Modern Trends in Arabic Sentiment Analysis: A Survey

Hala Mulki*[*] — Hatem Haddad[**] — Ismail Babaoğlu[*]

[*] *Department of Computer Engineering, Selcuk University, Turkey.*
[**] *Department of Computer & Decision Engineering, Université Libre de Bruxelles, Belgium.*

ABSTRACT. *The growth of the Arabic textual content on social media platforms has been caused by the continuous crises in the Arab World evoking the need to analyze the opinions of the public against the ongoing events. Arabic Sentiment Analysis (ASA) is, therefore, becoming the focus of many recent NLP studies. With several Arabic NLP resources being publicly available along with the emergence of deep learning techniques, researchers could handle the complex nature of Arabic language more efficiently. In the last decade, various ASA systems have been built. Yet, their achievements have not been investigated or compared against each other. This survey covers the ASA research carried out during the past five years. We compare and evaluate the performances and give insight into the ability of the created Arabic resources to support the future ASA research.*

RÉSUMÉ. *La croissance du contenu arabe dans les médias sociaux a été causée par les crises dans le monde arabe, évoquant la nécessité d'analyser les réactions du public à l'égard des événements en cours. L'analyse des sentiments de la langue arabe est au centre d'intérêt de plusieurs études en TAL. Avec l'émergence de plusieurs TAL ressources en arabe accessibles au public ainsi que l'émergence de techniques d'apprentissage approfondi, les chercheurs pourraient gérer la nature complexe de la langue arabe plus efficacement. Au cours de la dernière décennie, plusieurs systèmes d'analyse du sentiment dans le contenu arabe (ASA) ont été développés. Cependant, leurs performances n'ont pas été étudiées ou comparées entre elles. Cette enquête couvrirait les travaux de ASA réalisés au cours des cinq dernières années. Nous comparons les résultats, évaluons les performances et donnons un aperçu de la capacité des ressources arabes créées à soutenir la recherche future dans le domaine ASA.*

KEYWORDS: *Arabic Sentiment Analysis, supervised learning, lexicon-based, word embeddings.*

MOTS-CLÉS : *analyse du sentiment dans le contenu arabe, apprentissage supervisé, approche basé sur le lexique,* embedding *de mots.*

1. Introduction

Online shared opinions towards events or products are becoming a rich source of information required for analytical studies. Sentiment Analysis (SA) is a Natural Language Processing (NLP) task that facilitates performing such studies by mining the opinionated content in a piece of text (Piryani *et al.*, 2017). The uprisings in the Arab world that started in 2010 have led to a significant growth in the online Arabic content shared across social media platforms. The ability to analyze such vast opinionated content has attracted the attention of NLP researchers. Consequently, multiple Arabic sentiment analysis (ASA) systems could be developed to capture the sentiment at the different analysis levels; some of them used traditional machine learning approaches, others exploited semantic resources through lexicon-based models, while more recent studies employed the newly-emerged deep learning techniques. Through the presented ASA research, several semantic resources, annotated datasets and trained word embedding vectors were created. Therefore, it is crucial to conduct a comprehensive evaluation of what has been achieved in order to give insight into the potential improvements that could be done in terms of SA tools and resources.

In this paper, we present a survey of ASA research introduced during the last five years. The reviewed research works have been classified according to the used method and the analysis level. In addition, we compare the obtained results and evaluate the performances to recognize by which method, using which semantic resource and for which dataset a better performance can be achieved. Moreover, we shed light on the potential future exploitation of the produced tools and resources to develop more efficient ASA systems.

2. Sentiment Analysis

Sentiment refers to the people's opinions or emotions towards entities, events or ideas. It represents the opinionated content that implies a positive, negative or neutral polarity expressed within a written text (Turney, 2002). According to Liu (2012), sentiment analysis (SA) or opinion mining aims to develop automated techniques to analyze the opinions embedded in a piece of text.

Most of the proposed SA models adopt a general pipeline as follows:

1) Preprocessing: it reduces the noisy nature of the input text, especially for data derived from social media. NLP techniques such as tokenization, normalization, stemming, lemmatization, Part Of Speech (POS) tagging, denoising and stopwords removal are usually used. Consequently, the none-sentimental content represented by special characters, punctuation, duplicate-characters, typos etc. could be eliminated. Moreover, stemming and lemmatization reduce the size of features to be used in the subsequent phase as inflected words are returned to their roots or lemmas (Assiri *et al.*, 2015);

2) Feature Extraction and Selection: the preprocessed data forwarded to this phase facilitates syntactic features extraction since some preprocessing tasks like POS tagging, stemming and lemmatization, negation and emotion tagging can be considered key indicatives of the sentiment (Assiri *et al.*, 2015). In addition, through tokenization, the common bag-of-words and n-grams feature schemes are produced. Feature vectors can then be formulated via binary weighting due to the presence/absence of a word or n-gram in a specific input text. Furthermore, the relative importance of a term or an n-gram which is usually decided by its frequency of occurrence in the dataset, reduces the features' dimensionality by keeping terms of specific frequency values. On the other hand, sentiment lexicons provide another set of features where a term's sentiment score or intensity value define the text features. All these features are called "hand-crafted"; they have been used in most of the presented SA works (Mohammad *et al.*, 2013; Al-Osaimi and Badruddin, 2014; Abdulla *et al.*, 2013; Salameh *et al.*, 2015). More recently, a novel type of features has emerged, the so-called text embeddings where words, phrases and sentences are mapped into real-valued, low-dimensional feature vectors to be used within deep learning systems (Collobert *et al.*, 2011; Al Sallab *et al.*, 2015; Mdhaffar *et al.*, 2017);

3) Sentiment Classification: in natural languages, the sentiment is normally included in the subjectivity concept as the latter represents the language's aspects of opinions and impressions (Liu, 2012). Therefore, SA involves performing a subjectivity classification task first so that a unit of text (term, phrase, sentence or document) is classified as either objective or subjective. Then, the subjective text is classified into the polarity it implies which might be positive, negative, neutral or even mixed. Sentiments can be annotated at various levels of granularity: word or phrase, aspect, sentence and document. Regardless of the level at which sentiment is captured, the sentiment classification process is conducted using one of two main approaches: Machine Learning (ML) or rule-based. Both approaches exploited syntactic, lexicon-derived or embeddings features and were applied successfully in many SA research (Turney, 2002; Salameh *et al.*, 2015; Altowayan and Tao, 2016).

3. Arabic Sentiment Analysis Challenges

Sentiment analysis has become a very active area of NLP research since the advent of Web 2.0 technologies. Nevertheless, most of the presented SA research has been dedicated towards Indo-European languages while under-represented languages such as Arabic were remarkably less tackled. Despite the recent growth of the public Arabic content across social networks and with the continuous development of Arabic NLP tools, ASA research still faces challenges, most of which are related to the Arabic language itself. Arabic has three main variants: Classical Arabic used in Quran, Modern Standard Arabic (MSA), which is the formal type of Arabic, and the informal Arabic known as colloquial or Dialectal Arabic (DA) which combines several different dialects (Al-Kabi *et al.*, 2013). With such variety, where each form of Arabic has its own complexities which are represented by special linguistic and morphological features, SA has to handle further issues beyond those already existing for textual data.

