IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 4, 2021, accepted September 12, 2021, date of publication October 15, 2021, date of current version October 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3120746

MARSA: Multi-Domain Arabic Resources for Sentiment Analysis

AREEB ALOWISHEQ[1,2], NORA AL-TWAIRESH[2,3], MAWAHEB ALTUWAIJRI[4], AFNAN ALMOAMMAR[1], ALHANOUF ALSUWAILEM[1], TARFA ALBUHAIRI[5], WEJDAN ALAHAIDEB[4], AND SARAH ALHUMOUD[1]

[1]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh 13318, Saudi Arabia
[2]National Center for Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA), Riyadh 12391, Saudi Arabia
[3]STC's Artificial Intelligence Chair, Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh 11495, Saudi Arabia
[4]General Directorate of Information Technology, Princess Nourah Bint Abdulrahman University, Riyadh 11564, Saudi Arabia
[5]Elm, Riyadh 12333, Saudi Arabia

Corresponding author: Areeb Alowisheq (aaalowisheq@imamu.edu.sa)

This work was supported by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University.

ABSTRACT The Arabic language has many spoken dialects. However, until recently, it was primarily written in Modern Standard Arabic (MSA), which is the formal variant of Arabic. Social media platforms have changed the face of written Arabic where users converse freely in various dialects, thus offering a massive number of resources for the study of dialectal text. The Arabic dialects differ from MSA in morphology, syntax, and phonetics. Consequently, since the effectiveness of NLP tasks—like sentiment analysis—is dependent on the availability of representative resources, there is currently a great need for such resources in these dialects. In this paper, we present MARSA—the largest sentiment annotated corpus for Dialectal Arabic (DA) in the Gulf region, which consists of 61,353 manually labeled tweets that contain a total of 840 K tokens. The tweets were collected from trending hashtags in four domains: political, social, sports, and technology to create a multi-domain corpus. The importance of such a corpus is to facilitate the study of domain-dependent sentiment analysis in Arabic. In addition to this corpus, the annotators extracted indicator words to form affect lexicons for each domain. We draw insights from these lexicons regarding contextual polarity of certain words. Furthermore, we present benchmark experiments on the MARSA corpus in order to establish a baseline for further studies.

INDEX TERMS Corpus, sentiment analysis, dialectal Arabic.

I. INTRODUCTION

The Arabic language is spoken by more than 400 million people in the world and is ranked as the fifth most-spoken language [1]. Arabic is mainly divided into Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA). Today, CA is mainly found in pre-Islamic poems, traditional texts, and in the Holy Quran, while MSA is the formal language that is used in news, books, and education. DA, on the other hand, is derived from MSA and had rarely been used in written form—except recently, with the social media revolution. DA is generally classified into five major geographical dialects: Egyptian, Gulf, Iraqi, Levantine, and Maghribi [2]. Furthermore, each of these geographical dialects has several local varieties.

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

Social media platforms have become one of the major communication mediums nowadays due to the higher accessibility to the internet and rapid adaptation to smart phones. One of the most used social media platforms is the microblogging platform Twitter, with more than 353 million users in 2020 [3]. Interestingly, Arabic is the fourth most used language on Twitter [4]. Arab users on Twitter primarily communicate and express opinions in DA. Consequently, research on DA analysis has gained significant interest recently. Harvesting the content available online for value and meaning is a rapidly growing demand in multiple sectors. Employing sentiment analysis to discern trends, opinions and attitude in social media aids in understanding large number of users and costumers in an automated way to provide better services.

A corpus comprising data entries and their labels is an essential resource for creating sentiment analysis learning models or classifiers, enhancing machine's linguistic

intelligence in order to improve understanding of available data. Mainly, corpus annotation is accomplished in three different ways [5]. First, there is the manual approach in which a group of individuals with linguistic proficiency, consisting of at least two members, perform the annotation. Second, there is the crowdsourcing approach that utilizes assistive interface tools. Third, there is the automatic approach in which the correct annotation label is deduced from a type of rating indicator, such as star-ratings in review systems or emojis on social media platforms.

