Sentiment Analysis for Arabic Social Media

Manal Essam, Mohamed El-Menshawy and Hamdy M. Mousa
Department of Computer Science,
Faculty of Computers and Information,
Menoufia University, Egypt
manal.esam@ci.menofia.edu.eg, mohamed.elmenshawy@ci.menofia.edu.eg, hamdy.mousa@ci.menfia.edu.eg

***Abstract*—With the rapid spread of social media networks, such as Twitter and Facebook, in Arabic societies, it leads to the explosive growth of Arabic posts, reviews, comments, or tweets. Each one of these generates a huge volume of digital opinionated data on different topics such as politics, economics, societies marketing, and businesses. Analyzing valuable subjective information from opinionated data would assist in a better understanding, making decisions, and predicting global issues and events. Therefore, sentiment analysis coincides with social media networks and has become the most interesting research field in performing the analysis process and detecting sentiment polarities to extracted opinionated social-media data. However, there are several challenges faced the sentiment analysis process, especially with Arabic social data. Sentiment analysis of Arabic social media is indeed in its infantile stage and it has not obtained thoroughly attention wherein several challenging issues still need to address. Some of these challenges result from the complexity of Arabic natural language (e.g., complex morphological and lack of lexicon lists and datasets) and other challenges result from social media platform itself (e.g., slang words and colloquial terms). In this manuscript, we first study the impact of social media challenges on the current challenges of Arabic natural language. Our findings show that such challenges add more complexities to the Arabic sentiment analysis process. Based on these findings, we embark to review the emerged and contributed proposals, which give rise on analyzing opinionated data, extracted from Arabic social media networks. Our review methodology is based on a set of criteria, which we propose to assess and highlight the advantages and limitations of these proposals. The interesting point here is to help researchers identify the social sentiment analysis problems along with a comprehensive survey on the sentiment analysis levels (document, sentence, and aspect levels) and classification approaches (supervised, unsupervised, and hybrid approaches). Finally, we compare these contributed proposals in terms of the average accuracy and suggest a new hybrid approach based on our findings.**

Keywords*—*Arabic social media, Aspect level, Sentiment analysis, Twitter sentiment analysis

I. Introduction

Opinions represent human psychology beliefs and feelings and are fundamental influencers of their behaviors. Our beliefs and views are conditioned considerably by how other humans see and perceive the real world [1]. Whenever we require making a decision, we frequently seek out opinions of others. It is not only true for individual persons but also for organizations. For instance, business organizations need to discover consumer opinions about their offered products and services. Meanwhile, individual consumers need to know the opinions of users that have purchased a certain product to decide whether to advance in purchase the product or not. In the past, if individual persons need opinions, they directly ask their friends and family. If organizations need to find out consumer opinions, they carry out text surveys [2] through private groups, shaping their local media wherein users are typically consumers. In other terms, information merely flows from the publisher to the users. However, conducting such surveys is a hard task or an impossible task. Social media, defined as Web-based and/or mobile-based applications, extend this local media model by allowing each user to be a potential producer and consumer at the same time [3]. Any user can then interact with other users by commenting, sharing content, or expressing positive appraisal through the 'like' button (referred to as thumbs up).

Under the umbrella of social media term, there are a lot of microblogging services and platforms such as Facebook and Twitter, which are driven via user-generated contents (UGC). The content user produces within its timeline is consumed and driven by its connections (or followers). Social media can be used for different purposes [4]:

- Remaining in touch with friends and family, for example, through Facebook.
- Microblogging with catching up the latest news, for example, through Twitter.
- Remaining in touch with a professional network, for example, through LinkedIn.
- Sharing multimedia content, for example, through Instagram, YouTube, and Vimeo.
- Finding out answers to questions, for example, through Stack Overflow, Stack Exchange, and Quora.
- Finding out items of interest, for example, through Pinterest.

Facebook reported approximately 1.5 billion active users monthly in the second quarter of 2015. Twitter reported a volume of 500+ million tweets every day in 2013. On a smaller scale, Stack Overflow declared that more than 10 million programming questions asked on their platform since the website opened in 2015 [4]. The numbers show how the popularity of such social media has exponentially grown with more users that share more and more opinions, feelings, and information towards any matter of interest via different platforms. Since social media is a vital tool, Arab societies essentially depend on social networking media especially Facebook and Twitter to express opinions and ask advice from others about daily life issues. Numerically, 16 million Egyptian users use the Facebook platform. This number represents 1.4% of global Facebook users. Egypt has rated the first country among the Arab world that uses Facebook and seventeenth over the world regarding the audience size [5]. Around 317 million active users use Twitter in the Arab world [6], producing nearly 849 million tweets by March 2016. Compared to the

Arab world, the largest number of active Twitter users is in Saudi Arabia with more than 2.6 million of all users in the region [7]. Called Arab Spring is epitomized the significance of Twitter.

From the application perspective, we naturally desire to mine and analyze this wealth of Arabic social data to, for example, gauge public opinion and recognize current trends [7]. This is the automated task of sentiment analysis (also called opinion mining). Specifically, sentiment analysis (SA) is a multidisciplinary field of study, which intends to extract opinions from a natural language text by making use of computational methods [1]. However, although there are many users of Arabic social media, a few proposals have been put forward to study the SA of Arabic social media. El-Beltagy et al highlighted the major challenges that face the sentiment analysis task of Arabic social media [8]: (1) inaccessibility of dialectal Arabic parser; (2) lack of sentiment lexicons; (3) the need for the named entity recognition; and (4) handling compound phrases and idioms. Based on the Al-Twairesh et al's survey [9], El-Beltagy et al.'s challenges are only related to the complexities of processing the Arabic natural language itself without considering the challenges of the social media platforms. For instance, in the Twitter platform, a tweet should be no more than 140 characters, which enforces users to use abbreviations and slang [10]. Tweets could often contain spelling mistakes and informal grammars, URLs, and emoticons. Most posts on social media are written in Arabic dialects [7]. In addition to using Arabic dialects, the Facebook platform allows making large comments that could mention many entities and the conclusion with a positive sentiment for each comment is not enough.

Our motivation is to investigate the challenges resulted from Arabic sentiment analysis and the challenges produced from Arabic social media platforms to show the impact of social media challenges on the process of Arabic sentiment analysis. Based on our investigation, we present a set of evaluation criteria. By using these criteria, we review and evaluate the emerged and contributed proposals to analyze opinionated data extracted from Arabic social media networks, although these proposals are very few compared to current proposals of sentiment analysis of the English language. The interesting point here is to help researchers identify the sentiment analysis problems with rich information about the sentiment levels (document level, sentence level, and aspect level) and classification approaches (supervised, unsupervised, and hybrid). Finally, we compare these contributed proposals in terms of the average accuracy, computed from the reported accuracies and suggest a new hybrid approach based on our findings.

The organization of the manuscript is as follows. InII SectionII II, we present background material on Arabic language and sentiment analysis aspects to help understand the remaining sections. In Section IIIIII, we present the sentiment analysis process of Arabic social media, Arabic social media challenges, the impact of Arabic social media challenges, and a set of evaluation criteria. In Section IV, we review the current proposals and evaluate them using our criteria. In Section V, we discuss our investigation of Arabic sentiment analysis of Arabic social media and compare between reviewed proposals. In Section 0VI, we conclude the paper and suggest future work directions.

II. Background on Arabic Language and Sentiment Analysis

A. *Arabic Language Challenges*

Arabic is one of the six official languages of the United Nations and one of the Semitic languages. It is the official language of 27 countries and is spoken by more than 500 million people in the Arab world [7]. In addition, it is the official language of 1.4 billion Muslims and has been spoken for 2000 years. On the web, Arabic has ranked the fourth most used language and the fastest growing language during the last five years. "The growing rate of Arabic social media users, that will be reached 226 million by 2018, agreeing to the Arab Knowledge Economy Report 2015-2016 [11]". From this qualitative report, Internet usage rates will rise to 55% by 2018 compared with 37.5% in 2014 and about 7% higher than the expected global growth rate of 3.6 billion users.

The Arabic language has classical and informal forms. The classical language is taken from Qur'an and has phonological, morphological, lexical, and syntactically. The formal and accepted grammar rules of the Modern Standard Arabic are driven from the classical language and this is why it is termed formal language and used in education, news, and books. The spoken language or informal language used in daily life differs to some extent from Arabic country to other countries and produces many Arabic dialects. The informal language is the most used form in the current social media [12] [13] as the result of the free writing rules. There are some challenging issues facing researchers work on Arabic sentiment analysis. We summarize these challenges from [9] as follows:

- Complex morphological language (MOL): the Arabic words reveal several forms: derivational, inflectional and citicization where one lemma can get hundreds or thousands of surface forms with several affixes, clitics, and diacritics. Therefore, it is hard to extract or identify the root of Arabic words rightly.
- Dialogistic language (DIL): the Arabic language is a case of diglossia where the standard language (formal language) utilized in writing differs from the one utilized in daily life radically. The writing form is called Modern Standard Arabic (MSA) and the daily life form is called Dialectical Arabic (DA), colloquial and informal language. The two forms have different spelling rules and pronunciations. For instance, the word "جميل" is pronounced "jamil" in the MSA Iraq, "gamıl" the MSA Egypt, "žamıl" in the MSA Levantine, and "yamıl" in the MSA Gulf [14].
- Lack of sentiment lexicons (LOL): the recent efforts to form Arabic lexicon lists are very small compared to the English lists due to the complex nature of Arabic morphology. The recent studies covered a small number of dialects such as Egyptian and Levantine [15] beside the formal language. Also, there are some studies to translate English lexicons to Arabic, but these studies suffer from poor coverage due to the different meaning of translated words [16] [9].