Here, we highlight the major challenging issues encountered while conducting ASA:

– *Complex morphology*: being a Semitic language, Arabic adopts the root-and-pattern representation where a single set of consonants called the “root” is used to derive a variety of words by adding vowels (a,o,i) (ا، و، ي) or short vowels (diacritics) in addition to other consonants (Habash, 2010). The inflectional morphology, however, is observed through the ability of Arabic language to express a word in several grammatical categories while keeping the same meaning. The word’s inflected forms can be obtained for several categories such as person, tense, voice (active/passive), number, gender, etc. Consequently, with such high derivational and inflectional morphology, handling Arabic texts through customizing current English SA systems and tools might be limited (Habash, 2010). Thus, special preprocessing tasks supported by Arabic-oriented morphological analyzers should be combined in ASA systems;

– *Lack of resources*: despite the abundant online Arabic content, there is a lack of Arabic sentiment datasets and sentiment lexicons. During the last decade, some datasets have been constructed either for MSA or DA, nevertheless, the number of sentiment datasets which are publicly available remains little (Assiri *et al.*, 2015). Besides, most of these datasets do not have enough amount of data which affects the evaluation of ASA systems when compared to English SA models since the sentiment analysis accuracy depends on the size of the manipulated data. On the other hand, the difficulties that accompany the construction and annotation process of sentiment lexicons have hindered the provision of large-scale and highly-coverage Arabic lexicons, especially with the existence of different Arabic dialects and domains;

– *Negation and sarcasm*: negation in Arabic is expressed using specific negation words which indicate the meaning “not”; some of them are: “ما”, “لم” and “لا”. Negation should be accurately detected and handled as it can convert the meaning of a sentence yielding a quite opposite polarity. This task becomes more difficult and challenging when dealing with DA where negation words are so different from formal MSA ones and have several meanings such as “مو” meaning “not” in the Levantine dialect that can be used for negation (e.g. السلطة مو تازة[1]) or interrogative (e.g. تجي بوكرا، مو[2]) which might mislead the sentiment classifier. Another ambiguity faced by ASA models is the sarcasm issue in which the explicit polarity totally opposites the meant sentiment as in e.g. بعد الانتظار لساعتين، نفدت كل التذاكر، كم انا محظوظ[3], where the word “محظوظ”, which means “lucky”, indicates a positive sentiment while in the example it actually refers to the opposite;

– *Arabizi usage*: Arabizi is considered a newly-emerged Arabic variant written using the Arabic numeral system and Roman script characters (Assiri *et al.*, 2015). It is commonly used while expressing DA across social media and poses a challenge

[1] “The salad is not fresh.”

[2] “You’re coming tomorrow, aren’t you?”

[3] “After waiting for two hours, all tickets were sold; Lucky me.”

to sentiment analysis when it is mentioned along with Arabic (e.g. 3an jad كتير الفلم 7elou.[4]). This requires proper tools to interpret Arabizi into either MSA or DA before conducting the sentiment classification task;

– *Dialects variances*: DA forms the majority of the online opinionated Arabic content as it is commonly used across social media platforms. DA combines various dialects which differ according to the geographical location. Each dialect has its own vocabulary, syntactic and grammatical rules in addition to special idioms. On the other hand, despite that all dialects are derived from MSA and hence do share some vocabulary, common words or expressions among two dialects might have drastically different sentiments. For example, "يعطيك العافية" is a compliment of a positive sentiment that means "May God grant you health" in the Levantine dialect, while this very same phrase has an aggressive meaning of "Burn in hell" in the Tunisian dialect. Considering these variances, an ASA system that targets one dialect might not be efficient for another as it is developed with a dialect-dependent tools such as the morphological analyzer, stopwords/negation words and sentiment lexicons.

4. Arabic Sentiment Analysis Research

Earlier ASA studies had to handle the complex nature of Arabic through limited feature types and resources. However, with more MSA and DA morphological analysis and disambiguation tools becoming available, the ASA task was facilitated as these tools could provide a wide variety of syntactic and stylistic features such as 1-best tokenization, POS tags, stems, lemmas and diacritization in one fell swoop. On the other hand, exploitation of web forums and social media enabled the provision of sentiment datasets and lexicons needed for developing and evaluation of MSA and multi-dialectal SA systems (Rushdi-Saleh *et al.*, 2011; Mourad and Darwish, 2013; Badaro *et al.*, 2014; Nabil *et al.*, 2015).

ASA research has been conducted at different linguistic levels: word or phrase, aspect, sentence and document. The following subsections review the recent major studies achieved at each level.

4.1. *Words-level Sentiment Analysis*

Determining the semantic orientation of sentiment-bearing words or phrases in a corpus is essential for sentiment lexicon construction. Sentiment lexicons are fundamental for computing the sentence or document sentiment through lexicon-based methods or as features for machine learning methods. Sentiment lexicons can be compiled by means of three strategies: manually with the assistance of a linguist and native speakers, automatically based on another dictionary (dictionary-based) or us-

[4] "The film is really amazing."

Paper	Construction method	Size	Arabic variant	Assigned polarity
(El-Beltagy and Ali, 2013)	Corpus-based	4,392	Egyptian/MSA	Pos/Neg
(Abdulla *et al.*, 2014)	Manually Semi-automatic Corpus-based	4,815 9,100 8,618	MSA & DA	Pos/Neg
(Duwairi *et al.*, 2015)	Manually	2,376	MSA	Pos/Neg
(Assiri *et al.*, 2017)	Corpus-based + Dictionary-based	14,000	Saudi	Pos/Neg
(Abdul-Mageed *et al.*, 2014)	Manually	3,982	MSA /DA	Pos/Neg

Table 1. *The lexicons constructed and evaluated within the reviewed ASA studies.*

ing the corpus itself (corpus-based) or semi-automatically where manual interference is needed to normalize the automatically-built lexicon (Liu, 2012).

Considering the lexicons built within the works reviewed here (see Table 1), Abdulla *et al.* (2014) presented three sentiment lexicons built using manual, semi-automatic and automatic methods. The first lexicon has 4,815 entries. It was manually constructed through translating seed words from SentiStrength English lexicon using an English-Arabic dictionary with their polarity assigned manually. The translated seeds were then expanded by adding synonyms of each word under the same polarity in addition to the most common MSA words derived via Term Frequency (TF) weighting, emotions and dialectal terms from different Arabic dialects. The second one was built through the direct translation of SentiStrength using Google translate. Human interference was needed to normalize and clean the translated Arabic version yielding a lexicon of 9,100 entries. As for the third lexicon, it was compiled using a corpus-based automated approach in which the most common positive and negative terms were derived from the annotated corpus via TF weighting. For terms having both polarities, the polarity of the term whose TF is greater was adopted.

In Duwairi *et al.* (2015), the authors adopted the same scenario used in Abdulla *et al.* (2014) to manually build an MSA sentiment lexicon of 2,376 words. The lexicon was expanded through adding synonyms using Sakhr dictionary (Reyes and Rosso, 2014), stems and emotions.

Based on an assumption that sentiment terms often appear with other terms of same polarity, El-Beltagy and Ali (2013) presented an Egyptian lexicon built using a corpus-based method. As a first step, a list of 380 sentiment words seeds was used. Then following their hypothesis, the authors expanded this list by looking for patterns containing these seeds and their accompanied single terms.

Assiri *et al.* (2017) constructed a Saudi dialect lexicon by integrating the dictionary-based and corpus-based methods. A list of Saudi seed words was expanded using the method by El-Beltagy and Ali (2013), then terms from a pre-created lexicon by Badaro *et al.* (2014) were added to the Saudi lexicon after they were subjected to normalization and cleaning processes, then a collection of Saudi terms were manually added resulting in a lexicon of 14,000 sentiment terms.

4.2. *Aspect-level Sentiment Analysis*

Aspect-level SA identifies sentiment targets crucial for applications such as question-answering and recommendation systems (Liu, 2012). Due to the complexity of the Arabic language, aspect-level SA was less tackled by ASA research. In Farra and McKeown (2017), the authors proved that handling the richness of the Arabic language through specific morphological representations makes important targets (entities) and the sentiment towards them better identified. This was done using a framework of two cascaded sequence labeling CRF models: target-specific model and sentiment-specific model. While the first model is responsible for recognizing the entities, the second one predicts the sentiments towards these entities. Both models were trained to provide a sequence of entity/sentiment labels for the input tokens. The merit provided by this system is that the training phase involves learning the syntactic relations between entities and sentiment-bearing words. For this purpose, MADAMIRA morphological analyzer (Pasha *et al.*, 2014) was exploited as it enables an advanced tokenization with which multiple morphological representations could be formulated. For instance, clitics such as the definition article "ال التعريف" that usually indicates an entity could be split off the word, combined with a detailed POS feature and fed into the entity-specific model leading to an improved recall of the recognized entities. The proposed system was evaluated using 1,177 online comments with annotated targets and a lack of punctuation obtained from Arabic Opinion Target (Farra *et al.*, 2015) which is a part of Qatar Arabic Language Bank (QALB) corpus (Zaghouani *et al.*, 2014). The experimental results concluded that both models have achieved better results compared to multiple lexical baselines.

4.3. *Document and Sentence-level Sentiment Analysis*

Document and sentence-level SA works form the majority of the recent ASA research. Therefore, we dedicate the next section to cover the important SA studies conducted at this level. The reviewed studies are organized according to the methods used to build the ASA models.

5. Document and Sentence-level Sentiment Analysis Approaches

Arabic Sentiment Analysis can be conducted using traditional machine learning approaches such as supervised/unsupervised and hybrid, deep learning approaches (supervised, unsupervised) or rule-based approaches (Lexicon-based). The following subsections introduce a detailed explanation of some of these methods in addition to the state-of-the-art research related to it.