This paper presents MARSA, the largest DA corpus annotated for sentiment classification purposes in the Gulf dialect. This corpus comprises of 61,353 tweets. Other than being in DA, the importance of the created corpus is that it is a multi-domain corpus covering the following domains: sports, politics, technology, and social issues. This facilitates research into the contextual polarity of certain words and phrases, where a word can be positive in a certain domain, and negative in another. It also enables the training of domain-specific classifiers, as demonstrated in this paper, which can enhance the performance of sentiment analysis.

MARSA was annotated manually with 11 annotators completing the job in four months. A manually annotated corpus requires high human labor because annotators must assess each data entry and classify it under one of the provided labels. The tweets were classified into five labels: positive, negative, neutral, sarcasm, and both. *Positive* and *negative* were used to label tweets with the corresponding affect, while a tweet that does not hold any polarity toward either positive or negative was labeled *neutral*. In addition, *sarcasm* was used to label tweets where the meaning of the words in a tweet were opposite of what the user intended to say, which in terms of sentiment means that positive words were used to covey a negative sentiment and vice-versa. Last, *both* was used to label tweets that contain both positive and negative sentiments.

The remainder of the article is structured as follows. Section II provides an overview of the related work on sentiment corpora in Arabic. In Section III, we describe our approach and observations while creating the corpus and lexicons. Section IV explains the challenges we faced in annotating the corpus. In Section V, we report on the results of our benchmark experiments on the corpus, while Section VI presents our conclusion.

The lexicons are publicly available at the group's repository [6]. The corpus is available on request.

II. RELATED WORK

Research on Arabic Sentiment Analysis (SA) has gained much attention over the last few years—numerous efforts have been made in the field and the number of studies on Arabic SA has significantly increased [7]–[9]. Despite this expansion of Arabic SA corpora, there is still a gap in the field because constructing such corpora is costly in terms of time and effort. However, there exist research that has constructed corpora for Arabic SA with different genres of text. Early work focused on reviews, as in [10]–[15]. Recently, the focus has shifted to social media platforms, such as Twitter, due to the proliferation of these outlets among users. Since the focus of this paper is on Arabic tweets, we review the corpora and lexicons that we found in the literature on the SA of Arabic tweets. As stated in [9], resource quality significantly influences the classification performance.

Refaei *et al.* [16] constructed and released a corpus of Arabic tweets annotated for SA, which is available in the LREC repository of shared resources. It consists of 6,894 tweets: 833 positive, 1,848 negative, 3,685 neutral, and 528 mixed. It was annotated for morphological features, simple syntactic features, stylistic features, and semantic features.

One of the earliest datasets on the SA of Arabic tweets was the Arabic Sentiment Tweets Dataset (ASTD) [17], which is an Arabic tweet corpus written in the Egyptian Dialect. It consists of approximately 10,000 tweets that are classified as objective, positive, negative, and mixed. It presents baseline models in order to provide benchmarks for future work.

Similarly, [18] presented the AraSenti-Tweet corpus, which is a corpus of Arabic tweets written in the Saudi Dialect and annotated for sentiment. This corpus was manually annotated in order to produce a gold standard using four classes: positive, negative, neutral, and mixed. Subsequently, it was used in the development of different benchmark SA classifiers.

SemEval is a yearly series of semantic evaluation tasks that are held to foster competition in several tasks related to semantic analysis systems. Since SemEval 2013, a task was dedicated to Twitter sentiment analysis. This task endorses SA research of short informal texts and provides a benchmark for the comparison of different approaches. In SemEval 2017 [19] and 2018 [20], Arabic tweet datasets that are annotated for sentiment classification were also included, serving as excellent benchmarks for the Arabic SA research community.

In [21], a manually annotated Arabic Speech Act and Sentiment corpus of tweets (ArSAS) is presented. It is considered to be the first corpus of Arabic speech act on Twitter because it is annotated for six different classes of speech act: assertion, expression, recommendation, respect, question, and miscellaneous. Moreover, the tweets are also annotated for four classes of sentiment: positive, negative, neutral, and mixed. The corpus contains more than 21,000 Arabic tweets.

