technology (IIT), 2013 9th international conference on IEEE, pp 215–220

El-Halees A (2012) In: The 13th international Arab conference on information technology ACIT, pp 265–271

El-Halees A et al (2011) Arabic opinion mining using combined classification approach

Elhawary M, Elfeky M (2010) In 2010 IEEE international conference on data mining workshops, IEEE, pp 1108–1113

El-Makky N, Nagi K, El-Ebshihy A, Apady E, Hafez O, Mostafa S, Ibrahim S (2015) The 3rd ASE international conference on social informatics (SocialInformatics 2014), At Harvard University, Cambridge, MA, USA

ElSahar H, El-Beltagy SR (2014) In: CICLing, vol 1, pp 79–91

Fei Z, Liu J, Wu G (2004) In: Computer and information technology, 2004. CIT'04. The fourth international conference on IEEE, pp 1147–1152

Gamon M, Aue A, Corston-Oliver S, Ringger E (2005) Pulse: mining customer opinions from free text. Int Conf Adv Intell 6:121

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford. http://s3.eddieoz.com/docs/sentiment_analysis/Twitter_Sentiment_Classification_using_Distant_Supervision.pdf

Go A, Huang L, Bhayani R (2009) Twitter sentiment analysis. CS224N Project Report, Stanford. http://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf

Habash N, Rambow O, Roth R (2009) In: Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt, pp 102–109

Hammad ASA (2013) An approach for detecting spam in Arabic opinion reviews. Ph.D. thesis, Islamic University of Gaza

Hill S, Nalavade A, Benton A (2012) In Proceedings of the sixth international workshop on data mining for online advertising and internet economy. ACM, vol 12, pp 4–12

Ibrahim HS, Abdou SM, Gheith M (2015) arXiv preprint arXiv:1505.03105

Jansen BJ, Zhang M, Sobel K, Chowdury A (2009) Twitter power: Tweets as electronic word of mouth. J Am Soc Inf Sci Technol 60(11):2169

Korayem M, Crandall D, Abdul-Mageed M (2012) In: International conference on advanced machine learning technologies and applications. Springer, New York, pp 128–139

Kumar A, Sebastian TM (2012) Machine learning assisted sentiment analysis. Int Conf Comput Sci Eng 12(13):123

Lu B, Tsou BK (2010) In: machine learning and cybernetics (ICMLC), 2010 international conference on IEEE, vol 6, pp 3311–3316

Mahyoub FH, Siddiqui MA, Dahab MY (2014) Building an Arabic sentiment lexicon using semi-supervised learning. J King Saud Univ Comput Inf Sci 26(4):417
Mohamed E, Kübler S (2010) In: Language resources and evaluation
Mohammad SM, Salameh M, Kiritchenko S (2015) How translation alters sentiment. J Artif Intell Res 1:1
Mountassir A, Benbrahim H, Berrada I (2012) 3 e Séminaire de Veille Stratégique, Scientifique et Technologique (VSST'12)
Mouthami K, Devi K, Bhaskaran V (2013) Sentiment analysis and classification based on textual reviews. Inf Commun Embed Syst (ICICES) 1:271
Oraby S, El-Sonbaty Y, El-Nasr MA (2013) In: International joint conference on natural language processing, pp 471–479
Ortigosa A, Martin JM, Carro RM (2014) Sentiment analysis in Facebook and its application to e-learning. Comput Hum Behav 31:527
Pak A, Paroubek P (2010) Twitter based system: using Twitter for disambiguating sentiment ambiguous adjectives. Int Workshop Semant Eval 5:436
Pak A, Paroubek P (2011) In: 2011 22nd international workshop on database and expert systems applications, IEEE, pp 111–115
Pang B, Lee L, Vaithyanathan S (2002) ACL-02 conference on empirical methods in natural language processing, vol 10, pp 79
Prasad S (2010) Micro-blogging sentiment analysis using bayesian classification methods. CS224N Project Report, Stanford. http://nlp.stanford.edu/courses/cs224n/2010/reports/suhaasp.pdf
Refaee E, Rieser V (2014) In: Language resources and evaluation, pp 2268–2273
Refaee E, Rieser V (2014) In: Proceedings of the workshop on free/open-source Arabic corpora and corpora processing tools, p 16
Rushdi-Saleh M, Martín-Valdivia MT, Ure na-López LA, Perea-Ortega JM (2011) OCA: opinion corpus for Arabic. J Am Soc Inform Sci Technol 62(10):2045
Saif H, He Y, Alani H (2012) In: International semantic web conference. Springer, New York, pp 508–524
Shoukry A, Rafea A (2012) In: Collaboration technologies and systems (CTS), 2012 international conference on IEEE, pp 546–550
Singhal P, Bhattacharyya P (2016) Sentiment analysis and deep learning: a survey. http://www.cfilt.iitb.ac.in/resources/surveys/sentiment-deeplearning-2016-prerna.pdf
Soliman TH, Elmasry M, Hedar A, Doss M (2014) Sentiment analysis of Arabic slang comments on facebook. Int J Comput Technol 12(5):3470
Stone PJ, Dunphy DC, Smith MS (1966)
Thelwall M (2013) In: Proceedings of the CyberEmotions, pp 1–14
Tian F, Zheng Q, Zhao R, Chen T, Jia X (2009) Can e-Learner's emotion be recognized from interactive Chinese texts? Int Conf Comput Support Cooperative Work Des 13:546
Vohra MS, Teraiya J (2013) Applications and challenges for sentiment analysis: a survey. Int J Eng 2(2):1
Wahsheh HA, Al-Kabi MN, Alsmadi IM (2013) In: Applied electrical engineering and computing technologies (AEECT), 2013 IEEE Jordan conference on IEEE, pp 1–6
Wang W, Wu J (2011) Emotion recognition based on CSO&SVM in e-learning. Int Conf Nat Comput 7:566
Wang W, Chen L, Thirunarayan K, Sheth AP (2012) Harnessing twitter "big data" for automatic emotion identification. Int Conf Soc Comput (SocialCom) 4:587
Whitelaw C, Garg N, Argamon S (2005) In: Proceedings of the 14th ACM international conference on information and knowledge management. ACM, pp 625–631
Yassine M, Hajj H (2010) In: Data mining workshops (ICDMW), 2010 IEEE international conference on IEEE, pp 1136–1142
Yessenov K, Misailovic S (2009) Methodology, pp 1–17
Yoshida Y, Hirao T, Iwata T, Nagata M, Matsumoto Y (2011) Transfer learning for multiple-domain sentiment analysis-identifying domain dependent/independent word polarity. AAAI Conf Artif Intell 25:1286
Zaidan OF, Callison-Burch C (2011) In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, vol 2 Association for Computational Linguistics, pp 37–41
Zeroual I, Lakhouaja A, Belahbib R (2017) Towards a standard part of speech tagset for the Arabic language. J King Saud Univ Comput Inf Sci 29(2):171
Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B (2011) Combining lexicon-based and learning-based methods for twitter sentiment analysis. Technical Reports. http://www.hpl.hp.com/techreports/2011/HPL-2011-89.html
Zhao S, Zhong L, Wickramasuriya J, Vasudevan V (2011) Analyzing Twitter for social tv: sentiment extraction for sports. Int Workshop Future Telev 2:11