5.1. *Supervised Learning-based Approaches*

Supervised learning methods require a labeled corpus to train the classifier how to predict the text polarity (Biltawi *et al.*, 2016). The learning process is carried out by inferring that a combination of a sentence's specific features yields a specific polarity class (Shoukry and Rafea, 2012). The most common features used are bag-of-words and bag-of-n-grams features in addition to various linguistic features extracted by morphological analyzers. Having the features extracted, sentiment classification is then performed using supervised learning classification algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), Decision Tree (DT) and NEUNET (NN) (Biltawi *et al.*, 2016). Research works that adopted supervised approaches were concerned about which preprocessing tasks, features or classification algorithms can lead to a better classification performance either for MSA or DA.

Considering the wide spread of the Egyptian dialect across Twitter, enriching the Arabic sentiment resources with a pure Egyptian sentiment corpus along with Egyptian-specific preprocessing tools was the aim of Shoukry and Rafea (2012). They collected a dataset of 1,000 positive/negative Egyptian tweets to test their supervised SA model. The preprocessing included removing usernames, hashtags, URLs and non-Arabic letters. To classify a tweets' sentiment, SVM and NB were employed in two experiments. In the first one, stopwords were kept, while in the second they were omitted. Results revealed that SVM performed better than NB in both experiments achieving an accuracy of 72%, compared to 65% scored by NB.

To examine the impact of combining emotion icons in SA, Al-Osaimi and Badruddin (2014) introduced an SA model for multi-dialectal tweets. The collected corpus included 3,000 positive, negative and neutral tweets. Term Frequency-Inversed Document Frequency (TF-IDF) was used to extract the features. Sentiment classification was conducted using NB and KNN algorithms. Results showed that preserving emotion icons enhanced the model's performance as the best accuracy achieved by NB classifier increased from 58.28% to 63.79%.

The recently-emerged form of Arabic (Arabizi) was investigated in Duwairi *et al.* (2014). The study seeked to convert the dialectal and Arabizi content into MSA. A dataset of 1,000 positive/negative/neutral tweets written in Jordanian and Arabizi was collected. For preprocessing, stemming, tokenization, stopwords filtering tasks were applied in addition to the conversion of Jordanian and Arabizi to MSA. Morphological features, negations and emoji were included in the features set. The authors observed that, if stemming and stopwords removal are disabled, better performance can be achieved, while negation detection and conversion from Arabizi to MSA did not achieve a remarkable improvement in the evaluation measures. KNN, SVM and NB classifiers were used, where NB was the best with an accuracy of 76.78%.

In Salamah and Elkhlifi (2016), an under-represented Arabic dialect was investigated where a dataset of 340,000 Kuwaiti tweets were collected and manually annotated for positive and negative polarity. Tweet-related features and opinions-oriented ones were extracted. The opinion-oriented features were obtained from 22 manually-

Paper	Algorithm/features	Dataset	Evaluation
(Shoukry and Rafea, 2012)	SVM, NB unigrams+bigrams	1,000 tweets Egyptian pos/neg	Best: SVM acc=72%

(Al-Osaimi and Badruddin, 2014)	NB, KNN
	TF-IDF
	unigrams
		3,000 tweets
		multi-dialects
		pos/neg/neut
			Best: NB
			acc=58.28% (-emoji)
			acc=63.79% (+emoji)

(Duwairi *et al.*, 2014)	KNN, SVM, NB syntactic, negation emoji	1,000 tweets Jordanian/Arabizi pos/neg/neut	Best: NB acc=76.78%
(Salamah and Elkhlifi, 2016)	SVM, J48, RT, DT tweet-related emotion-bearing words	340,000 tweets Kuwaiti pos/neg	Best: SVM F1=71.5%
(Abdul-Mageed, 2015)	SVM, NB, IB1 POSs tokens	1,552 sentences MSA subj/obj	Best: SVM acc=85%

Table 2. *Summary of supervised learning-based ASA research works.*

built classes that combine emotions-bearing words. SVM, J48, Random Tree (RT) and decision Tree (DT) classifiers were used. SVM scored the best results with an F1-score of 71.5% compared to 42%, 48% and 51% achieved by J48, DT and RT respectively. A summary of the above mentioned research papers is listed in Table 2 where Best, acc and F1 indicate the best method, the scored accuracy and F1-score respectively.

One of the pioneering works about subjectivity detection in MSA was presented by Abdul-Mageed (2015). The author hypothesized that using specific tokens would favorably impact the subjectivity classification task. The proposed model was trained with a collection of words having certain POS tags such as ADJ, ADV and NOUN_PROP. The experiments were conducted using Penn Arabic Treebank dataset (Popescu and Etzioni, 2007) with several ML techniques applied: SVM, NB, and Instance-based learning. These techniques were compared against each other with two types of features' settings: frequency and presence vectors. In all experiments, the preprocessing step was essential as the study highlighted that the rich morphology of MSA imposes using the compressed form of words in order to obtain a better model generalization. The obtained results emphasized the positive impact of using certain tokens rather than all the words for training; moreover, similar to the SA task, SVM was found of the best performance for subjectivity classification, compared to other ML methods where it scored a high accuracy equals to 85%.

Hybrid approaches combine lexical and linguistic features together with lexicon-derived features to perform sentiment analysis. This involves incorporating the term's polarity score defined by a sentiment lexicon in the features set needed to train a supervised sentiment classifier (Biltawi *et al.*, 2016). Abdul-Mageed *et al.* (2014) studied the efficiency of standard and genre specific features when used to express MSA and

DA seeking for the best scheme to represent lexical information within SA context. To do that, the authors constructed an adjective sentiment lexicon to enrich the lexical features. Their SA system SAMAR has two classification stages: subjectivity and polarity classification. Four MSA/dialectal datasets were collected manually including reviews and tweets. Syntactic features, extracted via AMIRA morphological analyzer (Diab, 2009), were adopted in addition to an extra feature resulted from the matches between the input tokens and the adjectives contained in the manually-built lexicon. Moreover, a novel feature that distinguishes MSA from DA was added. The used lexicon includes 3,982 labeled adjectives. Experimental study showed that using SVM trained with the different features enabled beating the baselines for most datasets either for subjectivity classification with a best accuracy of 73% or for the sentiment classification with an accuracy of 70.30% for DAR dataset.

To compensate for the lack of publicly available resources, Salameh *et al.* (2015) suggested using publicly available English NLP tools and lexical resources. This study presented an ASA model that employs an English SA system with an English lexicon on a translated Arabic content. Four datasets of positive/negative/neutral tweets and social media posts written in MSA/dialectal were used. Preprocessing included normalization, tokenization and POS tagging to produce syntactic and stylistic features. An English SA model NRC-Canada designed by Mohammad *et al.* (2013) was modified to handle the Arabic text along with a translated version of NRC Hashtag Sentiment Lexicon. The Arabic content translated to English was targeted using the system developed in Kiritchenko *et al.* (2014). The obtained accuracy values for Levantine datasets were 78.65% for the Syrian dataset and 63.89% for BBN.

Baly *et al.* (2017) introduced a hybrid model OMAM whose features were inspired from the English SA model (Balikas and Amini, 2016). An equivalent set of surface, syntactic and semantic features were obtained with the assistance of MADAMIRA from Pasha *et al.* (2014) and SAMA by Maamouri *et al.* (2010) morphological analyzers. Additional features were provided by ArSenL (Badaro *et al.*, 2014), AraSenti (Al-Twairesh *et al.*, 2016) and ADHL (Mohammad *et al.*, 2016) lexicons. Preprocessing phase included replacing emotions, URLs and hashtags with special tokens. The model was applied on dialectal Arabic tweets provided by SemEval-2017 (Rosenthal *et al.*, 2017). Results indicated that SVM classifier trained with the previous features achieved an F1 score of 42.2%, a recall of 43.8% and an accuracy of 43%.