In [22], using SA as a case study, the authors investigated whether it is possible to adapt classification models that have been trained on MSA data for texts written in DA. The DA used in this study was the Levantine DA. Hence, a new corpus of tweets written in the Levantine DA was presented and annotated for sentiment. Subsequently, several experiments on sentiment classification were performed using this corpus. The results showed that a model trained on the MSA corpus does not perform well on the DA corpus, suggesting that dialects should be treated as separate languages.

The Arabic Tweets Sentiment Analysis Dataset (ATSAD) is presented in [23], where distant supervision was employed

through the use of emojis as noisy labels in order to collect a dataset of 36,000 tweets that were labeled as positive and negative; subsequently, a subset of 8,000 tweets was annotated manually. To evaluate the corpus, emoji-based annotation was compared to human annotation. In addition, the human-annotated dataset was used to improve the annotation of the automatically-labeled dataset through self-training approaches.

Table 1 presents a summary of the highlighted Arabic corpora mentioned earlier. We can see from this table that the largest corpus found contains 38,037 tweets, the corpus we present in this paper exceeds this number. Moreover, all the papers mentioned previously do not present a multi-domain corpus. In this paper, we aim to fill this gap.

An essential SA resource is sentiment lexicons, where words are labeled in accordance with their sentiment polarity (positive, negative, neutral). Sentiment lexicons are created either manually or automatically. In the manual approach, words that are extracted from datasets are manually labeled as positive, negative, or neutral. These lexicons are usually more accurate than sentiment lexicons that are constructed automatically—however, they are limited in size. Several Arabic sentiment lexicons that were constructed manually include [24]–[30]. These manually constructed lexicons require human effort and time; hence, automatic approaches have been proposed. [31] proposed an automatic approach using graph reinforcement applied on machine translation tables of an English lexicon translated into Arabic, while [32] performed an automatic mapping of the Arabic WordNet (AWN) 2.0 to the English SentiWordNet (SWN) 3.0 through union gloss-synset string matching. [....33] used a seed list to expand on AWN 2.0 synset relations. Similarly, [34] applied automatic gloss-synset matching between AraMorph English gloss terms and SWN synset terms adjusted using heuristics and manual back-offs. [35] used the translation of an English lexicon (MPQA) and term expansion utilizing synonyms, followed by Pointwise Mutual Information (PMI) between terms and seed words in a large set of reviews. [36] also used English lexicons translation and PMI on a large-scale dataset of Arabic tweets.

Although these Arabic sentiment lexicons have shown comparable performance for Arabic SA, they are not all domain-specific. Consequently, a word that has opposite sentiment in different domains cannot be detected and causes ambiguity for the sentiment classifier. Therefore, in this paper, we aim to fill this gap by proposing an Arabic sentiment lexicon that is domain-specific.

III. CORPUS CREATION

The MARSA corpus comprises tweets collected from Twitter. It is manually annotated for sentiment and the tweets are categorized into four domains: social, political, sports, and technology. One important byproduct of the process was the curation of sentiment lexicons—one for each of the four domains. This section describes the corpus creation process, which consists of four stages: data collection, preprocessing, annotation, and inter-annotator agreement. The following sub-sections explain these four different stages in detail.

TABLE 1. Existing arabic corpora.

Year	Author	Pos	Neg	Neut	Mix	Total
2014	Refaee [16]	833	1,848	3,685	528	8,868
2015	Nabil [17]	799	1,684	6,691	832	10,006
2017	Al-Twairesh [18]	4,957	6,155	4,639	1,822	17,573
2018	Elmadany [21]	4,643	7,840	7,279	1,302	21,064
2018	Rosenthal [19]	2,479	3,492	4,155	NA	10,006
2018	Mohammad [20]	NA	NA	NA	NA	1,800
2020	Qwaider [22]	18,173	18,695	NA	NA	36,868
2020	Abu Kwaik [23]	14,887	15,589	NA	7,561	38,037

A. DATA COLLECTION

Over half a million tweets, around 658,000, were collected between November 2015 and February 2016 using Twitter API and R scripts. The tweets were collected from trending hashtags in Saudi Arabia in four different domains: social, political, sports, and technology (tech), Table 2. The tech domain focused on hashtags related to the weakness of internet connections that were targeted at telecommunication companies. The sports domain focused on hashtags that were created and active during football matches. The social domain focused on hashtags about issues affecting the Saudi society, such as royal orders, Saudi budget, issues affecting the income of Saudi citizens, etc. It also included hashtags about shocking stories or controversial issues that initiated substantial reaction as well as hashtags that speculated or reported on school closings due to weather conditions. The political domain focused on covering political events, including news about terrorism or military activities, as well as on hashtags initiated by Saudi government opponents.