- Lack of corpora and datasets (LOD): large annotated corpora are still a scarce resource for the Arabic language compared to the available English datasets. Most of the Arabic datasets are in-house and the available datasets usually derived from news or movie reviews [17] [18].
- Named entity recognition (NER): most Arabic names are derived from positive adjectives, such as the name” سعيد” corresponds to the adjective “happy”. Although NER is not required to detect sentiment [19], it is demanded in the case of Arabic sentiment analysis to distinguish between entity names and sentiment words [13] [9].
- The phrases and idioms (PAI): Arabic speakers usually use popular compound idioms and phrases to convey their feelings. The challenge here is that these terms hold implicit sentiments that cannot be observed by a sentiment system if these terms were not found in the lexicon list or have been contained in the experimental dataset. For example, the idiom “ لا يا شيخ” that is used to express disagreement in someone’s speech and the idiom “ما شاء الله” that is used to express loving things or persons.
- ARABZI (Arabish, Arabzi, or Romanized Arabic): this refers to a way of writing Arabic language using foreign or Latin characters, such as “وانتى طيبه” can be written as “w enty taiba”. Social media users commonly use it in [13] to make a switch between English and Arabic in their posts, causing it difficult to refer if a written word is English or Arabish. The recent proposals for Arabic sentiment analysis have not considered this challenge [9].
- Comparative opinions (COO): the opinion expressions can be stated as a comparison between two objects. The comparative expressions are different from direct expressions semantically and syntactically. Therefore, analyzing comparative opinions is deemed a challenge in English sentiment analysis [2]. In addition, there is only one proposal for analyzing Arabic comparative opinions [20].
- Scrasm detection (SCD): scrasm means that someone saying a positive opinion, but it means negative or vice versa [1]. Scrasm detection is a tedious task in English sentiment analysis; there are a small number of efforts. For Arabic sentiment analysis, no proposal was evaluated.
- Opinion spam detection (OSD): opinion spam means the chance of opinionated words being fake or untruthful to mislead the reader. Spam opinionated words usually observed with the business domain to encourage or damage a product. The Arabic spam detection task is understudied [21].
- Opinion target and opinion holder extraction (shortly, OTE), the primary task of SA is to extract the sentiment polarity of text, but the opinion target and opinion holder extraction is a necessary task for some domains and applications [22] [1].

B) *Sentiment Analysis*

Sentiment Analysis (SA) is the field of studying and analyzing opinions, sentiments, appraisals, attitudes, and emotions of human society towards entities like products, issues, events services, organizations and their attributes (aspects) [2]. In addition, sentiment analysis is an application of the natural language processing (NLP), data analytics and computational linguistic to automate the identification or classification of sentiment from opinionated text [23]. NLP is the scientific field for studying human languages from a computational perspective [24].

1) *Sentiment Analysis Process of Arabic Language*

The sentiment analysis is a pipeline process that incorporates four activities. We used the UML activity diagram to illustrate these activities and their association in Fig. 1.

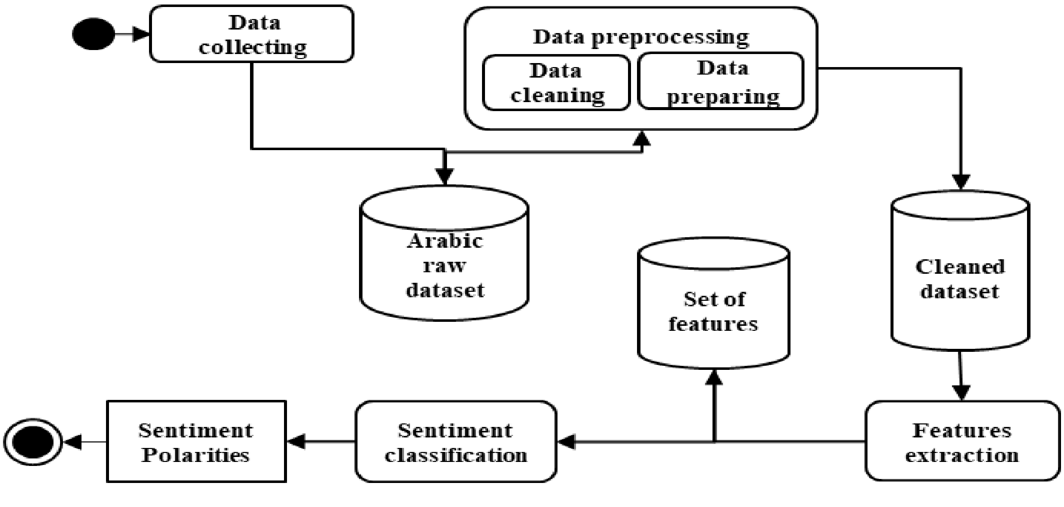

Fig. 1. The UML activity diagram of the sentiment analysis process

- The data collecting activity is the process of crawling or gathering data according to a specific topic from different resources via manual or automatic software tools. These resources might include existing corpus, databases, or lexicons and internet websites, online portals, and news web pages. The output of this activity is a dataset, normally called raw data.

- In second activity (data preprocessing), the raw data is first cleaned and then prepared using a set of NLP methods. Such methods do not intend to change the meaning of data text. They mainly identify the problems resulted from the data collection activity and modify the raw data to avoid interfering these problems during the data analysis activities. These methods include tokenization, lemmatization, stemming, and normalization. For example, the tokenization method divides sentences into words, while the lemmatization method converts words into their original roots. The stemming method converts words into their surface form and the normalization method converts different words and letters to a standard form. The output of this activity is a cleaned dataset.
- The feature extraction is the process of getting a representation of the cleaned data in the form of a set of most frequent and important features with aiming to reduce the size of data. These features usually include n-gram words (e.g., unigram/bigram and trigram), term frequency, term position, sentiment words and phrases, and part-of-speech. The n-gram features are the most widely used features in the literature of sentiment analysis.
- The sentiment classification activity is the process of classifying the extracted features into their sentiment polarities (positive and negative) by using machine learning (ML) approaches. These approaches are divided into three categories: supervised, unsupervised, and hybrid. The supervised approach (also called the corpus-based approach) depends on a set of supervised methods such as Support Vector Machine (SVM) and naïve Bayes (NB). The unsupervised approach (also called lexicon-based approach) is based on lexicons (i.e., a set of opinionated words along with their sentiment polarities) and the hybrid approach, as its name means, is a combination of two or more methods from supervised and unsupervised approaches. As we expected, the output of this activity is a set of sentiment and performance results including only positive polarity and negative polarity of features and accuracy metrics.

To illustrate the functionality of the sentiment analysis activities, we consider the following collected raw Arabic text, seen as a whole document:

أداء الممثلين كان مبدعا. من قامت بدور ريم ابدعت جدا وتفوقت، ومن قامت بدور نور ابدعت ايضا، ولكن لوحظ عليها التكلف لكنه بسيط جدا إذا ما قورن بعمرها

TABLE I illustrates the Arabic Sentiment Analysis Process. In this example, the SVM method needs a set of labels along with the set of features as input. In addition, the classifier returns a sentiment positive polarity for the whole of the document.

2) Sentiment Classification Approaches

Sentiment analysis is fundamentally a text classification issue. The results of the sentiment classification activity are generally articulated as a two-class classification (positive polarity and negative polarity) [1]. In addition, some approaches consider other class polarities such as neutral and mixed [25]. The sentiment classification is performed using three approaches: supervised, unsupervised, and hybrid.

The supervised approach uses generally a set of training data and their corresponding labels to be able to learn and later predict. In the context of sentiment analysis, the supervised approach uses corpora of relevant features as training data and their annotations (labels) to train and validate its classifiers [26]. Getting Arabic annotated data is often unavailable or needs a lot of efforts and time [24]. The methods that can predict the results from training data are, for example, SVM, NB, Artificial Neural Networks (ANN), and K-Nearest Neighbor (K-NN). The features are used with this approach [27] include:

TABLE I. ARABIC SENTIMENT ANALYSIS PROCESS

The data preprocessing activity	
Tokenization	“أداء” “الممثلين” “كان” “مبدعا” “من” “قامت”“” بدور” “ريم” “ابدعت” “جدا ““وتفوقت ““ومن ““قامت ““بدور” “نور” “ابدعت” “ايضا” “ولكن ““لوحظ ““عليها” “التكلف” “لكنه ““بسيط ““جدا” “إذا” “ما” “قورن ““بعمرها”
Stop words removal	أداء” “الممثلين” “مبدعا” “قامت” “بدور” “ريم” “ابدعت” “جدا ““وتفوقت ““قامت““بدور” “نور” “ابدعت” “ايضا” “لوحظ ““التكلف” “بسيط “جدا” “قورن ““بعمرها”
Stemming	أداء” “ممثلين” “مبدع” “قام” “دور” “ريم ““أبدع ““جدا” “تفوق” “قام” “دور” “نور” “أبدع” “لوحظ التكلف” “بسيط” “جد” “قورن” “عمر”
The feature extraction activity	
Unigram	“أداء”“ ممثلين” “مبدع” “قام” “دور” “ريم ““أبدع ““جدا” “تفوق” “قام” “دور” “نور” “أبدع” “لوحظ التكلف” “بسيط” “جد” “قورن” “عمر”
The sentiment classification	
SVM	Sentiment positive polarity

- Terms and their frequency: these terms are individual words (unigram) and their frequency counts (n-gram). One of the most common weighting schemes used is term frequency-inverse document frequency (TF-IDF).
- Part of speech (POS): the POS of words is remarkable indicators. For example, the adjectives are a crucial indicator of sentiment (opinion) (e.g., happy, sad, etc.). Therefore, they are treated as a special feature.

- Sentiment words and phrases: sentiment words directly carry positive or negative sentiment polarity such as good, honest and beautiful are positive opinionated words and, bad, dirty and horrible are negative opinionated words. Most opinion words are adjectives and adverbs; however, nouns and verbs can express opinions (e.g., garbage and love).
- Sentiment shifters: the most important sentiment shifters are negation words, which change sentiment polarity from positive to negative or vice versa (e.g., I do not hate this camera).

The unsupervised approach is based on the opinionated words with their positive and negative sentiment polarity for classification. The opinionated or sentiment words with their polarities are called lexicons. In addition to the sentiment polarity, some lexicon lists add the weight of the opinionated words to determine the strength of its class. The lexicon list could be created from existing dictionaries or from a corpus. In the context of sentiment analysis approach, the unsupervised model does not require labeled datasets but it needs a wide-coverage lexicon that covers a maximum number of sentiment words with their semantic orientation to determine an overall sentiment polarity of text. This classifier identifies the score of its opinionated words and calculates the sum of them to give an overall semantic orientation of the text. There are popular algorithms used with this classifier to calculate the sentiment polarity of opinionated words, such as Point Mutual Information (PMI), an aggregation function, K-NN, etc. [2].

The hybrid approach, as its name means, is a combination of two or more methods from supervised and unsupervised approaches. Initially, the lexicon-based classifier is used to label unannotated data using a list of lexicon words. Therefore, the approach saves the time-consuming in labeling data that often prepared manually. The output data from lexical classifier will be used as a training dataset for the supervised approach. As the result, the hybrid approach benefits from the advantages of two approaches (the accuracy of supervised approach and the saving time and effort in unsupervised approach).

3) *Sentiment Analysis Levels*

The sentiment analysis process is evaluated to classify the opinionated text into positive, negative or neutral sentiment polarity about a specific object (e.g., topic, product, or person). This process is considered at three levels [9] [24] [1]: document, sentence, and aspect.

Document-level: the main task of this level is to classify a whole text as positive, negative or neutral sentiment polarity. It assumes that a whole document conveys specific sentiment polarity about a single entity. For example, given an object tweet, the system decides whether the tweet holds a positive or negative sentiment about the object. Thus, it is not ideal for documents that hold or compare more than a single entity.