With the key role of the lexicon-derived features in improving the performance of hybrid SA systems, there was a crucial need for a large-scale, domain-independent, high-coverage and publicly-available Arabic lexicon. To meet that need, Al-Moslmi *et al.* (2017) introduced the Arabic senti-lexicon to assist in sentiment classification of multi-domain, multi-dialectal Arabic reviews. The quality of the constructed lexicon towards SA task was assessed through training the model with five types of feature sets most of which were lexicon-derived. Features included sentiment words' polarity-based, sentiment words' presence-based, frequency POS-based, sentence level-based and other features related to words and sentences statistics. SVM, NB, LLR, KNN and neural network (NEUNET) were employed. To evaluate the presented model, the au-

Paper	Hybrid features	Algorithm	Dataset	Evaluation
(Abdul-Mageed *et al.*, 2014)	Linguistic syntactic adjective polarity score from Adj-Lex	SVM several kernels	DAR: 2,798 TGRD: 3,015 THR: 3,008 MONT: 3,097 MSA/dialectal pos/neg/neut	Best: SVM linear kernel acc=70.3% (DAR)
(Salameh *et al.*, 2015)	linguistic word N-grams Char N-grams score from translated-Lex	SVM	1,111 dialectal tweets pos/neg 1,200 Levant comments pos/neg/neut 2,000 Syrian tweets pos/neg/neut 2,681 dialectal tweets pos/neg	acc=85.23% (dialects)
(Baly *et al.*, 2017)	linguistic syntactic emotion presence tweet-related score from MSA/dialects-Lex	SVM	3,355 dialectal tweets pos/neg/neut	acc=43%
(Al-Moslmi *et al.*, 2017)	N-grams sentence-level syntactic score from ArabicSenti-Lex	SVM, NB, LLR, KNN NN	8,861 reviews dialects pos/neg	Best: LLR, NN F1=97%

Table 3. *Summary of Hybrid ASA research works.*

thors created a dataset called Multi-domain Arabic Sentiment Corpus (MASC) including 8,861 positive/negative customer reviews written in several Arabic dialects. Data was first preprocessed in terms of tokenization, normalization, stemming and stop-words removal. The model was trained on each feature set solely, then on all of them combined in one set. Results indicated that, SVM achieved the best results when only POS-based features are included. However, when all features are used for training, LLR, NN and NB were of better performance where LLR and NN achieved an F1-score of roughly 97%, while NB achieved 96% compared to 82.07% and 77.97% F1-scores achieved by SVM and KNN respectively. A summary of the above-mentioned research papers of hybrid SA models is listed in Table 3 where Best, acc and F1 indicate the best method, the scored accuracy and F1-score respectively.

5.2. *Lexicon-based Approaches*

In lexicon-based methods, neither labeled data nor training step are required to design the sentiment classifier. The polarity of a sentence or a document is determined via the lexicon-derived sentiment scores of its constituent words (Liu, 2012). A sentiment lexicon combines a list of subjective words and phrases along with their positive or negative score which denotes the sentiment polarity and strength of a word/phrase (Piryani *et al.*, 2017). Sentiment lexicons can be general-purpose or domain-specific, built either manually or automatically (Abdulla *et al.*, 2013). For each entry in the lexicon, the sentiment weight or score is assigned by one of these weighting algorithms:

– Straight Forward Sum (SFS) method: adopts the constant weight strategy to assign weights to the lexicon's entries such that negative words have the weight of –1 while positive ones have the weight of 1. The polarity of a given text is thus calculated by accumulating the weights of negative and positive terms and the total polarity is determined by the sign of the resulted value. Thus, for a tweet such "غوغل مبدعة شي خرافي"[5], the polarity is calculated as follows: google+incredibly+creative= 0+1+1=+2. The tweet has a positive polarity;

– Double Polarity (DP) method: assigns both a positive and a negative weight for each term in the lexicon. For example, if a positive term in the lexicon has a weight of 0.6, then its negative weight will be: – (1–0.6) =–0.4. Similarly, a negative term of a weight equals to –0.9 would have a 0.1 positive weight. Polarity is calculated by summing all the positive weights and all the negative weights in the input text. Consequently, the final polarity is determined according to the greater absolute value of the resulted sum (Piryani *et al.*, 2017). Thus, the positive score of the previous tweet is [0+0.5+0.8]=1.3 while the negative score=[0+(–0.5)+(–0.2)]=–0.7. Since the positive score is greater than the negative one, this indicates the positive polarity of the tweet assuming that "خرافي"[6] has a positive score of +0.8 and "مبدعة"[7] is of +0.5 positive score.

Lexicons of uniform weight along with the SFS method have been commonly used in most lexicon-based SA research. However, since SFS depends only on the counts of positive and negative words of a sentence to determine its polarity, it might lead to miss-classified instances under the label "neutral". This is encountered when the number of positive words in a sentence equals that of the negative words (Liu, 2012). For example, if a negative word such as "terribly" and a positive one like "exciting" are contained in the same sentence "the movie is terribly exciting", then the sentence's polarity score computed via SFS and uniform weight strategy would be: movie+terribly+exciting=0+(–1)+1= 0 which refers to a neutral polarity, while the previous sentence is obviously bearing a positive sentiment.

[5] "Google is incredibly creative."

[6] "Incredibly."

[7] "Creative."

Several attempts were introduced to develop a novel weighting algorithm with which a better sentiment classification can be achieved. One of these attempts was presented by El-Beltagy and Ali (2013) where the authors noticed that sentiment terms often appear with other terms of same polarity. Based on this theory, they constructed a corpus-based lexicon (details in Section 4.1). Using the resulted lexicon, SFS and DP methods were adopted to determine the positive/negative/neutral sentiment of the input text. The model was also tested with the uniform weight scheme with negation switch policy, intensification words weighting and person names removal applied. Two manually-collected and annotated Egyptian datasets were used. The first one, called Dostour, combines 100 comments, while the second represents a Twitter dataset of 500 tweets. The best performance was achieved by the second weighting strategy with DP method where an accuracy of 83.3% was scored for Twitter dataset while for Dostour dataset, an accuracy of 63% was achieved.

Aiming to evaluate manually-built against the automatically-built lexicons for the SA task, Abdulla *et al.* (2014) examined performing sentiment analysis of MSA/dialectal Arabic using three lexicon variants built via different construction methods (see Section 4.1). In addition, an integrated lexicon resulted from merging the three constructed ones was also utilized for the final system evaluation. Two datasets were used in the experiments, the first contains 2,400 positive/negative comments from Maktoob collected by Al-Kabi *et al.* (2013), while the second combines 2,000 positive/negative/neutral tweets obtained from Abdulla *et al.* (2013) . Data was normalized and a light stemming was applied. Light stemming removes common affixes from words without reducing them to their stems or roots and thus retains the variety of words having same root and different meanings. Sentiment classification was then performed using the four lexicons one by one with SFS method and switch negation policy applied. Experiments showed that the stemming degraded the performance with manually-built and dictionary-based lexicons. In contrast, the accuracy was improved when stemming was applied on the corpus-based lexicon which forms more than half the size of the integrated lexicon. Hence, the best results were achieved with the integrated lexicon achieved since for Maktoob dataset, an accuracy of 74.6% was scored, compared to 70.2% with non-stemming option while for Twitter dataset, the scored accuracy was 70.2%, against 60.75% with non-stemming option.

In Duwairi *et al.* (2015), the authors claimed that when dealing with MSA data, the likelihood of finding a stem in the sentiment lexicon is higher than that of finding the original word. This has been investigated using an MSA sentiment lexicon constructed manually as it was explained in Section 4.1. A dataset of 4,400 positive/negative tweets was manually collected and annotated to evaluate the model. The data was preprocessed such that stopwords were removed while negations were kept. Stemming of the input data was conducted by MSA Khoja stemmer[8]. To investigate the stemming impact, experiments were conducted with/without stemming. SFS method with switch negation policy were employed to calculate the sentiment score of the input tweets. The results revealed that, for such MSA data, stemming has im-

[8] http://zeus.cs.pacificu.edu/shereen/ArabicStemmerCode.zip.

proved the sentiment classification performance where the accuracy improved from 23% to 46%, while F1-score increased from 31.3% to 55.51%.