B. PREPROCESSING

The data was cleaned from irrelevant content, such as user mentions, URLs, emojis, and non-Arabic characters. We also removed content that did not affect meaning, such as elongations, diacritics, and punctuation marks (except for underscores). In addition, we normalized different Arabic letter forms—for example, the different forms of *alif* (ا,آ,أ) were converted into (ا), the letter *ta* (ة) was converted to (ه).

Initially, the collected tweets contained many duplicate tweets that were subsequently removed; however, spam presented the main challenge because spam tweets constituted the majority of the corpus in the beginning. This persuaded us to develop a spam detector [37]. It was trained on an annotated sample of the data where the annotators labelled spam tweets as noise. The resulting spam detector was applied to the entire corpus. The details were published in [37]. After removing duplicate and spam tweets, the corpus size decreased from 658,000 to 142,434 tweets, with 22% of tweets left.

TABLE 2. Corpus description before preprocessing.

Domain	Date	Keyword/Hashtag	Description	Count
Tech	Nov, Dec	STC, MBC,# انترنت_السعوديه_ضعيف #كلمة_لشركة الإتصالات# موبايلي# #حذف_قناة_العربية_من_الريسيفر	about internet connection weakness	34,242
Sport	Nov, Dec, Feb	#القادسية والهلال, #الرائد والاهلي, #النصر_الإتحاد, الهلال و هجر, #الهلال_الشباب, #الهلال_التعاون #الاتحاد_الخليج, #الهلال_التعاون# الهلال_نجران, #الهلال_الاهلي# برشلونه_روما, الهلال#	Hashtags created and active during football matches.	241,221
Social	Nov, Dec	#وامر_ملكيه #الميزانية_السعودية #رسوم_الأراضي_البيضاء #رفع_اسعار_البنزين_والكهرباء	Royal orders affecting budget, income, prices.	251,000
		#العثور_على_جوري_الخالدي #لماذا_يا_خوله_العنزي #السماح_بالسينما_في_السعودية بريده_تغرق#الرياض_تغرق, #امطار_بريدة	Shocking stories or controversial issues that initiated substantial reaction. Hashtags to speculate or report school closings due to weather conditions.	
		#تعليق_الدراسه_في_الرياض #تعليق_الدراسه_في_القصيم تعليق_الدراسة_في_الرياض,# وكاله_تعليق_الدراسه# #إعفاء_عزام_الدخيل		
Political	Nov, Dec, Jun, Feb	تفجيرات_باريس, #اليمن, ##فرنسا, داعش #هجمات_باريس, # باقية_وتتمدد, # العالم_ضد_الارهاب,# #حصار_غزة جريمة وإبادة, محمد_سلمان##طعن_مبتعثة_بأمريكا, #رجال_الحد_الجنوبي, #الطائرات_الروسية #دعم_البضائع_التركية, #سلمان_عبدالعزيز #كيف_واجهت_السعودية_القاعده, داعش# حزب_الله, #إسقاط_مقاتلة_روسية, # الحوثي قاسم_يهاجم_الملك_عبدالله, # #الحشد_الشعبي #عاصفة_الحسم, #نمر_النمر #جريمة_اعدام_الشيخ_النمر, #تنفيذ_القصاص_علي_47_ارهابي, امن_السعوديه_خط_احمر, # #عاصفة_الحزم القصاص_بـ47_إرهابي,# #داعشي يقتل ابن عمه	Political events that included news about terrorism or military activities. It also included hashtags initiated by Saudi government opponents	131,860

TABLE 3. Examples of tweets from different labels.