Sentence-level: This level is divided into two tasks. The first task is subjectivity classification, to distinguish or classify the objective sentences from the subjective sentences. The second task is sentiment classification, to classify the subjective sentences into positive, negative and neutral sentiment polarity. In addition, we should note that subjectivity sentence is not conveyed sentiment direction all time and the objective sentences can detect sentiment direction. Therefore, the sentence level is not enough. For example, “I bought the camera last week and the battery has damaged”.

Aspect-level: together the document and the sentence level do not illustrate what people exactly like because they do not identify or extract opinion targets (entities) and the sentiment of these entities. Therefore, analyzing opinion texts at the document level or the sentence level is often deficient for most domains and applications. For extracting such details, we go to the aspect level. It implements finer-grained analysis, instead of seeing language structure (documents, sentences, and clauses), aspect level straightly deals with opinionated words. The core idea at this level is considering an opinion comprises of sentiment (positive and negative) and target of opinion (entity). For example, although the camera is wonderful, her battery does not work well. This sentence has positive tone but we cannot consider the entire sentence as a positive. To best our knowledge there is limited research on Aspect-based sentiment analysis for Arabic text in general. This can be explained due to the lack of available datasets prepared for aspect-based sentiment analysis level and to the slow progress in sentiment analysis of Arabic text research in general [22] [16].

III. Sentiment Analysis of Arabic Social Media

Social media is a web-based and mobile-based Internet application that give the users the chance to create access and change their private views that is ubiquitously accessible [3]. In addition to social networking media (e.g., Twitter and Facebook), we will also use ‘social media’ for other blogs such as really simple syndication (RSS) feeds, blogs, wikis, and news, all contain unstructured text and accessible through the web. Social media is an important resource for SA research field because social media content is most resource contain volume, variety and have most dynamic base of human activity, giving good opportunities to analyzing individuals, groups and society views (opinions) [7]. Especially Facebook and Twitter are considered the most active communication tools [6] because users share every daily life activity freely, easily and fast. Scientist's society and industry are growingly recovering new directions for gathering analyzing this wealth of opinionated data automatically from social networking tools. Social media data, in particular, twitter posts (tweets), gives more challenges than structured data, taken from newspapers, books, and scientific journals [7].

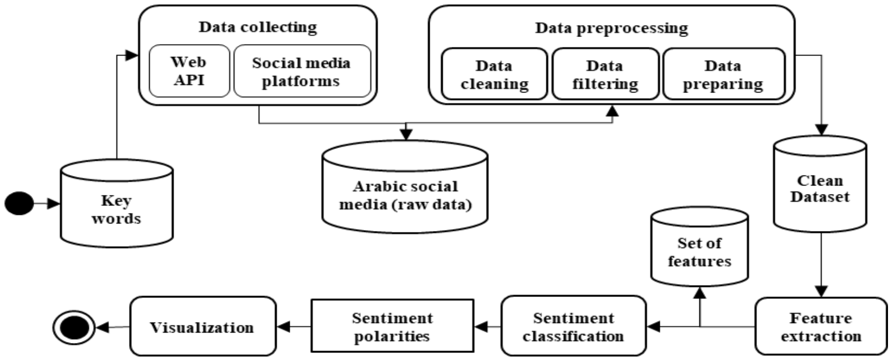

Fig. 2. The UML activity diagram of the sentiment analysis process of Arabic social media

A) *Sentiment Analysis Process of Arabic Social Media*

The sentiment analysis process of Arabic language discussed Section B)1) cannot be applied directly to analyze the sentiment analysis of Arabic data extracted from social media platforms. Initially, the data collecting activity in Fig. 1 needs to extend by using keywords or seed words to crawl or gather the opinionated data from social media platforms (see Fig. 2). In Fig. 2, social media platforms assign Web APIs to crawl data to only authorized users. For example, the Twitter platform has several API tools, including Twitter API, stream API, and tweet crawler to retrieve tweet contents. To use these tools, the developers and researchers need to obtain (consumer key, consumer secret, access token, access token secret) credential parameters from the Twitter developer webpage.

Secondly, other activities in Fig. 2 need to extend as well to be able to address the challenges resulted from Arabic social data extracted in the previous activity. Moreover, we need to incorporate a new activity, called visualization activity. The description of these extended activities and new activity is as follows (see Fig. 2):

- In the second activity (data preprocessing), the social media raw data is first cleaned, filtered and then prepared using a set of NLP methods. Social media data is a complex unstructured text, due to the informal and free writing text format provided by the employed platform where each post contains sentimental data (or opinionated text) along with non-sentimental data such as URLs, images, hashtags, mentions, videos, misspelling text, and duplicated words and letters. After cleaning non-sentimental data, we filter only sentimental text because non-sentimental data is considered the irrelevant text. The NLP methods applied in the cleaning and preparing text include tokenization, lemmatization, stemming, and normalization.
- The visualization is the process of visualizing the results of sentiment polarities related to the particular domain and illustrates the sentiment changes over time. Various graphic representation forms can evaluate this activity.

As we did before, we consider the following collected raw Arabic tweet to illustrate the sentiment analysis process of Arabic social media:
★ RT @Oman_Falcon: أبل قالت أنها هتستبدل مجانا 5 إيفون اللى اشتكى مستخدموه من وجود عيب تقني في زرار التشغيل بعد أن وقف الجهاز عن العمل هذا الشيء جميييل ما شاء الله حتى لا تظلمونا انا اوافك على هذا الشيء هيييح amazing w allahy …

TABLE II. Sentiment Analysis of Arabic Social Media Process

The data preprocessing activity	
• **Cleaning: we remove RT, mentions, punctuations, and white spaces.** • **Filtering: we remove non-Arabic words, and stop words.**	أبل قالت أنها هتستبدل مجانا آيفون5 االلى اشتكى مستخدموه من وجود عيب تقني في زرار التشغيل بعد أن وقف الجهاز عن العمل هذا الشئ جميييل ما شاء الله حتى لا تظلمونا انا اوافك على هذا الشيء هيييحamazing w allahy
	أبل قالت أنها هتستبدل مجانا آيفون االي اشتكى مستخدموه من وجود عيب تقني في زرار التشغيل بعد أن وقف الجهاز عن العمل هذا الشئ جميييل ما شاء الله حتى لا تظلمونا انا اوافك على هذا الشيء هيييح أبل تعلن هتستبدل مجانا آيفون اشتكى مستخدموه وجود عيب تقني زرار التشغيل وقف الجهاز العمل الشيء جميييل ما شاء الله لا تظلمونا اوافك الشيء هييح
• **Preparing: we make the normalization and tokenization processes.**	أبل تعلن هتستبدل مجانا ايفون اشتكى مستخدموه وجود عيب تقنى زرار التشغيل توقف الجهاز العمل الشيء جميل ما شاء الله لا تظلمونا اوافك الشيء هيح “ابل” “تعلن” “هتستبدل” “مجانا” “ايفون” “اشتكى” “مستخدموه” “وجود” “عيب” “تقنى “زرار” “التشغيل” “توقف” “الجهاز” “العمل” “الشيء” “جميل” “ما شاء” “الله” “لا ””تظلمونا” “اوافك” “الشيء” “هيح”
The feature extraction activity	
• **Feature extraction: our features are unigram features.**	وجود” “عيب” “تقنى “زر” “التشغيل” “توقف” “الجهاز” “ اشتكى” “مستخدموه” “ ”ابل” “تعلن” “تستبدل” “مجانا” “ايفون“ ”تظلمونا” “اوافك” “الشئ” “هيح“ “ ”ماشاء” “الله” “لا ””العمل” “الشئ” “جميل
The sentiment classification	
• **Classification: we use here SVM.**	Positive sentiment polarity

Table II demonstrates the sentiment analysis of Arabic social media process. From the previous example, we notice that the most dominant reason for lagging sentiment analysis of Arabic social media is the lacking of methods that can be used to process the Arabic text.

B) *Challenges of Arabic Social Media*

The challenges of the sentiment analysis process applied in Arabic social media, particularly social Twitter network platform result from informal and free writing text format producing unstructured text. The user-generated content does not then follow any cuffs, causing unlimited challenges. We summarize these challenges as follows [7]:

- Unstructured language (UNL).
- Orthographic mistakes (ORM).
- Slang words (SLW).
- Ironic sentences, abbreviation, and contractions (IRC).
- Dialectal (DIA).
- Idiomatic phrases (IDP).
- Spelling inconsistencies and connected words (SPI).
- Lack of capitalization (LOC).
- Repeat characters for disbelief or belief sentiment (REC).

The unstructured language (text) from social media as the result of using free writing features, for example, writing tweets is freely and openly without any constraints except the tweet length only 140 characters. Therefore, the text has different types of data URLs, mentions, images, and written text without separation. Social media posts contain slang words are produced by users to convey sentimental feelings in low letters, such as (هيييح) this word gives positive sentiment but not relate to any Arabic forms (MSA or DA) for analyzing the word root. The social media posts contain many orthographic mistakes because the user writes posts as they speak errors from keyboard typing and errors from the weakness of writing e.g., (اوفق-اوفك). Therefor, the morphological analysis of social media data is more complex. The spelling of Arabic words is inconsistency where Arabic words can derive from the same root, such as (يلعب-تلاعب). The two words are derived the same root (لعب). Therefore, the spelling inconsistency will affect the detection of sentiment polarity. Social media users usually use Arabic ironic and colloquial expressions, contractions, and abbreviations to convey their sentiments, such as (ماشاء الله– لا يارجل - هه). These terms hold different sentiment polarity. Lack of capital letters in Arabic effect on capture the features of words and sentences because the words are connected. The social media users write posts derive from the same root, such as (يلعب-تلاعب). The two words are derived the same root (لعب). Therefore, the spelling inconsistency will affect the detection of sentiment polarity. The social media users write posts with their native dialectical and with no diacritics. This causes complexities when analyzing sentiment polarity with various dialects. The users usually duplicate letters and words too strengthen their sentiment, hence removing these letters may affect the accuracy of the classification approaches that based on the weight of words. Social media users write their posts with Latin characters to code switch between the two languages. There are not efforts dealt with this challenge.