Paper	Scoring method	Lexicon/Features	Dataset	Evaluation
(El-Beltagy and Ali, 2013)	SFS, DP	Egyptian size:4,392 unigrams	1: 100 comments 2: 500 tweets Egyptian pos/neg/neut	Best: DP 1: acc=83.3% 2: acc=63%
(Abdulla *et al.*, 2014)	SFS	MSA/dialectal size: 19,800 unigrams	1: 2,400 comments 2: 2,000 tweets MSA/dialectal pos/neg/neut	+stemming 1: acc=74.6% 2: acc=70.2%
(Duwairi *et al.*, 2015)	SFS	MSA unigrams	4,400 tweets MSA pos/neg	+stemming F1 =55.51%
(Assiri *et al.*, 2017)	WLBA, SFS DP	Saudi/dialects size:14,000 lexicon term length, negation and supplication	1: 4,700 Saudi tweets pos/neg 2: 500 Egyptian tweets pos/neg/neut	Best: WLBA 1: acc=81% 2: acc=76%

Table 4. *Summary of lexicon-based ASA research works.*

Unlike the above-mentioned methods, which employed pre-weighted lexicons to determine the sentiment score, Assiri *et al.* (2017) introduced a polarity weighting algorithm called WLBA which assigns weights to the polarity words by learning from the data itself. This algorithm considers the polarity words' context as it explores and counts how frequently a pair of (polarity, non-polarity) words co-occurs. Later, it assigns a weight to the polarity word due to its associations' count with the non-polarity word in the whole corpus. A Saudi lexicon was built using corpus-based and dictionary-based approaches (see Section 4.1). Upon applying the model on Egyptian dataset from (El-Beltagy and Ali, 2013) and a manually-collected Saudi dataset of 4,700 tweets, results showed that WLBA achieved a poor performance compared to SFS and DP for both datasets due to ignoring complex structural and lexical specifications of the Saudi corpus. However, when features like negation and supplication were accurately handled via rule-based methods, WLBA outperformed other methods with an accuracy of 81%, compared to 72% and 43% scored by SFS and DP methods respectively. Additionally, for the Egyptian dataset, the achieved accuracy was 76%, compared to 71% and 68% scored by SFS and DP method respectively. Table 4 lists a summary of the above-mentioned lexicon-based research papers where Best, acc and F1 indicate the best method, the scored accuracy and F1-score respectively.

5.3. *Deep Learning Approaches*

Deep learning approaches are representation learning methods which learn discriminative features automatically from the data through either an unsupervised manner or via a supervised strategy, more specifically, self-supervised learning which is an instance of supervised learning whereby the training labels are determined by the input data (here document statistics such as the usage of words) (Gomez *et al.*, 2017). In addition, for a specific task like document classification, such embeddings can be learned within a neural network model trained on annotated data (Mikolov *et al.*, 2013). These methods can learn continuous and real-valued multiple levels of text representation using multi-layer nonlinear neural networks where each layer transforms the representation at one level into a representation at a higher and more abstract level (Mikolov *et al.*, 2013). The learned representations can be divided into two types:

– Word embeddings: where every word in the corpus is mapped to a real-valued low-dimensional vector in the embedding space using one of the word mapping algorithms such as word2vec (Mikolov *et al.*, 2013) and GloVe (Pennington *et al.*, 2014);

– Document embeddings: generate continuous representations of larger blocks of text such as sentences, paragraphs or whole documents using a document mapping algorithm such as doc2vec (Le and Mikolov, 2014).

Both representation types can be used as features for further classification tasks such as sentiment classification. Indeed, text embeddings features have been successfully applied in recent ASA research as they could capture the fine-grained semantic and syntactic regularities within the input text (Le and Mikolov, 2014). In addition, the automatic feature extraction, the low-dimensionality and less data sparsity of the embedding vectors have made deep learning-based SA models competitive to hand-crafted-based ones (Mikolov *et al.*, 2013).

In Altowayn and Tao (2016), the authors replaced the hand-crafted features with efficient features produced without much effort to be adopted for the sentiment analysis task. With the application of minor preprocessing, word embeddings features were used as discriminative features to train several supervised classifiers. The used embeddings were generated using Continuous bag of words (CBOW) learning algorithm (Mikolov *et al.*, 2013) and an MSA/DA training corpus of 190 million words. The authors indicated that their embeddings model could handle dialects efficiently as it mapped different writing shapes of dialectal words close to each other in the embedding space. To perform the SA task, fixed-sized embedding vectors were learned for a combination of three datasets of multi-dialectal tweets: ASTD (Nabil *et al.*, 2015), ArTwitter (Abdulla *et al.*, 2013) and QCRI (Mourad and Darwish, 2013), in addition to other two datasets representing book reviews: LABR (Aly and Atiya, 2013) and MSA news articles derived from the translated MPQA corpus (Banea *et al.*, 2010). Results showed that, for subjectivity classification of the MPQA dataset, the presented model has slightly improved the performance compared to hand-crafted features-based systems of Banea *et al.* (2010) and Mourad and Darwich (2013) where it achieved an accuracy of 77.87% and F-score of 76.14%. As for the polarity classification, best metrics values were scored by the logistic regression algorithm with an accuracy of 81.88% and F-measure of 81.58%.

Arabic word embeddings are usually learned using large-scale training corpora so that they could cover the vocabulary of the dataset to be sentimentally classified. Thus, the learning process is considered. costly in terms of the time needed for training. This could be avoided if pretrained Arabic word embeddings were included in a neural SA model. Gridach *et al.* (2017) have investigated this idea where an ASA model was developed using word embeddings provided by Zahran *et al.* (2015) and previously trained with MSA/dialectal corpora using three word representations: Glove, SG and CBOW. These representations were examined as initializing vectors of the input words fed to a deep learning SA model built using Convolutional Neural Networks (CNNs). The proposed model CNN-ASAWR was developed as a variant of Collobert *et al.* (2011) system. The trained model was applied on two MSA/dialectal datasets: ASTD (Nabil *et al.*, 2015) and SemEval-2017 (Rosenthal *et al.*, 2017). Results showed that Arabic pre-trained word representations can be considered as universal feature extractors used for the sentiment classification task as better performances were achieved. In ASTD dataset for instance, the best F-measure scored by CNN-ASAWR was 72.14%, compared to 62.60% achieved by Nabil *et al.* (2015) while for SemEval-2017, an F-measure of 63% was achieved against 61% scored by the system of El-Beltagy *et al.* (2017) which ranked first in SemEval competition.

With the lack of lexical and semantic resources especially for under-represented Arabic dialects, paragrah embeddings represent an alternative expressive features for DA. Based on that, the authors in Mdhaffar *et al.* (2017) investigated representing Tunisian comments by distributed paragraph representations to be used as features in a Tunisian SA model. Their model was evaluated using a combination of publicly available MSA/multi-dialectal datasets: OCA (Rushdi-Saleh *et al.*, 2011), LABR (Aly and Atiya, 2013) and a manually annotated Tunisian Sentiment Analysis Corpus (TSAC) obtained from Facebook comments. Doc2vec algorithm by Le and Mikolov (2014) was applied to generate document vectors of each comment. The produced vectors were then fed SVM, Bernoulli NB (BNB) and Multilayer Perceptron (MLP) classifiers with various combinations of MSA, dialects and Tunisian used as training sets. The best results were scored by MLP classifier when TSAC corpus was solely used as a training set where it achieved an accuracy equals to 78% and an F1-score of 78%.

Each deep learning architecture has specific merits which are usually related to its building unit. Baniata and Park (2016) investigated the impact of using a combination of CNN and Bidirectional-Long Short Term Memory (BiLSTM) on SA of MSA/dialectal tweets. They relied on the fact that the phrase representation of every sentence captured by CNN can be further enhanced by using BiLSTM network which can capture the contextual information and thus yields an improved performance. Two configurations were examined: CNN-BiLSTM, which involves generating the sentence representation to be improved later by the context information derived from both direction, and BiLSTM-CNN, where contextual information is first captured then fed to CNN to assist in generating the sentence representation. The used CNN model

Paper	**Embedding**	**Dataset**	**Classifier**	**Evaluation**
(Al Sallab *et al.*, 2015)	Recursive parsing tree	LDC-ATB MSA pos/neg	DNN, DBN DAE, RAE Linear-SVM	Best: RAE acc=74.3%

(Altowayan and Tao, 2016)	word2vec
	(CBOW)
		LDC-ATB
		ASTD, ArTwitter
		QCRI, LABR
		MPQA
		MSA/dialects
		pos/neg
			LR, SGD
			GNB, RF
			Linear-SVM
			Nu-SVM
				Best (MSA):
				Linear-SVM
				acc=77.87%
				Best (dialects):
				LR
				acc=81.88%

(Gridach *et al.*, 2017)	word2vec
	(skip-gram)
	(CBOW)
	Glove
		ASTD
		SemEval 2017
		MSA/dialects
		pos/neg/neut
			CNN	F-score=72.14%
				(ASTD)
				F-score=61%
				(SemEval2017)

(Baniata and Park, 2016)	word2vec pretrained vectors	LABR MSA/Dialects pos/neg	CNN-BiLSTM BiLSTM-CNN	Best: CNN-BiLSTM acc=86.43%

(Mdhaffar *et al.*, 2017)	Doc2vec	113,196 tweets
		OCA, LABR
		Tunisian/dialects
		pos/neg
			SVM
			MLP
			BNB
				Best: MLP
				prec=78%
				recall=78%

(Al Sallab *et al.*, 2017)	Recursive syntactic parsing tree	Tweets QALB ATB MSA/dialects pos/neg	AROMA (modified RAE) DNN, DBN DAE-DBN, RAE NB, Linear-SVM	Best: AROMA acc=86.5%

Table 5. *Summary of Deep learning-based ASA research works.*

contained layers of filter sizes 3, 4 and 5 with the activation function ReLu used in both configurations. Both ensembles were evaluated using LABR dataset (Aly and Atiya, 2013). The data was normalized first through removing diacritics, punctuations and non-Arabic characters, and the vocabulary size was reduced by keeping words of frequency greater than 10. Word embeddings were then obtained based on pre-trained word vectors by Al-Rfou *et al.* (2013). It was noted that CNN-BiLSTM architecture achieved an accuracy of 86.43%, whereas BiLSTM–CNN architecture has suffered from of overfitting after the fifth epoch yielding an accuracy of 66.26%.