Label	Example
Positive	الحمد لله الف مبروووك جمهور الزعيم الفوز عقبال الكاس ان شاا الله الهلال_التعاون Thank God congratulations Al-Zaeem's (Al-Hilal) fans, wishing the cup God willing Al-Hilal_Al-Taawoun
Negative	تحذير اذاعه ام بي سي اف ام سبب رئيسي لهبوط الذائقه بشكل عام Warning MBC FM radio is a main cause for the drop in tastefulness in general
Both	التعاون يبذل فيشكر الحظ والتحكيم في اعتقادي اهم الاسباب في عدم الفوز ههاردلك سكري قصيم مبروك للهلال التاهل الهلال Al-Taawoun gives and should thank luck refereeing in my opinion is the main reason in not winning hard luck sukkari Al-Qassim (Al-Taawoun) Congratulations Al-Hilal Al-Hilal _Qualifying
Neutral	لقاا الامين العام للجنه الوطنيه لمكافحه المخدرات في صباح الخير ياعرب Meeting of the Secretary-General of the National Commission for Narcotics Control in Good Morning Arab show
Sarcasm	السماح_بالسينما_في_السعوديه السينماا بالكثير زياره او زيارتين وبيجولونها صاله رسوم متحركه لحضانه الاطفال هع Allowing_cinema_ in_Saudi_Arabia cinema with at most one or two visits and it will be transformed into a cartoon hall for children's daycare Haha
Cannot be determined (ND)	اخذ رايتك واغطسها بنوتيلا واكلها الهلال_التعاون Take your flag dip it in Nutella and eat it Al-Hilal_Al-Taawoun

C. ANNOTATION

Out of 142,434 tweets, the annotators manually labeled 107,581 tweets with one of six labels: positive, negative, both, neutral, sarcasm, cannot be determined (Table 3). To do so, they used the following guidelines:

- Positive: There is a clear indicator that the opinion is positive even if it is not strong.
- Negative: There is a clear indicator that the opinion is negative even if it is not strong.
- Both: A tweet has a mixed positive and negative sentiment with the same strength.
- Neutral: There is no opinion in the tweet (i.e., news).
- Sarcasm: A tweet says something positive while its meaning is negative (or vice versa).
- Cannot be determined (ND): The existence and direction of the polarity is not clear.

A simple annotation interface was created. The interface is shown in Figure 1. It shows a tweet and asks the annotator to select one of the labels.

The affect lexicons for each domain were created by asking annotators first to extract indicator words from the tweets labeled as positive and negative and then to enter them into a designated field on the same interface. The indicator word is an affect word that determines the polarity of the tweet, as shown in Table 4.

Annotators were recruited from either graduates or undergraduates at Imam Mohammad Ibn Saud Islamic University and King Saud University. They were native Arabic speakers who spoke the Gulf Arabic dialect. They were also, comfortable with using technology. In addition, they were all Twitter users, which means that they were aware of this platform's culture and jargon. They were trained by being provided with annotation guidelines, accompanied with examples for each label to minimize user recall and aid efficiency. Several meetings were held to clarify ambiguities and to familiarize annotators with the task. The annotation process was monitored by a research team member.

For each domain, the total number of annotated tweets is shown in Table 5. There was a total of 11 annotators and each tweet was annotated by 2 annotators. This process took approximately four months.

FIGURE 1. The ASA annotation system interface.

TABLE 4. Examples of indicator words.

	Tweet	Affect words
Pos	احب الهلال حب الوفا حب القدامى الاولين الهلال_الاهلي الهلال	أحب، الوفاء، حب
	I love Al-Hilal loyal love first old love Al-Hilal Al-Hilal_Al-Ahli	I love, loyal, love
Neg	اتبا لهذا الفكر الذي مبني على الغدر والخيانه حسبنا الله ونعم الوكيل داعشي_يقتل_ابن_عمه	تبا
	Damn this thought, which is based on treachery and betrayal. God suffices us. ISIS_agent_kills_his_cousin	Damn

TABLE 5. Number of annotated tweets in each domain.