C) *Impact of Arabic language Challenges on social media challenges*

We start by combining the Arabic SA and social media challenges in an integrated framework, as shown in TABLE III. We then deeply investigate the impact of these challenges on each other using the one-to-one technique. For example, in the one-to-to technique, we study the impact of considering the UNL challenge with the MOL challenge. We found that the UNL challenge adds more complexity in the MOL analysis such as when we analysed the morphology of informal words, which required specific NLP capabilities such as normalization, stemming, etc. The available NLP tools for Arabic are evaluated for the formal language; therefore, the accuracy of preparing the informal text is low. We repeat this technique of the UNL challenge with other Arabic language challenges one-by-one. Another example that shows the impact of the ORM challenge on the MOL challenge. We found that the ORM challenge adds more complexity in the MOL analysis. For instance, when we analyze the morphology of the word (اطبع, print), we get two different roots (اطبع-طبع), but the situation is going worth in social media, as the user writes this word with orthographic mistakes as follows: for example, اتبع, followed its morphology is two roots (اتبع- تبع). Therefore, when we classify the latter two roots, they give sentiment polarities (negative or positive based on the context) that are different from the expected polarities (in our case, the polarity is neutral). As a result, the sentiment classification accuracy will be affected. In addition, when we consider the SLW challenge with the MOL challenged, we find that the slang words do not relate to MSA or any DA, such as the slang word هييح does not have any Arabic root, which in turn we cannot find its polarity. In this case, we need to build a specific lexicon list for these social media slang words. In addition, the social media users use the dialectical Arabic form rather than the MSA form, as the dialectical form is different from the MSA form in terms of the syntactical and lexical structures, and does not follow orthographic rules. As the result, the complexity of the MOL challenge is increased.

We repeat our one-to-one technique on the other challenges. Our findings of this technique are summarized in the following table III.

Where the symbol "+" means the challenge becomes more complex. In summary, on the one hand, the complexity of Arabic language is the main source of the most of Arabic social media challenges regarding the MOL, DIL, LOL, LOD, SPO, COO, and NER challenges. On the other hand, we found that the social media Twitter platform adds other challenges into the Arabic language challenges. We call the more complex challenged with unlimited challenges. Therefore, the called unlimited challenges make the sentiment analysis process, especially preprocessing activity more complex and give us an insensitive to define a set of criteria to review and compare current proposals.

TABLE III. A SUMMARY OF STUDYING THE IMPACT OF SOCIAL MEDIA CHALLENGES ON ARABIC LANGUAGE CHALLENGES

	MOL	DIL	LOL	LOD	NER	PAI	Arabizi	COO	SCD	OCD	OTE
UNL	+	+			+		+				
ORM	+				+						+
SLW	+		+								
SPI											
IRC									+	+	
DIA	+	+	+	+		+			+		
LOC					+						+
DIE	+		+								
REC	+										

D) Evaluation Criteria

In this section, we introduce our evaluation criteria with respect to studying in Section C). Indeed, we use these criteria to evaluate the reviewed proposals in Section IV.

- Language: we use the language criterion to be able to classify the current proposals according to the adopted language into proposals-based modern standard language and proposals-based dialectical language. This criterion will help determine the proposals that have been considered the dialectal language, commonly used in social media and implicitly highlight that the proposals-based MSA will not recommend to interesting readers as their approaches are mainly proposed for MSA, which need technical modifications or improvements to apply.
- Methods: the main motivation of this criterion is to classify the proposals resulting from applying the language criterion (i.e., proposals-based dialectical language) according to the learning methods into supervised, unsupervised, and hybrid. The reason is to capture the main benefits of the hybrid methods, which at the same time combine the advantages of both the supervised and unsupervised methods in one framework.
- Levels: this criterion is to classify the proposals resulting from applying the method criterion into three levels: document, sentence, and aspect. Intuitively, we aim to determine the proposals that consider the sentiment analysis process on the aspect level, to avoid the shortcomings of the document and sentence sentiment analysis.
- Data resources and platforms: introducing this criterion enables us to determine the sources of the employed datasets, incorporating the opinionated social data, to train and test the developed algorithms. Since the focus is on social media, we need to assess the capabilities and limitations of the famous and common platforms (e.g., Twitter and Facebook).
- Performance: this criterion enables us to evaluate the proposals according to the accepted performance. For evaluating the performance of the sentiment classification activity, the Error Matrix or Confusion Matrix can be used. Each column symbolizes predicted classifications and each row symbolizes realistically defined classifications. Based on this matrix, four indexes can be calculated to show the performance: Accuracy, Recall, Precision, and F-measure.
  - Accuracy is the percentage of the testing set of items that are classified correctly via the classifier. It calculates as follows where $\mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \mathrm{P}, \mathrm{N}$ refer to the number of true positive, true negative, false positive, positive, and negative items, respectively:

$$Acc = \frac{TP+TN}{\mathrm{P+N}} \quad (1)$$

  - Precision can be defined as a measure of exactness (i.e., what the percentage of items labeled as positive are sharply accurate). It can calculate as follows:

$$Pre = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (2)$$

  - Recall is a measure of completeness (i.e., what the percentage of positive items are labeled as such). It can calculate as follows:

$$Rec = \frac{TP}{TP+FN} = \frac{TP}{P} \quad (3)$$

  - F-measure gives the equal weight to the precision and recall. It can calculate as follows:

$$F = \frac{2x\ Pre\ x\ Rec}{Pre\ +Rec} \quad (4)$$

IV. METHODOLOGY AND LITERATURE REVIEW

Most of the research communities gave rise to apply the sentiment analysis into the English language [9] and cover the three sentiment analysis levels (document, sentence, and aspect) using different classification approaches (supervised, unsupervised, and hybrid). Such English communities have a huge volume of resources such as the dataset, available at http://dumps.wikimedia.org and the lexicon list available at http://SentiWordNet.com [3]. However, the Arabic sentiment analysis is flourished and not reaches mature. Moreover, the research direction of the sentiment analysis of Arabic social media is still in its infancy [24] [13].

In our research methodology, we use the following keywords “sentiment analysis survey”, “Arabic sentiment analysis challenges”, “social media”, and “Arabic language challenges” using the following databases: Google scholar, Springer, IEEE explorer to find related papers. We obtained 50 papers that cope with [24] [13], where is there are very few proposals in the literature. We then removed irrelevant papers after reviewing their abstract, methodology, and conclusion by three co-authors. Finally, our database consists of 23 papers: 3 survey papers and 20 conference and journal papers. According to the sentiment analysis process illustrated in Fig. 1, we classify the sentiment classification approaches into three main groups (supervised, unsupervised, and hybrid approach) and their distributions as shown in Fig. 3. This section continues by reviewing the contributed proposals related to sentiment analysis of Arabic social media. In our review, we present each proposal and use our evaluation criteria to identify the advantages and disadvantages of these proposals.

A) Supervised learning approaches

In this paper, the authors [7] investigated the challenges related to the sentiment analysis of Arabic social media (e.g., Arabic dialectical, slang expression and orthographic mistakes). They gave rise to highlight the challenges resulted from informal/dialectical Arabic Saudi language. The SA method is initially cleaning, normalizing, and then stemming Arabic tweets. The obtained tweets are manually annotated to generate the training corpus that contains 4000 annotated tweets. The syntactic features (or n-grams) are extracted from the corpus. These features (when n=1, 2, and 3) are fed the supervised support vector machine (SVM). The SVM classifier computes the sentiment polarity of Arabic tweets. Therefore, the sentiment level is of the kind the document level. The highest F-measure of the SVM classifier is 73% with n=1. The accuracy is reduced when the number of n-grams is increasing.

The authors in [18] developed an Arabic corpus, called OCA. This corpus is available for the scientific community to run the sentiment analysis process. The OCA corpus contains 500 movie reviews, collected from different web pages and blogs automatically using a simple bash script. In the preprocessing phase, the collected reviews are tokenizing, stemming, and stop words removed. Then, the N-gram and TF-IDF features are extracted from each review to train the SVM and NB classifiers. These classifiers, as we known, returns the sentiment polarity of each review. Therefore, the sentiment level is of the kind the document level. The performance of NB and SVM is measured using the accuracy, precision, and recall metrics with n-gram features and stemming and without stemming. The highest accuracy result is 95% with SVM and trigram features but without stemming reviews.

The authors in [28] proposed a technique to identify the ‘opinion aspect’ from Arabic tweets. It is clear that the sentiment level is the kind of aspect level. The technique is executed in three phases. Prior to start these phases, they collected 500 Arabic tweets in general topics using The Twitter API and manually annotated them. In the first phase: such collected tweets are processed to remove unwanted data in three steps. In the first step, they removed URLs, mentions, and RT from each tweet. In the second step: the leading and trailing spaces, the line breaks, and the leading and trailing non-alphabetic characters such as “,:, \, *, and - are removed. In the third step, they kept the Unicode of characters, removed repetition words, and made normalization. In the second phase: the features are extracted from cleaned tweets to be used in the sentiment classification phase. These features are POS tags of words, named entities, POS patterns of words tags, and Twitter specific features. In the third phase: three ML classifiers are trained using the extracted features to determine each word in the tweet is either an opinion target or not. These classifiers are SVM, NB, and K-NN. Two experiments are conducted with different features to evaluate the effectiveness of POS tags with/without POS patterns. The highest F-measure of: 1) K-NN is 91.5% with POS pattern features and 85.9% without POS pattern features, 2) NB is 85.4% with POS pattern features and 85.2% without POS pattern features, and 3) SVM is 69.7% with pattern features and 66.6% without pattern features.

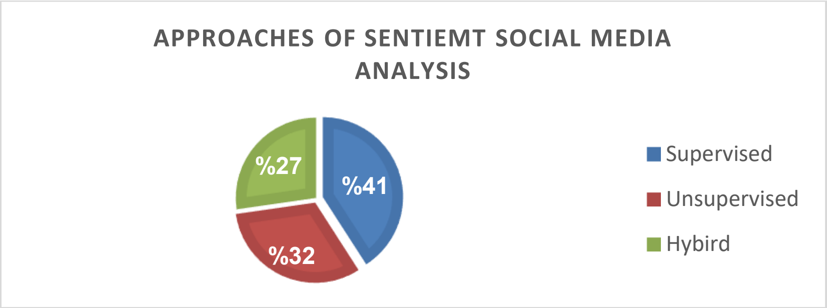

Fig. 3. Sentiment Analysis Approaches Distribution

In [29], the authors developed an Arabic sentiment analyzer for classifying comments of Arabic news pages on Facebook. Initially, the authors built Arabic corpus that consists of 2400 Egyptian dialectical comments, which manually annotated by a human expert. The sentiment analysis process includes three phases: preprocessing, feature extraction, and classification. In the preprocessing phase, the stop words and long comments are removed. In the feature extraction phase, two groups of features are extracted. The first feature group includes common words between posts and comments, the number of words in comments, and the number of comments in posts. The second feature considered all words in posts and comments and give each word value from four labels to union them. That resulted from splitting posts and comments. In addition, the negation and intensifier features are added to both feature groups to improve the performance. In the classification phase, three ML methods are applied to classify the extracted features into three sentiment polarities: supportive, attacked, and neutral. These classifiers are SVM, NB, and DT. The sentiment level is of the kind the document level. The highest precision metric is 73.4% with the SVM classifier.