The variety of deep learning architectures has evoked the question about which architecture can perform better for ASA analysis. Therefore, Al Sallab *et al.* (2015) explored four deep learning models of different architectures and compared their performances within the context of ASA. The first three models are: Deep Neural Network (DNN), Deep Belief Network (DBN) and Deep Auto Encoders (DAE). While DNN model employs the back propagation in a conventional neural network with several layers, DBN avoids overfitting through a pretraining phase before feeding a discriminative fine tuning step whereas DAE provides a compact representation of the input sentence with a reduced dimensionality. These models were trained using the ordinary Bag-of-Words features along with lexicon features derived from ArSenL lexicon (Badaro *et al.*, 2014). As for the fourth model, Recursive Auto Encoder (RAE), it was suggested to address the lack of context handling procedures issue found in the previous three models. RAE can parse raw sentence words in the best order for which the error of recreating the same sentence words in the same order is as minimum as possible. This is done via a recursive parse tree where the sentence words are parsed recursively till finding the best words' order. The evaluation was performed using Linguistic Data Consortium Arabic Tree Bank[9]. Upon comparing the performances of the four models in positive/negative sentiment classification against an SVM model with hand-crafted features, it was noted that the performance of DNN, DBN and DAE was close to SVM's, while DAE provided a better representation for the input sparse sentence vector. The RAE model outperformed all the other models achieving an accuracy of 74.3% and F1-score of 73.5%, compared to an accuracy of 45.2% and F1-score of 44.1% scored by linear SVM. This indicates the privilege of recursive models compared to one-shot models in terms of learning accurate semantic representations.

According to Al Sallab *et al.* (2015), the efficiency of RAE-based models was attributed to their ability to perform SA without the need for opinion resources or extensive NLP. However, standard RAE models become insufficient to handle Arabic lexical sparsity and ambiguity which limit the model's ability to generalize and causes over-fitting. These issues were addressed in Al Sallab *et al.* (2017) where A Recursive Deep Learning Model for Opinion Mining in Arabic (AROMA) was developed. To enable modeling the semantic interactions at the morpheme level and to reduce the lexical sparsity and ambiguity, the training data was subjected to morphological tokenization using MADAMIRA (Pasha *et al.*, 2014) before it was fed to AROMA. In addition, semantic embedding with/without unsupervised pre-training alongside sentiment embedding were used to provide improved word distributed representations. Furthermore, instead of using the greedy algorithm to define the order of the model's recursion, AROMA employed phrase structures to automatically generate syntactic parse trees with which a better modeling of composition was achieved. The presented model was evaluated using three datasets annotated for pos/neg polarities: an MSA dataset from Abdul-Mageed *et al.* (2011) called ATB, dialectal Tweets dataset by Refaee and Reeser (2014) an MSA/DA comments derived from Farra *et al.* (2015) and referred to as QALB. The experiments involved using different combinations of the contributions augmented to the standard RAE. The results indicated that compared to the standard RAE, AROMA with all the contributions combined could improve the classification accuracy significantly by 12.2%, 8.4% and 7.2% for the ATB, QALB and Tweets datasets, respectively. Moreover, AROMA was evaluated against several ML and DL models where it overcome all of them as it scored an accuracy increment of 7.3%, 1.7% and 7.6% for the same previous datasets respectively.

[9] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=2005T20.

6. Discussion and Conclusion

The complex morphological nature of the Arabic language along with the wide usage of dialects require a careful design of the SA models such that inflected words, various writing styles, typos and varying grammatical nature are handled efficiently. Considering these issues, supervised learning-based methods adopted special preprocessing procedures such as stemming, lemmatization and tokenization. Another hand-crafted features such as n-grams and bag-of-words in addition to syntactic and stylistic features were used by these models. With such a variety of features, feature vectors tend to be of high dimensionality and sparsity which may drown the classifier with noisy features or lead to memory issues as it has been reported in Duwairi *et al.* (2014). To reduce the features' size, feature selection methods like TF-IDF weighting with a specific threshold was adopted as in Shoukry and Rafea (2012) and Al-Osaimi and Badruddin (2014). In the same context, some works suggested reducing the text size through stemming or stopwords removel. However, it has been proved that stemming has no impact on the classification performance, especially for DA since available stemming tools mostly target MSA (Duwairi *et al.*, 2014). Moreover, with the lack of reliable stopwords lists for dialects, keeping stopwords was proved to be better than eliminating them as they can assist in capturing the sentiment (Shoukry and Rafea, 2012). Supervised SA systems are generally robust and accurate. However, performances may vary from one system to another due to the used classifier. It has been noted that SVM usually outperforms other classifiers (Shoukry and Rafea, 2012; Salamah and Elkhlifi, 2016); this can be attributed to the fact that SVM can efficiently handle feature vectors of high dimensions and sparsity through its overfitting protection property.

Supervised SA models provide an accurate reliable performance, yet the labor-intensive task of preparing a sentimentally annotated corpus along with the training overhead and memory issues are important disadvantages that cannot be ignored. In contrast, lexicon-based SA models are easy to design since they do not need a labeled input data. Moreover, the training overhead is avoided by using a sentiment lexicon that acts as a rule-based classifier. However, one major drawback of these models is that they are not aware of language subtleties such as sarcasm, negations, etc. Because uniform weight scheme lexicons ignore the contextual-related information since a sentence's polarity is recognized by the polarity scores of its constituent words. To enhance the SA performance of these methods, new weighting schemes were developed based on the word's co-occurrence information (context) as in El-Beltagy and Ali (2013) and Assiri *et al.* (2017). On the other hand, the used lexicons also suffer from low-coverage and Out Of Vocabulary (OOV) issues especially for dialects which degrades the classification performance remarkably. Increasing the lexicon-coverage has been tackled in Abdulla *et al.* (2014) and Duwairi *et al.* (2015) through creating a large-sized lexicon initially constructed using seed words derived from publicly available lexicons or from the corpus itself then enriched with stems, synonyms and dialectal terms. Nevertheless, these solutions were insufficient to overcome the lexicon's dialect- and domain-dependency problems unless a very large-sized lexicon is built which is considered a difficult task.

To exploit the merits of the two previous methods, hybrid models have emerged. Lexicon-derived features are combined with linguistic ones to obtain a better sentiment classification performance. What makes these models better than both supervised and lexicon-based models is its ability to involve external semantic resources and datasets as in Salameh *et al.* (2015) and Baly *et al.* (2017) Furthermore, the wide variety of the hand-crafted features provide a coherent representation of the contextual information (Abdul-Mageed *et al.*, 2014; Al-Moslmi *et al.*, 2017).

The good performance of hybrid methods has been achieved at the cost of the laborious tasks of designing the features and building the lexicons. Deep learning-based methods alleviate such efforts through learning the features automatically from the data itself using deep neural networks. Text embedding features such as word/document embeddings generated via word2vec/doc2vec methods have proved their efficiency for SA when they were used to train ML classifiers (Altowayan and Tao, 2016; Mdhaffar *et al.*, 2017). Moreover, a better classification performance can be obtained if various architectures of deep neural networks, whose units adopt the compositional manner to represent the input text, are used to design the classifier as in Baniata and Park (2016), Al Sallab *et al.* (2015) and Al Sallab *et al.* (2017). It is obvious that DL methods are superior to traditional ML methods in terms of SA performance and features extraction cost. Nevertheless, SA using deep neural networks architectures requires more training time (Joulin *et al.*, 2016). This can be handled by applying an appropriate tuning of hyperparameters in addition to specific preprocessing and postprocessing procedures.