Domain	Total number of tweets
Tech	14,350
Sports	44,637
Social	32,069
Political	16,525
Total	107,581

D. CORPUS STATISTICS

The annotation results are presented in Table 6. The table shows the number of tweets classified by domain and the six labels from Table 3. The *Conflict* column shows the number of tweets in each domain for which annotators disagreed regarding labels. At the end of this stage, annotators were asked to review the tweets for which there was a disagreement on and the results presented in the table show the number of conflicts after this review. Therefore, the resulting corpus contained 61,353 tweets, labelled as positive, negative, both, neutral, or sarcasm.

The next section discusses how annotator agreement was measured for the corpus. As shown in Table 6, the number of negative-labeled tweets exceeded the positive ones in all domains, except sports. We could interpret the greater positive sentiment in the sport domain to be the result of the enthusiasm that fans have when supporting their teams during football matches.

TABLE 6. Annotated tweets statistics.

	Pos	Neg	Both	Neut	Sar	ND	Conf
Tech	441	3,410	79	1,214	402	6,046	2,758
Sport	12,205	8,258	767	6,026	1,081	4,194	12,106
Social	2,896	5,234	101	7,468	1,882	4,621	9,867
Pol.	1,785	3,829	166	4,018	91	993	5,643
Total Tweets	17,327	20,731	1,113	18,726	3,456	15,854	30,374
Total Tokens	224,991	336,400	18,149	219,618	41,544	191,034	367,216

However, in the political domain, opinions were highly polarized, and individuals typically engaged in hashtags in order to confront and insult opponents rather than to show support to their affiliation (side). Furthermore, trending hashtags were rather negative in nature, such as #جريمة_اعدام_الشيخ_النمر (the crime of executing Sheikh Al-Nimr) and #داعش (ISIS).

In social and technology domains, a similar negative tendency was observed. This can be explained by how individuals use Twitter to vent on and complain about issues in both domains. The higher overall negative sentiment, in general, can be attributed to negativity bias or negativity effect, where people tend to psychologically be affected by negative things more than positive ones [38].

Negativity bias has been observed in social media interactions with varying findings. A recent study on US political hashtags [39] showed that participant comments on news articles that contain these hashtags had more negative language in comparison with the control group. This resonates with our observations for tweets within the political domain. In addition, Jenders *et al.*'s [40] analysis of retweets showed that negative messages are more likely to be retweeted. Similar findings were reported in an analysis of tweets about traffic and transportation [41], [42]. However, other studies have found that there is a bias toward positive tweets [43], [44].

Reflecting on the related work that was presented in Table 1, we can also observe the prevalence of negative tweets over positive ones. Therefore, the negativity observed in both this corpus and others raises an important question—is the popularity of a hashtag on Twitter correlated to the volume of negative interactions? This notion is supported by our corpus, especially because the tweets were collected from trending hashtags, which means that they attracted more participation than other tweets.

E. LEXICONS STATISTICS

With respect to affect lexicons, each domain has two lexicons: a positive lexicon and a negative lexicon. These were curated manually by annotators during the annotation process, as explained in Section III-C.

Table 7 shows the sizes of the positive and negative lexicons for each domain. As expected, their sizes correspond to the number of tweets in each domain as shown in Table 6.

TABLE 7. Affect lexicons sizes.

	Positive	Negative
Tech	262	1,675
Sport	2,046	2,981
Social	798	2,676
Political	595	2,097
Total	**3,084**	**6,377**

TABLE 8. Common words in affect lexicons.

Neg/Pos	Tech	Sport	Social	Political
Tech	-	125	98	66
Sport	415	-	311	217
Social	470	586	-	151
Political	304	406	404	-

TABLE 9. Jaccard index of common words in affect lexicons.

Neg/Pos	Tech	Sport	Social	Political
Tech	-	0.057	0.102	0.083
Sport	0.098	-	0.123	0.090
Social	0.121	0.116	-	0.121
Political	0.088	0.087	0.092	-

Table 8, shows the number of affect words that are common between two different domains. The shaded numbers represent the number of common positive words between two domains, while the non-shaded represents the number of negative ones.