In [30], the authors investigated the subjectivity-sentiment analysis task of MSA reviews based on the sentence level. The task consists of two phases. In the first phase, reviews are classified to subjective and objective. In the second phase, the subjective sentences are classified into positive and negative polarities. Before implementing this task, they constructed an Arabic review corpus with each review in the corpus is manually tagged by two college educated native speakers to one label from 4 possible labels: objective (obj), subjective positive (s-pos), (s-neg), and (s-neutral). The total number of reviews is 1281 obj sentences and 1574 subj sentences. The subj sentences are further classified into 491 s-pos sentences, 689 s-neg sentences, and 349 s-neutral sentences. In addition, they built an Arabic lexicon list of 3982 adjectives labeled with the pos, neg, and neutral sentiment polarities. In the preprocessing phase, each review in the corpus is tokenized into words and the words are segmented without clitics (or stemming). Then, the features are extracted to run the classifier. The extracted features are two groups: language independent (unique, n-gram, and adjectives) and morphological features (person, state, gender, etc.). The SVMlight method is used as a sentiment classification to perform the mentioned two tasks. The F-measure metric is used to measure the performance of the first phase in the subjectivity analysis task. The best result is 73.17% for the unigram and bigram features with the stemming preprocessing technique. For the second phase (sentiment analysis), the best result is 56.78% for the lemmatization preprocessing technique and the unigram features.

The authors in [25] investigated the different supervised ML approaches on their building corpus. They developed an Arabic Jordanian dataset consists of 1800 MSA and DA Arabic tweets. The tweets are cleaned by removing unwanted data like URLs, hashtags, mentions, repeated tweets, and foreign letters. Two human experts manually annotate the cleaned tweets to positive and negative labels. The corpus is used to extract n-gram features (unigram, bigram, and trigram) with full stemming, light stemming, and without stemming from tweets. In the sentiment classification, the SVM and NB supervised classifiers are trained to detect the sentiment polarity of tweets. As the result, the sentiment level is of the kind the document level. The accuracy and F-measure are used to measure the performance without removing the stop words. The results showed that: 1) SVM gives accuracy 88.72% and F-score 88.27% with full stemmer and bigram feature and TF-IDF weighting scheme, 2) NB gives 83.61% for accuracy and 84.73% for F-score with light stemmer and trigrams with a TF weighting scheme. Thus, the performance of SVM is better than the corresponding performance of NB.

The authors in [22] built a human annotated Arabic dataset, called HAAD. HAAD supported the aspect level of sentiment analysis and its related tasks such as aspect extraction and aspect categorization. The process of building HAAD is carried out in three steps: data collection, data annotation, and data annotation format. In the data collection step, data collected from Arabic reviews LABR dataset by seven groups of three graduated students. The collected data consists of 2838 aspect terms and 1296 out of them are distinct aspect terms. In the annotation step, the graduated students annotate the reviews according to four tasks: aspect extraction, aspect polarity, aspect category, and aspect category polarity. In the first two tasks (aspect extraction and polarity), the annotators annotate all single/multiple terms that refer to entity target and label each entity to positive, negative, conflict, or neutral labels. In the aspect category and polarity tasks, the annotators assign each review to the corresponding categories and label each category to positive, negative, conflict, or neutral. In the annotation format step, the dataset puts in XML format and makes available for research. To measure the performance of such tasks, the F-measure and accuracy methods have been applied. The results of their accuracy are 23.39%, 29.7064, 15.185%, and 42.5743%, respectively.

The authors in [21] proposed a new approach to mark spam opinions in Arabic reviews by integrating methods from data mining and text mining in one mining classification approach. The methodology of the approach is started by gathering data from online economic Arabic review websites such as www.booking.com, www.agoda.ae , and www.tripadvisor.com.eg. The collected reviews are integrated into a dataset, called TBA. The TBA dataset consists of 2848 records and each record is annotated to spam/non-spam labels by a human expert. In the preprocessing step, the TBA dataset cleaned from unnecessary data to remove non-Arabic text, the reviews are tokenized into words and stop words are removed. The words are then stemmed using the light stemming technique wherein tokens that have less than four characters are removed. In the classification step, the methods from data mining (review content feature and metadata) and text mining (hotel information features) are combined and oversampling technique is applied to overcome the data imbalance drawback. The SVM, NB, and ID3 classifiers are trained to detect spam and non-spam of each sentence review. Therefore, the sentence level is applied. This approach achieves the best result with the NB method, which gives 99.2% for accuracy.

In [31], the authors proposed an ML approach (called SAMAR) for SSA system for Arabic social media from different genres. They initially generated manually annotated Arabic dataset covering a variety of existing Arabic datasets (DARDASHA, TAGREED, TAHRIR, and MONTADA). For data preprocessing, the tokenization, lemmatization, and POS tagging were experimentally applied to select the preprocessing one of them with high accuracy. Also, they manually created a lexicon of 3982 adjectives labeled with positive, negative, and neutral. The system applied two-phase classification approach. In the first phase, the system separates between subjective and objective terms. In the second phase, the subjective terms are classified into positive and negative sentiment polarities. Therefore, the system is of kind sentence level. To train the classifier, the morphological features using the tokenization process, POS tagging features, standard features (unique and lexicon features), dialectical features, and genre-specific features are extracted. The SVM classifier is used in such two phases (subjectivity analysis and sentiment analysis) without reporting any performance metric.

In table IV, we first apply our evaluation criteria to assess the above discussed supervised learning approaches and second report results.

TABLE IV. THE SUMMARY OF ASSESSING THE SUPERVISED LEARNING APPROACHES AFTER APPLYING OUR EVALUATION CRITERIA

Ref.	**Dataset**	**Language**		**Sent. Level**			**Performance Metrics**			
		MSA	**DA**	**Doc.**	**Sen.**	**Asp.**	**Acc.**	**Pre.**	**Rec.**	**F-measure**
[7]	**Twitter**		✓	✓						✓
[18]	**Reviews**	✓		✓			✓			
[28]	**Twitter**	✓	✓			✓				✓
[29]	**Facebook**		✓	✓				✓		
[30]	**Reviews**	✓	✓		✓		✓			✓
[25]	**Twitter**	✓	✓	✓			✓			
[22]	**LABR dataset**	✓				✓	✓			✓
[21]	**Reviews**	✓		✓						✓
[31]	**Using social media datasets**	✓	✓		✓					

B) *Unsupervised learning approaches*

In [32], the authors developed a feature-based sentiment analysis approach to analyze MSA reviews from different social media pages. The approach is indeed based on a seed list of words. Initially, the reviews are collected to form the Corpus. Then, the Corpus is processed to clean, tokenize, and normalize. For the feature extraction activity, the POS tagging features are extracted from reviews to indicate if the word is a noun, it is a feature of opinion, or the word is adjective, it is opinion. To indicate the sentiment polarity of each word, the similarity measurement method is used. In the classification activity, the reviews are classified into positive, negative, and neutral polarities based on the number of positive, negative, and neutral words within each review. The reviews with their polarities are used to assign the weight to each opinionated word based on its frequency in positive and negative reviews. The extracted lexicons (opinionated words with their weights) and reviews are cleaned (from stop words) and stemmed. The preprocessed reviews and lexicons are used to extract sentiment aspect (feature) by applying the syntactic dependency and POS patterns. The unsupervised technique is used to classify the features into positive, negative, and neutral sentiment polarities. To evaluate the performance, the accuracy method is used that gives 92.4% for the sentiment classification.

In [33], the authors proposed a sentiment analysis system that is divided into two modules: the sentiment lexicon module and sentiment classification module. For the sentiment lexicon module, the Arabic stem words are manually collected and translated into the English language. Then, the online English sentiment lexicons were used to determine the semantic orientation of each translated word. The sentiment values divided into positive with weight from 60% to 100%, neutral with weight from 60% to 40%, and negative with weight from 40% to 0%. The resulted lexicon includes about 120,000 Arabic terms. The sentiment module started by collecting modern standard Arabic reviews. The collected reviews are preprocessed to remove non-sentimental data and review words are stemmed to their roots. The extracted lexicons are used as a set of features after giving each word its assigned value. The sentences in collected reviews are classified into their sentiment polarities by summing the sentiment value of words in the sentences. Therefore, the sentiment level employed in this approach is clearly the sentence level. To evaluate the performance, the accuracy method is used, which gives 86.89%.

In [34], the authors proposed an automated approach that uses lexico-syntactic patterns to build dialectical or slang subjectivity lexicon. This lexicon is ready to utilize in the Arabic sentiment analysis process. The approach is held in two phases. In the first phase, it distinguishes the general lexico-syntactic patterns that are not based on any Arabic POS tagger and morphological analyzer, so as to increase the recall. It manually identifies 11 patterns according to how people carried out their opinions in dialectical Arabic sentiment analysis. These patterns are reduced to eight by cleaning them from noisy matched patterns. The cleaned patterns applied to a large Arabic dataset of 7.5 million Egyptian dialectical tweets. This dataset is cleaned by deleting unwanted data such as links, hashtags, and redundant tweets. The normalization technique is applied to the cleaned dataset. The subjectivity terms are extracted from the dataset and manually annotated with positive, negative, and non-sentiment tag by 3 graduated students. The matching between these patterns and the subjectivity terms gets a precision of 88.6%. The polarity classification is assigned to the extracted subjectivity terms. The tweet dataset was annotated using the lexicon list of

2K subjective terms selected from Egyptian dialect lexicon. The PMI method used to calculate the co-occurrence matrix between the extracted terms and tweets. Therefore, the approach is applied to the document level. To evaluate the performance of the approach, the precision method is used, which gives 84.5%.

In [35], the authors focused on extracting and analyzing Arabic business reviews according to their sentiment polarities adopting the lexicon-based approach. They constructed an Arabic lexicon list using a seed list of 1600 words with their 600 positives, 900 negatives, and 100 neutral labels. The lexicon words are used in an Arabic similarity graph using a large Arabic review Corpus to classify the polarity of neighboring words. The extracted lexicons are used in constructing the sentiment analyzer. In addition, the negation mechanism and sentence boundary features are used for sentiment classification. Then, the semantic orientation is calculated by summing up the scores of terms in the sentence and the score of negated terms is added to give the final score of the sentence (positive, negative, neutral, or mixed). Therefore, the system is worked at the sentence level. To test the system, it is compared to the English system performance. The results showed that the system achieves high precision and recall similar to the English system.

In [36], the authors proposed a sentiment analysis system using human computation. The system consists of two phases. The first phase is a game used to manually label a large Corpus of dialectical Arabic reviews from the www.Qaym.com website. In this game, the words in reviews are annotated by users into positive, negative, neutral, and entity (mainly noun), which in turn constructs a sentiment lexicon. The game users label the whole sentence into positive, negative, and neutral to help the system identify the pattern of the sentence and its polarity classification. The second phase is the sentiment analyzer. The sentiment analyzer is based on two proposed methods to classify the sentiment of reviews. The first method is used to extract patterns from the game and to perform the sentence matching mapped to the sentence polarity. The second method is based on the summation of sentimental words in the sentences. Therefore, the system is applied to the sentence level. The accuracy of the method first approach is 56.41% and the second method gives 60.32% for the accuracy metric.