The previous comparison analysis of methods could be supported by tracking the classification performance of a specific dataset using traditional ML, rule-based and DL methods. For instance, given the Jordanian ArTwitter dataset, it could be noted that performing SA using a lexicon-based method (Abdulla *et al.*, 2014) resulted in a classification accuracy of 70%, while adopting distributed representations extracted via word2vec as in Altowayan and Tao (2016) has increased the accuracy by 11.88%. On the other hand, within the same category of methods such as the DL methods, it could be deduced that the effective handling of the special properties of the Arabic language has a positive impact on the SA performance as it can be seen in Al Sallab *et al.* (2015) and Al Sallab *et al.* (2017) where the accuracy improved from 74.3% using standard RAE to 86.5% when this model was equipped with Arabic-specific modifications.

The development of ASA models has involved the provision of annotated corpora (see Table 6), semantic resources and pretrained word vectors. This enriched the repository of NLP Arabic tools and resources. Regarding the research reviewed in this paper, most of the proposed datasets were of informal dialectal content as they were harvested from social media platforms. For single-dialect datasets such as Saudi (Assiri *et al.*, 2017), Kuwaiti (Salamah and Elkhlifi, 2016) or Jordanian (Duwairi *et al.*, 2014), they are rarely reused by other studies. However, pure Egyptian or

Paper	Dataset name/type	Size	Arabic Variant	Polarity	Publicly Available
(Shoukry and Rafea, 2012)	tweets	1,000	Egyptian	Pos/neg	No
(Al-Osaimi and Badruddin, 2014)	tweets	3,000	DA	pos/neg /neut	No
(Salamah and Elkhlifi, 2016)	tweets	340,000	Kuwaiti	pos/neg	No
(Abdul-Mageed *et al.*, 2014)	TGRD THR MONT DAR	3,015 3,008 3,097 2,798	MSA & DA	pos/neg /neut	No
(Salameh *et al.*, 2015)	Syr BBN	2,000 1,200	Syrian Levantine	pos/neg /neut	Yes Yes
(Al-Moslmi *et al.*, 2017)	reviews	8,861	DA	pos/neg	Yes
(El-Beltagy and Ali, 2013)	tweets comments	100 500	Egyptian	pos/neg /neut	No
(Abdulla *et al.*, 2014)	tweets comments	2,000 2,400	MSA/Jordanian MSA/DA	pos/neg pos/neg	Yes No
(Duwairi *et al.*, 2015)	tweets	4,400	MSA	pos/neg	No
(Assiri *et al.*, 2017)	tweets	4,700	Saudi	pos/neg	No
(Mdhaffar *et al.*, 2017)	TSAC	16,970	Tunisian	pos/neg	Yes

Table 6. *The datasets constructed and/or evaluated within the reviewed ASA studies.*

Egyptian-dominated datasets as El-Beltagy and Ali (2013), QCRI (Mourad and Darwish, 2013) and ASTD (Nabil *et al.*, 2015) have been used as a baseline in many other studies. This is due to the fact that Egyptian dialect forms the majority of the textual content on social media which makes it a preferable dialect to investigate by many research works. Yet, recent studies have shed light on other dialects spoken by the "Arab spring" countries such as Syrian and Tunisian (Salameh *et al.*, 2015; Mdhaffar *et al.*, 2017). Regarding MSA/multi-dialectal datasets such as Rushdi-Saleh *et al.* (2011), Aly and Atiya (2013), Abdulla *et al.* (2014), Abdul-Mageed *et al.* (2014) and Al-Moslmi *et al.* (2017), they are widely reused especially that modern ASA systems are designed with the objective of being dialect/domain-independent systems. On the other hand, the presented ASA systems have provided several Arabic sentiment lexicons. Most of which support MSA/multi-dialects such as lexicons in Abdulla *et al.* (2014), Abdul-Mageed *et al.* (2014) and Al-Moslmi *et al.* (2017) or Egyptian (El-Beltagy and Ali, 2013) which is publicly available.

Finally, some of the reviewed deep learning-based models produced Arabic word vectors trained either on MSA/multi-dialectal corpora as in Al-Rfou *et al.* (2013) and Altowayan and Tao (2016) or with Tunisian corpus (Mdhaffar *et al.*, 2017) or using an Egyptian corpus as in Zahran *et al.* (2015). According to Gridach *et al.* (2017), involving pretrained word vectors in a deep learning model enhances the quality of the model's embeddings vectors and thus improves further classification tasks. Based on that fact, the performance of Arabic deep learning-based SA models can be improved by exploiting pretrained Arabic word vectors trained on external corpora (Baniata and Park, 2016; Gridach *et al.*, 2017). This can be facilitated if the produced Arabic word vectors were made publicly available as those from Al-Rfou *et al.* (2013).

All the presented studies tried to address one or more challenging issues of ASA. Although valuable efforts were spent, there is a lot to do towards developing new tools and resources able to support MSA and DA more efficiently. For instance, Named Entities (NEs), especially person names, either in MSA or DA form a dilemma to any ASA system as they might considered as an adjective of a specific sentiment. Instead of excluding NEs (El-Beltagy and Ali, 2013; Duwairi *et al.*, 2014), they could be used as sentiment indicatives through assigning polarities to them. Thus, the polarity of a sentence could be predicted given the polarity of an NE contained in it. This idea is applicable for data collected during a short period of time in which opinions towards an NE are rather fixed. For instance, the location name "حلب" referring to Aleppo city in Syria has been often mentioned within negative contexts during December 2016 when Eastern Aleppo was under siege. Another difficult and interesting topic to investigate is SA of DA in terms of providing a universal system through which dialects' variances are ignored and common words between dialects along with their synonymous relations are considered. One possible way to perform that is by using word/phrase embeddings composed using syntactic-ignorant compositional functions and learned within a deep neural model.

7. References

Abdul-Mageed M., Subjectivity and sentiment analysis of Arabic as a morophologically-rich language, PhD thesis, Indiana University, 2015.

Abdul-Mageed M., Diab M., Korayem M., "Subjectivity and sentiment analysis of modern standard Arabic", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, p. 587-591, 2011.

Abdul-Mageed M., Diab M., Kübler S., "SAMAR: Subjectivity and sentiment analysis for Arabic social media", *Computer Speech & Language*, vol. 28, n° 1, p. 20-37, 2014.

Abdulla N. A., Ahmed N. A., Shehab M. A., Al-Ayyoub M., "Arabic sentiment analysis: Lexicon-based and corpus-based", *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, IEEE, p. 1-6, 2013.

Abdulla N., Mohammed S., Al-Ayyoub M., Al-Kabi M. *et al.*, "Automatic lexicon construction for arabic sentiment analysis", *2014 International Conference on Future Internet of Things and Cloud (FiCloud)*, IEEE, p. 547-552, 2014.

Al-Kabi M. N., Abdulla N. A., Al-Ayyoub M., "An analytical study of arabic sentiments: Maktoob case study", *8th International Conference for Internet Technology and Secured Transactions (ICITST)*, IEEE, p. 89-94, 2013.

Al-Moslmi T., Albared M., Al-Shabi A., Omar N., Abdullah S., "Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis", *Journal of Information Science*, 2017.

Al-Osaimi S., Badruddin K. M., “Role of Emotion icons in Sentiment classification of Arabic Tweets”, *Proceedings of the 6th international conference on management of emergent digital ecosystems*, ACM, p. 167-171, 2014.

Al-Rfou R., Perozzi B., Skiena S., “Polyglot: Distributed word representations for multilingual nlp”, *arXiv preprint arXiv:1307.1662*, 2013.

Al Sallab A. A., Baly R., Badaro G., Hajj H., El Hajj W., Shaban K. B., “Deep learning models for sentiment analysis in arabic”, *ANLP Workshop*, vol. 9, 2015.

Al Sallab A. A., Baly R., Hajj H., Shaban K. B., El-Hajj W., Badaro G., “AROMA: A Recursive Deep Learning Model for Opinion Mining in Arabic As a Low Resource Language”, *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 16, n^o 4, p. 25:1-25:20, 2017.