In [37], the authors developed a semantic approach to identify user opinions and business insights from Arabic social media using their Arabic sentiment ontology. This ontology covers groups of opinionated words in different Arabic dialects with their weights. The authors use their approach to identify the sentiment of Arabic tweets in different issues. Technically, the approach consists of four modules: data collecting, filtering, preprocessing, and classification. The data collecting module crawls tweets using Tweet Archivist tool. The filtering module filters out any irrelevant tweets. The preprocessing module removes URLs, fixes typos, shortens long words, and removes stop words. The classification module classifies the tweets into positive, negative, and neutral sentiment polarity by calculating the weights of each opinionated word in tweet according to the predefined Arabic sentiment ontology. The approach is obviously applied at the document level. For evaluating the performance of this approach, the precision metric is applied and reached 75%.

In [19], the authors identified the main challenges and open research trends that face Arabic sentiment analysis of social media such as the lack of available Arabic resources (e.g., datasets and lexicons). They proposed a case study to determine the semantic orientation of Egyptian dialect tweets and comments. Particularly, the approach consists of three steps. It firstly constructs a sentiment lexicon list manually, started by 380 seed sentiment words. This lexicon list is used later to collect more words with respect to the idea of neighboring words have the same polarity. The lexicons manually filter by deleting false words. This step is repeated with new terms until constructing 4392 sentiment lexicons.

TABLE V. THE SUMMARY OF ASSESSING THE UNSUPERVISED LEARNING APPROACHES BY APPLYING THE DEFINED EVALUATION CRITERIA

Ref.	**Dataset**	**Language**		**Sent. Level**			**Performance Metrics**			
		MSA	**DA**	**Doc.**	**Sen.**	**Asp.**	**Acc.**	**Per.**	**Rec.**	**F-measure**
[32]	**Reviews**	✓		✓		✓	✓			
[33]	**Twitter**	✓	✓		✓		✓			
[34]	**Twitter**		✓	✓				✓		
[35]	**Reviews**	✓			✓			✓	✓	
[36]	**Twitter**	✓	✓	✓				✓		✓
[37]	**Twitter**		✓	✓				✓		
[19]	**News page, Twitter**	✓	✓	✓			✓			

The second step is mainly to assign weight to each lexicon using two weight algorithms. Two datasets are used: Twitter dataset contains 500 tweets and Dostour dataset contains 100 comments from Egyptian article. They manually annotated with their semantic polarity. The semantic orientation of tweets is calculated by making use of the two methods. The first method is an in-house NER technique that extracts names to avoid them. The second method concentrates on merely intensifiers and negations. Therefore, the system is worked at a document level. The accuracy method is used to measure the performance of the two methods. The best result is 83.8% accuracy for the first method with the Twitter dataset and 63% accuracy for the second method with the Dostour dataset.

In table V, we first apply our evaluation criteria to assess the above discussed unsupervised learning approaches ad second report a summary of the obtained results.

C) *Hybrid learning approaches*

In [23], the authors implemented a web-based tool using the R language to sentimentally analyze Arabic text tweets. Therefore, the tool is applied to the document level. The tool consists of three parts. The first part asks users to input the interesting topic, while the second part attaches the time input to the tweet collection. The time input consists of the date, which is put back a week before the current date to collect only recent tweets about the topic. The third part is related to the sentiment analysis process. The sentiment analysis process can use either lexicon-based methods or machine learning methods. The selected machine-learning method consists of four main steps: collecting the data, preprocessing data (e.g., stemming tweets and removing retweets), selecting the features (n-grams), and applying the machine learning techniques (SVM and NB). The lexicon-based method is directly determining the polarity from lexicons. The accuracy is 70% for NB and 34% for SVM. The accuracy is 66% for the point mutual information (PMI) method.

In [38], the authors investigated the impact of the preprocessing techniques on the performance of an Arabic sentiment analysis system using Egyptian dialectical tweets. This system then works at the sentence level. Firstly, the preprocessing phase consists of three steps where the first step applies the normalization technique to the tweets. In the second step, the normalized tweets are stemmed. The third step removes the Egyptian dialectal stop words from tweets. The impact of the preprocessing phase on the performance of the supervised and unsupervised approaches is measured as well. The supervised approach is carried out in five steps. The first step collects 1000 tweets from Twitter to build a Corpus and the second step cleans the Corpus and annotates the tweets. The third step preprocesses the Corpus, the fourth step extracts n-gram features, and finally, the fifth step computes tweets sentiment polarities using the SVM classifier. On the other hand, the unsupervised approach is carried out in three steps. The first step builds a sentiment lexicon using 400 annotated tweets, the second step preprocesses the tweets and lexicons, and finally, the third step classifies the remaining tweets into positive or negative using the sentiment lexicons found in the tweets. It is noticeable that the supervised approach with the SVM method achieves a high performance after applying the preprocessing step. Specifically, the accuracy result of applying the preprocessing step is almost 4.5% better than the corresponding one without applying the preprocessing step. The same improvement goes with the precision, recall, and F-measure methods. In addition, the unsupervised approach achieves an improvement of 7% in the accuracy and recall, while there was an improvement of 2% in precision and 5% in the F-measure.

In [20], the authors conducted the sentiment analysis task of comparative sentences in the Arabic language. The sentiment analysis of comparative sentences splits in two processes. The first process identifies comparative sentences from non-comparative ones. This process is based on three classifiers (linguistic, machine learning, and hybrid). The linguistic classifier classifies sentences into comparative/non-comparative using the linguistic properties of POS. The ML classifier uses three methods (SVM, NB, and K-NN). Finally, they integrate between the linguistic and ML approaches. In the second process, the authors developed a set of rules to classify three types of comparative statements. For evaluating this approach, initially, the authors prepared Corpus by crawling Arabic reviews from three domains: education, technology, and sports. Each review manually annotated to comparative/non-comparative labels. In the preprocessing step, the unnecessary data are removed from the Corpus. Then, the normalization, tokenization and stop words removal techniques are applied. For testing, the F-measure method is applied. The linguistic approach reports 63.37%. The ML approach reports 86.36% using the K-NN method. The combination of two approaches results in 88.78% for the accuracy.

In [39], the authors constructed the word2vec model from a large Arabic corpus, gathered from online Arabic newspapers. This gathering process uses ML and convolutional neural networks techniques. The full system is divided into two models: the word embedding model and sentiment analysis model. Initially, they collected Arabic text from existing Arabic dataset that contains 1.5 billion words saved in XML format. Then, the collected data cleaned by removing unwanted data, like digits and non-Arabic words. Some letters are normalized. In the word-embedding model, the word2vec model is applied. The input of this model is a list of sentences where each sentence is divided into a list of words. The output of the model predicts the center word according to the surrounding words. The sentiment analysis model starts by using their Arabic Twitter dataset. The POS, TF-IDF, lexicon features are extracted from the dataset. For the classification activity, SVM, NB, and NN methods are applied to determine the sentiment polarity of tweets. Therefore, the system applied to the document level. They improved the accuracy of the sentiment classification from 91% to 95%.

In [40], the authors investigated the challenges of Arabic social media platforms, especially Twitter platform with respect to the Saudi Arabia dialect. They recommend a hybrid approach as a solution to the Arabic sentiment analysis process. In this approach, the lexical classifier is applied to labeled training data and its output is used as a training data for the SVM classifier. The approach is passed through multiple phases. Initially, the data is gathered from Twitter to design the Corpus. In the preprocessing activity, the tweets are cleaned from noise data (e.g., URLs, mentions), some characters are normalized, stop words are removed, and words are stemmed. The cleaned tweets passed to a lexicon-based classifier to be labeled as positive or negative sentiment polarity. In the feature extraction activity, the n-gram is extracted. In the classification activity, the SVM classifier uses the extracted features for determining the sentiment polarities of the training dataset. Therefore, the system is applied to the document level. Finally, the F-measure and accuracy criteria are applied and achieved 84% and 84.01%, respectively.

In the following table VI, we first apply our evaluation criteria to assess the above-discussed hybrid learning approaches and second report a summary of the obtained results.

V. Discussion

Despite the Arabic sentiment analysis research field started in 2008 [13], it is still flourished. The positive point here is that there is a growing in the research rate, especially with the increasing number of Arabic social media users. By deeply investigating and comparing our results reported in table IV, table V and table VI, we get the following interesting points:

- Most of the current proposals focus only on the MSA language. In addition to, there is no a single system that can consider all aspects of both DIA and MSA with a reasonable accuracy.
- Most of Arabic SA research is mainly conducted on the document level and the sentence level and meanwhile there is a very few works dealing with the aspect level (see, for example, [18] [25] [41]).
- Most of the proposals use accuracy metrics to measure the performance (or the error rate of the system).
- The supervised approach is used more than other unsupervised and hybrid approaches. The most supervised machine learning methods are SVM and NB.
- The most features used to train classifiers are words and their frequencies (n-grams).
- Most of available Arabic SA resources is conducted on movie reviews and there is a new research direction directed toward the use of the Twitter platform.
- There is no Arabic dialectical lexicon list because simply there is no a dictionary assembling all Arabic dialects.
- Most of the Arabic datasets are in-house and the public dataset covers only movie reviews domain.
- The Arabic sentiment analysis process covered different genres (resources) of movie reviews, Facebook's comments, posts, tweets, news comments, and different topics (or domains) such as politics, economy, and social events.
- Detecting and analyzing tweets that are written in Arabizi has not yet thoroughly studied.

TABLE VI. The Summary of Assessing The Hybrid Learning Approaches By Applying The Defined Evaluation Criteria

Ref.	**Dataset**	**Language**		**Sent. Level**			**Performance Metrics**			
		MSA	**DA**	**Doc.**	**Sen.**	**Asp.**	**Acc.**	**Pre.**	**Rec.**	**F-measure**
[23]	**Twitter**		✓	✓			✓			
[38]	**Twitter**		✓		✓		✓	✓	✓	✓
[20]	**Reviews**	✓	✓		✓					✓
[39]	**Twitter, Online newspaper**	✓	✓	✓			✓			
[40]	**Twitter**		✓	✓			✓			✓

In the rest of this section, we focus on comparing the evaluation of the three sentiment classification approaches. This comparison is important to select the suitable approach to address the sentiment analysis challenges with the highest accuracy. These approaches focus merely on solving the Arabic sentiment challenges with social media platforms. For the comparison purpose, we calculate the accuracy improvement average for each classifier using the following equation where n is the number of papers in each sentiment classification approach:

$$\mathrm{AVG(ACC.)} = \sum\nolimits_{k=0}^{n} \frac{Accuracy\ of\ paper_k}{Number\ of\ papers} \qquad (5)$$

Intuitively, the highest $\mathrm{AVG(ACC.)}$ rate means the proposals that address this sentiment challenge have the high accuracies. The following figures present the average accuracy results for sentiment classification approaches. In Fig. 4, the supervised approaches that address the spam challenge have low accuracies. Moreover, the number of these approaches is very small. The supervised approaches that address the dialectical dataset challenge in Fig. 4 have high accuracies. Other figures (Fig. 5 and Fig. 6) can be analyzed in a similar way with respect to unsupervised and hybrid sentiment classification approaches. Note that the NLP overheads in these figures include scrams, abbreviation, negation, comparative sentences, ambiguity, or spam.

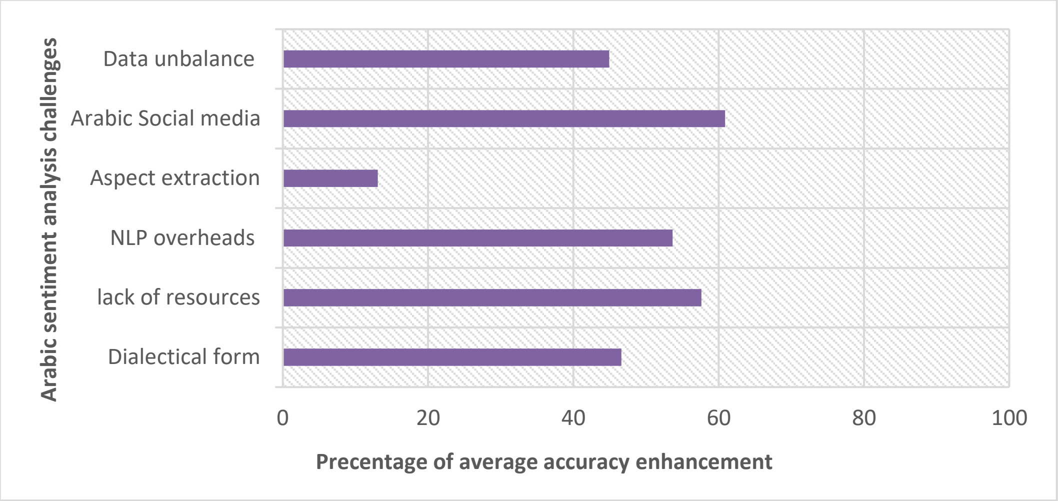

Fig. 4. The improvement in accuracy results of Arabic SA challenges with social media using the supervised approaches

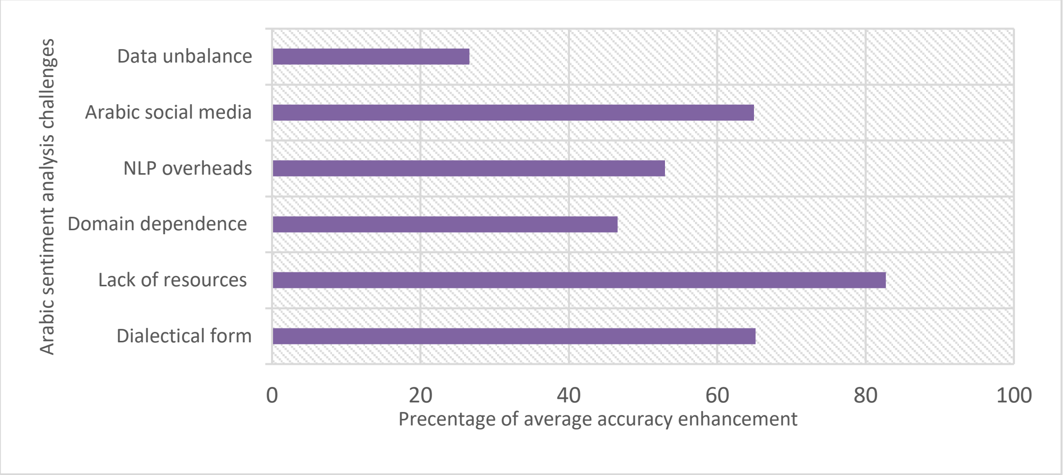

Fig. 5. The improvement in accuracy results of Arabic SA challenges with social media using the unsupervised approaches

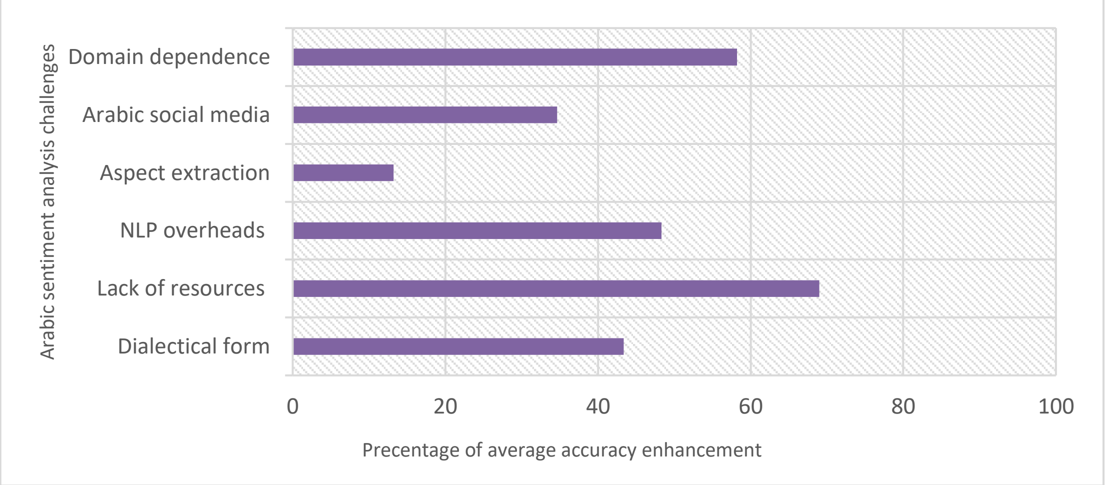

Fig. 6. The improvement in accuracy results of Arabic SA challenges with social media using hybrid approaches

To continue our comparisons, the following figures (Fig. 7, Fig. 8, and Fig. 9) show the highest accuracy of the sentiment analysis approaches that solve each Arabic SA challenge with social media. For example, we found that the supervised approaches have high accuracies in solving the dialectical form, lack of resources, and NLP overheads challenges. Although the supervised approaches have high accuracies, we recommend the hybrid approaches, because the hybrid approaches are domain independent and save time of labelling and annotating training data. The hybrid approaches also have approximately the same accuracy like the supervised approaches in tackling the main features of social media.

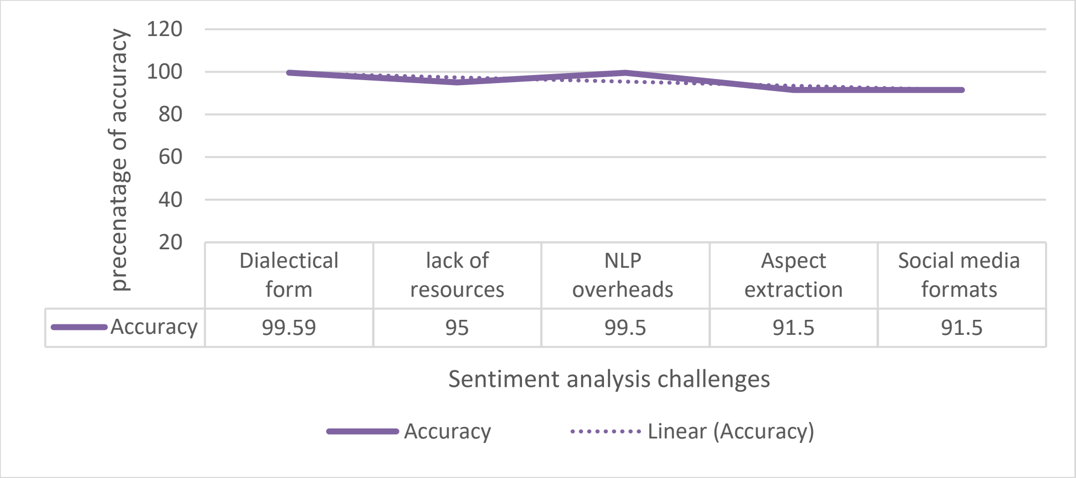

Fig. 7. The highest accuracy of each Arabic SA challenge with social media using the supervised approaches

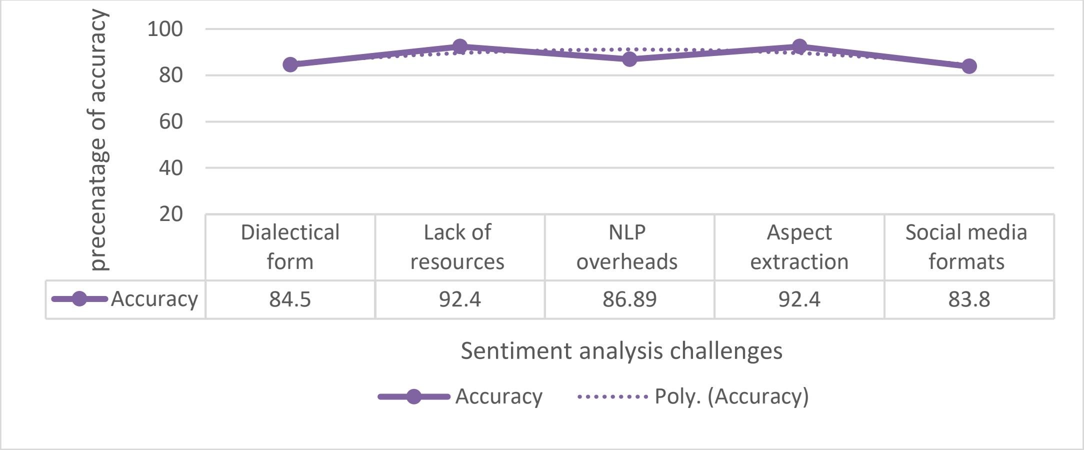

Fig. 8. The highest accuracy of each Arabic SA challenge with social media using the unsupervised approaches