Al-Twairesh N., Al-Khalifa H. S., Alsalman A., “AraSenTi: Large-Scale Twitter-Specific Arabic Sentiment Lexicons”, *ACL (1)*, 2016.

Altowayan A. A., Tao L., “Word embeddings for Arabic sentiment analysis”, *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, p. 3820-3825, 2016.

Aly M. A., Atiya A. F., “LABR: A Large Scale Arabic Book Reviews Dataset.”, *ACL (2)*, p. 494-498, 2013.

Assiri A., Emam A., Al-Dossari H., “Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis”, *Journal of Information Science*, 2017.

Assiri A., Emam A., Aldossari H., “Arabic sentiment analysis: a survey”, *International Journal of Advanced Computer Science and Applications*, vol. 6, n^o 12, p. 75-85, 2015.

Badaro G., Baly R., Hajj H., Habash N., El-Hajj W., “A large scale Arabic sentiment lexicon for Arabic opinion mining”, *ANLP 2014*, 2014.

Balikas G., Amini M.-R., “TwiSE at SemEval-2016 Task 4: Twitter Sentiment Classification”, *arXiv preprint arXiv:1606.04351*, 2016.

Baly R., Badaro G., Hamdi A., Moukalled R., Aoun R., El-Khoury G., Al Sallab A. A., Hajj H., Habash N., Shaban K., El-Hajj W., “OMAM at SemEval-2017 Task 4: Evaluation of English State-of-the-Art Sentiment Analysis Models for Arabic and a New Topic-based Model”, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, Vancouver, Canada, p. 603-610, August, 2017.

Banea C., Mihalcea R., Wiebe J., “Multilingual subjectivity: Are more languages better?”, *Proceedings of the 23rd international conference on computational linguistics*, Association for Computational Linguistics, p. 28-36, 2010.

Baniata L. H., Park S.-B., “Sentence Representation Network for Arabic Sentiment Analysis”, *Proceedings of the Korean Information Science Society*, 2016.

Biltawi M., Etaiwi W., Tedmori S., Hudaib A., Awajan A., “Sentiment classification techniques for Arabic language: A survey”, *7th International Conference on Information and Communication Systems (ICICS)*, IEEE, p. 339-346, 2016.

Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P., “Natural language processing (almost) from scratch”, *Journal of Machine Learning Research*, vol. 12, p. 2493-2537, 2011.

Diab M., “Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking”, *2nd International Conference on Arabic Language Resources and Tools*, vol. 110, 2009.

Duwairi R., Ahmed N. A., Al-Rifai S. Y., "Detecting sentiment embedded in Arabic social media–a lexicon-based approach", *Journal of Intelligent & Fuzzy Systems*, vol. 29, nᴼ 1, p. 107-117, 2015.

Duwairi R., Marji R., Sha'ban N., Rushaidat S., "Sentiment analysis in arabic tweets", *5th international conference on Information and communication systems (icics)*, IEEE, p. 1-6, 2014.

El-Beltagy S. R., Ali A., "Open issues in the sentiment analysis of Arabic social media: A case study", *9th international conference onInnovations in information technology (iit)*, IEEE, p. 215-220, 2013.

El-Beltagy S. R., El kalamawy M., Soliman A. B., "NileTMRG at SemEval-2017 Task 4: Arabic Sentiment Analysis", *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, Vancouver, Canada, p. 790-795, August, 2017.

Farra N., McKeown K., "SMARTies: Sentiment Models for Arabic Target entities", *arXiv preprint arXiv:1701.03434*, 2017.

Farra N., McKeown K., Habash N., "Annotating targets of opinions in arabic using crowdsourcing", *Proceedings of the Second Workshop on Arabic Natural Language Processing*, p. 89-98, 2015.

Gomez L., Patel Y., Rusiñol M., Karatzas D., Jawahar C., "Self-supervised learning of visual features through embedding images into text topic spaces", *arXiv preprint arXiv:1705.08631*, 2017.

Gridach M., Haddad H., Mulki H., "Empirical Evaluation of Word Representations on Arabic Sentiment Analysis", *International Conference on Arabic Language Processing*, Springer, p. 147-158, 2017.

Habash N. Y., "Introduction to Arabic natural language processing", *Synthesis Lectures on Human Language Technologies*, vol. 3, nᴼ 1, p. 1-187, 2010.

Joulin A., Grave E., Bojanowski P., Mikolov T., "Bag of tricks for efficient text classification", *arXiv preprint arXiv:1607.01759*, 2016.

Kiritchenko S., Zhu X., Mohammad S. M., "Sentiment analysis of short informal texts", *Journal of Artificial Intelligence Research*, vol. 50, p. 723-762, 2014.

Le Q., Mikolov T., "Distributed representations of sentences and documents", *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, p. 1188-1196, 2014.

Liu B., "Sentiment analysis and opinion mining", *Synthesis lectures on human language technologies*, vol. 5, nᴼ 1, p. 1-167, 2012.

Maamouri M., Graff D., Bouziri B., Krouna S., Bies A., Kulick S., "Standard Arabic morphological analyzer (SAMA) version 3.1", *Linguistic Data Consortium, Catalog No.: LDC2010L01*, 2010.

Mdhaffar S., Bougares F., Esteve Y., Hadrich-Belguith L., "Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments", *WANLP 2017 (*co-located with *EACL 2017)*, 2017.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., "Distributed representations of words and phrases and their compositionality", *Advances in neural information processing systems*, p. 3111-3119, 2013.

Mohammad S. M., Kiritchenko S., Zhu X., “NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets”, *arXiv preprint arXiv:1308.6242*, 2013.

Mohammad S. M., Salameh M., Kiritchenko S., “How Translation Alters Sentiment.”, *J. Artif. Intell. Res. (JAIR)*, vol. 55, p. 95-130, 2016.

Mourad A., Darwish K., “Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs”, *WASSA@ NAACL-HLT*, p. 55-64, 2013.

Nabil M., Aly M. A., Atiya A. F., “ASTD: Arabic Sentiment Tweets Dataset”, *EMNLP*, p. 2515-2519, 2015.

Pasha A., Al-Badrashiny M., Diab M. T., El Kholy A., Eskander R., Habash N., Pooleery M., Rambow O., Roth R., “MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic”, *LREC*, vol. 14, p. 1094-1101, 2014.

Pennington J., Socher R., Manning C. D., “Glove: Global vectors for word representation”, *EMNLP*, vol. 14, p. 1532-1543, 2014.

Piryani R., Madhavi D., Singh V. K., “Analytical mapping of opinion mining and sentiment analysis research during 2000–2015”, *Information Processing & Management*, vol. 53, n^o 1, p. 122-150, 2017.

Popescu A.-M., Etzioni O., “Extracting product features and opinions from reviews”, *Natural language processing and text mining*, Springer, p. 9-28, 2007.

Refaee E., Rieser V., “An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis”, *LREC*, p. 2268-2273, 2014.

Reyes A., Rosso P., “On the difficulty of automatically detecting irony: beyond a simple case of negation”, *Knowledge and Information Systems*, vol. 40, n^o 3, p. 595-614, 2014.

Rosenthal S., Farra N., Nakov P., “SemEval-2017 Task 4: Sentiment Analysis in Twitter”, *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval ’17, Association for Computational Linguistics, Vancouver, Canada, August, 2017.

Rushdi-Saleh M., Martín-Valdivia M. T., Ureña-López L. A., Perea-Ortega J. M., “OCA: Opinion corpus for Arabic”, *Journal of the Association for Information Science and Technology*, vol. 62, n^o 10, p. 2045-2054, 2011.

Salamah J. B., Elkhlifi A., “Microblogging opinion mining approach for kuwaiti dialect”, *Computing Technology and Information Management*, vol. 1, n^o 1, p. 9, 2016.

Salameh M., Mohammad S., Kiritchenko S., “Sentiment after Translation: A Case-Study on Arabic Social Media Posts”, *HLT-NAACL*, p. 767-777, 2015.

Shoukry A., Rafea A., “Sentence-level Arabic sentiment analysis”, *2012 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE, p. 546-550, 2012.

Turney P. D., “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews”, *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 417-424, 2002.

Zaghouani W., Habash N., Mohit B., The Qatar Arabic language bank guidelines, Technical report, Technical Report CMU-CS-QTR-124, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, September, 2014.

Zahran M. A., Magooda A., Mahgoub A. Y., Raafat H. M., Rashwan M., Atyia A., “Word Representations in Vector Space and their Applications for Arabic”, *CICLing (1)*, p. 430-443, 2015.