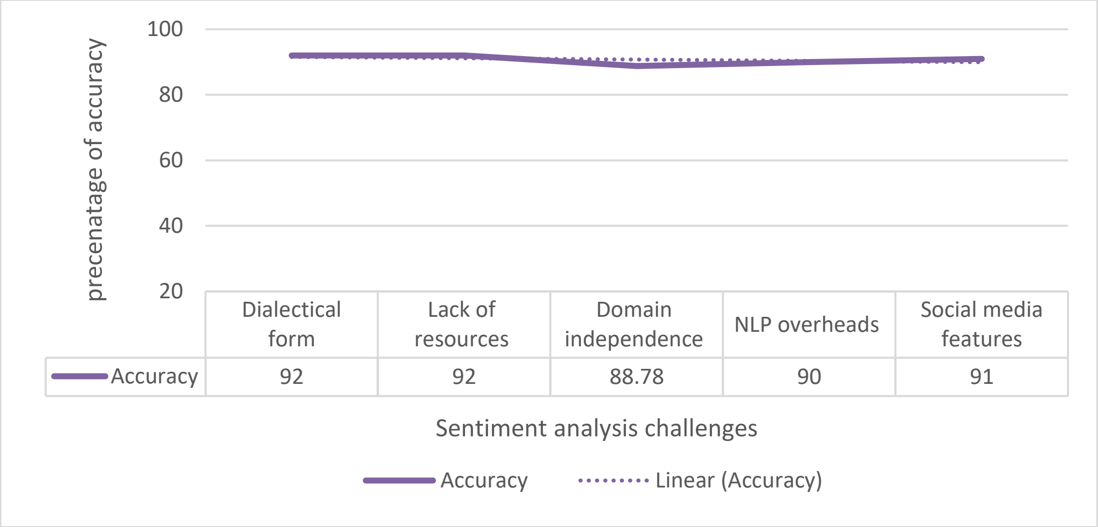

Fig. 9. The highest accuracy of each Arabic SA challenge with social media using the hybrid approaches

VI. Conclusions And Future Work

In this manuscript, we investigated the Arabic sentiment analysis challenges and the challenges generated from social media platforms. We studied the impact of social media challenges on the Arabic sentiment analysis challenges using our developed one-to-one mapping technique. We concluded that the challenges of social media added more complexities on the sentiment analysis process and the majority of these challenges are due to complex nature of Arabic language. Such conclusions helped us to develop a set of evaluation criteria to review and assess numerous proposals that analyze opinions of people in Arabic social media in the literature. We finally compared current sentiment classification approaches. Our findings are that the research on sentiment analysis of Arabic social media is still in its infancy stage. However, this research field is obtaining a satisfied attention from the research community, especially in the past four years. The Arabic sentiment analysis process uses different ML approaches for sentiment classification approaches (supervised, unsupervised, and hybrid). The hybrid approach showed to gain satisfying accuracies in solving the principal features of social media, as it combines the advantages of supervised and unsupervised approaches; although the supervised approaches are the most used approach in the literature. In addition, current proposals covered only the document and sentence sentiment levels on different topics without giving rise to the aspect sentiment level. The covered topics are crawled from different genres, news web pages, reviews, blogs, and social network platforms (Twitter and Facebook).

In the future work, we plan to develop a new hybrid approach at the aspect sentiment level, which is missing in the literature. This proposed approach should use the machine learning NB and SVM methods, as they achieve the best sentiment classification accuracies in the literature. This approach should also use: 1) the NLP techniques in the preprocessing activity to guarantee the satisfaction of high accuracy; and 2) the n-gram features, the most effective features in the three sentiment classification approaches. Given that, we plan to improve our approach using a deep learning technique and big data techniques such as Hadoop and MapReduce to obtain optimal (or near optimal) results.

References

[1] B. Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions", New York, USA: Cambridge University Press, 2015.

[2] B. Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool, 2012.

[3] B. Batrinca and P. C. Treleaven, "Social media analytics: A survey of techniques, tools and platforms", *AI and Society, Vol.30 (1),* pp. 89-116, 2015.

[4] M. Bonzanini, "Mastering Social Media Mining with Python", Birmingham, UK: Packt Publishing Ltd, 2016.

[5] M. Shearlaw, "Egypt five years on: was it ever a 'social media revolution'? ", 2017. [Online]. Available: https://www.theguardian.com/world/2016/jan/25/egypt-5-years-on-was-it-ever-a-social-media-revolution. [Accessed May 2018].

[6] M. Abdullah and M. Hadzikadic, "Sentiment Analysis on Arabic Tweets: Challenges to dissecting the language", *International Conference on Social Computing and Social Media,* pp. 191-202, 2017.

[7] G. Alwakid , T. Osman and T. Hughes-Roberts, "Challenges in Sentiment Analysis for Arabic Social Networks", *Third International Conference On Arabic Computational Linguistics, vol. 117, pp. 89-100,* Dubai, UAE, 2017.

[8] S. R. El-Beltagy and A. Ali, "Open issues in the sentiment analysis of Arabic social media: A case study", *The 9th International Conference on Innovations in Information Technology*, pp. 215-220, 2013.

[9] N. Al-Twairesh, H. Al-Khalifa and A. Al-Salman, "Subjectivity and sentiment analysis of Arabic: trends and challenges," in *11th International Conference on Computer Systems and Applications. pp. 148-155,* Doha, Qatar, IEEE Computer Society, 2014.

[10] F. Albogamy and A. Ramsay, "POS Tagging for Arabic Tweets," in *Recent Advances in Natural Language Processing, pp. 1-8,* Hissar, Bulgaria, 2015.

[11] "Arab knowledge economy report 2016", 19 3 2016. [Online]. Available: https://www.arab knowledge economy report 2016.COM. [Accessed 20 6 2018].

[12] N. Y. Habash, Introduction to Arabic Natural Language Processing, Morgan & Claypool, 2010.

[13] N. Boudad, R. Faizi and R. Thami, "Sentiment analysis in Arabic: A review of the literature", *Ain Shams Engineering Journal,* 2017.

[14] N. Habash, R. Eskander and A. Hawwari, "A morphological analyzer for Egyptian Arabic", *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology,* pp. 1-9, 2012.

[15] M. Abdul-Mageed and M. T. Diab, "SANA: A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis", in *LREC,* pp. 1162-1169, 2014.

[16] I. Obaidat, R. Mohawesh, M. Al-Ayyoub, A.-S. Mohammad and Y. Jararweh, "Enhancing the determination of aspect categories and their polarities in arabic reviews using lexicon-based approaches", *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT),* , pp. 1--62015,.

[17] M. Abdul-Mageed and M. T. Diab, "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis", *LREC,* pp. 3907-3914, 2012.

[18] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López and J. M. Perea-Ortega, "OCA: Opinion corpus for Arabic", *Journal of the Association for Information Science and Technology,* pp. 2045-2054, 2011.

[19] S. R. El-Beltagy and A. Ali, "Open issues in the sentiment analysis of Arabic social media: A case study," *9th international conference on Innovations in information technology (iit),*, 2013.

[20] A. El-Halees, "Opinion mining from Arabic comparative sentences," in *The 13th International Arab Conference on Information Technology ACIT*, pp. 265-271, 2015.

[21] A. b. Hammad and A. El-Halees, "An approach for detecting spam in arabic opinion reviews", *The International Arab Journal of Information Technology,* 2013.

[22] M. Al-Smadi, O. Qawasmeh, B. Talafha and M. Quwai0der, "Human annotated arabic dataset of book reviews for aspect based sentiment analysis", *3rd International Conference on Future Internet of Things and Cloud (FiCloud),* 2015, pp. 724-730.

[23] M. El-Masri, N. Altrabsheh, H. Mansour and A. Ramsay, "A web-based tool for Arabic sentiment analysis," *Procedia Computer Science,* pp. 38-45, 2017.

[24] A. Assiri, A. Emam and H. Aldossari, "Arabic Sentiment Analysis: A Survey," *IJACSA*, pp. 75-85, 2015.

[25] K. M. Alomari, H. M. ElSherif and K. Shaalan, "Arabic Tweets Sentimental Analysis Using Machine learning", *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 602--610, 2017.

[26] S. O. Alhumoud, M. I. Altuwaijri, T. M. Albuhairi and W. M. Alohaideb, "Survey on Arabic Sentiment Analysis in Twitter", *World Academy of Science, Engineering and Technology,* pp. 1-6, 2015.

[27] A. Shoukry and A. Rafea, "Sentence level arabic sentiment analysis," in *Collaboration Technologies and Systems (CTS), 2012 International Conference*, pp. 546-550, 2012.

[28] M. Alhazmi and N. Salim, "Arabic opinion target extraction from tweets," *ARPN Journal of Engineering and Applied Sciences,* vol. 10, no. 3, pp. 1023--1026, 2015.

[29] A. E.-D. A. Hamouda and F. E.-z. El-taher, "Sentiment Analyzer for Arabic Comments System", *International Journal of Advanced Computer Science and Applications(Int. J. Adv. Comput. Sci. Appl),* vol. 4, no. 3, pp. 99-103, 2013.

[30] M. Abdul-Mageed, M. Diab and M. Korayem, "Subjectivity and Sentiment Analysis of Modern Standard Arabic", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[31] M. Abdul-Mageed, M. Diab and S. Kübler, "SAMAR: Subjectivity and sentiment analysisfor Arabic social media," *Computer Speech \& Language,* vol. 28, no. 1, pp. 20-37, 2014.

[32] L. A.-E. Abd-Elhamid, D. Elzanfaly and A. Sharaf Eldin, "Arabic Feature-Based Level Sentiment Analysis," *journal of fundemental and applied sciences,* vol. 28, pp. 143-148, 2018.

[33] M. Al-Ayyoub and S. Bani Essa, "Lexicon-based sentiment analysis of Arabic tweets," *Int. J. Social Network Mining,* vol. 2, no. 2, pp. 101-114, 2014.

[34] H. ElSahar and S. R. El-Beltagy, "A Fully Automated Approach for Arabic Slang Lexicon," *International Conference on Intelligent Text Processing and Computational Linguistics*, Egypt , 2014.

[35] M. Elhawary and M. Elfeky, "Mining Arabic Business Reviews," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference*, 2010.

[36] A. S. Al-Subaihin and H. S. Al-Khalifa, "A system for sentiment analysis of colloquial Arabic using human computation", *The Scientific World Journal*, pp. 1-12, 2014.

[37] S. Tartir and I. Abdul-Nabi, "Semantic sentiment analysis in Arabic social media", *Journal of King Saud University – Computer and Information Sciences,* vol. 29, no. 2, pp. 229-233, 2016.

[38] A. Shoukry and A. Rafea, "Preprocessing Egyptian Dialect Tweets for Sentiment Mining," 2011.

[39] A. M. Alayba, V. Palade, M. England and R. Iqbal, "Improving Sentiment Analysis in Arabic Using Word Representation", *arXiv preprint arXiv:1803.00124,* pp. 1-6, 2016.

[40] H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis – a hybrid," *Journal of Information Science,* vol. 42, no. 6, pp. 782-797, 2016.

[41] L. A.-E. Abd-Elhamid, D. Elzanfaly and A. Sharaf Eldin, "Arabic Feature-Based Level Sentiment Analysis," *journal of fundemental and applied sciences,* vol. 28, pp. 143-148, 2018.