We can see that the greatest overlap is between the sport and social domains in terms of both positive and negative words, while the lowest overlap is between the technology and political domains. This was expected and correlates with domain sizes. In addition, the overlap ratios were calculated and are shown in Table 9. The Jaccard index J, was used to calculate the overlap ratios, which is the ratio of the intersection over the ratio of the union [45]:

$$J = (|A \cap B|)/(|A \cup B|) \tag{1}$$

The overlap ratios in Table 9 still show a greater overlap between the sport and social domains in terms of positive lexicons, while the social and political domains are slightly higher in negative ones.

Another interesting aspect to explore was the words considered positive in one domain and negative in another, as well as the words that annotators considered to be both negative and positive within the same domain. The number of these words for each domain is shown in Table 9. Examples of these words are explained below and shown in Table 11.

In Table 11, Example 1 shows the word رخيص (cheap), which was considered to be positive in the social lexicon but negative in the political lexicon. The word ''cheap'' is typically used to positively describe a service or a product in social discourse; however, at the same time, it also has negative connotations when describing a human being, which was the case in political discussions.

TABLE 10. Number of words considered to be both positive and negative.

Neg/Pos	Tech	Sport	Social	Political
Tech	7	14	6	5
Sport	3	37	9	7
Social	10	24	14	11
Political	3	7	3	4

TABLE 11. Examples of tweets with common words.

#	Positive tweet	Negative tweet
1	رفع_اسعار_البنزين_والكهربا مازال البنزين رخيص مقارنه بدول العالم	التشويش على الناس في فهم اي مقال باشاعه فكره مسبقه عنه اسلوب رخيص ينتهجه من يسمون مثقفين
2	الهلال_هجر اعيد واكرر انا الان اشوف ب الموسم ذا الميدا هو اخطر مهاجم	تعليق_الدراسه_في_القصيم وش رايكم ه ذا جو الواحد يدرس فيه وبكرى راح تكو ن الطرق اخطر من كذا بعد
3	تعليق_الدراسه_في_القصيم ياحظهم القصمان مدلعينهم هاليومين افرحو	تعليق_الدراسه_في_القصيم جيل مدلع المفروض الوزاره ما تعلق الدراسه

The word أخطر (more dangerous) was considered by annotators to be positive in the sport lexicon and negative in the social lexicon, as shown in Example 2 in Table 11.

Interestingly, the annotators also considered selecting the same indicator words as both positive and negative in the same domain. For example, in the social domain, the word ''مدلع'' (spoilt) was used positively in the first tweet and negatively in the second, as shown in Example 3 in Table 11.

F. INTER-ANNOTATOR AGREEMENT

The annotation process is prone to biases because annotators can have different perspectives and opinions about the sentiment of a tweet. To observe the inter-rater reliability and measure the consistency between annotators, we used Cohen's kappa coefficient measurement, Equation 2 [46]. Kappa, κ, is one of the most commonly used measures for agreement between two annotators on categorical variables. It corrects for agreement by chance and is widely used in computational linguistic annotation tasks [47]:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{2}$$

where $p_o$is the observed agreement among annotators and $p_e$is the expected agreement by chance. When$\kappa = 1$, there is complete agreement between annotators. If agreement is random, then $\kappa = 0$, while negative values indicate that agreement is less than random. Equation 3 depicts the calculation of $p_e$as follows:

$$p_e = \frac{1}{N^2} \sum_i ni1 \cdot ni2, \tag{3}$$

TABLE 12. Intepretation of cohen's kappa.

κ	Strength of agreement
< 0.0	Poor
0.0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1	Almost perfect

where N is the total number of tweets, i are the labels, $ni1$ is the number of times that the first annotator assigned label i to tweets, and $ni2$ is the number of times that the second annotator assigned label i to tweets.

The calculated kappa measure, κ, for the six labels between our pairs of annotators is $\kappa = 0.6526$. According to [48], this κ value is interpreted as indicative of substantial agreement between annotators. Table 12 shows this interpretation of Cohen's kappa values.

IV. ANNOTATION CHALLENGES

During the annotation of the data, annotators faced several challenges. These can be divided into operational and linguistic challenges. The main operational challenges are discussed first. Initially, there was an overestimation of the ability of annotators to annotate such a large corpus. The total duration of the annotation stage was four months. At the beginning, annotators faced a problem in finishing the annotation of their allocated tweets on time. This problem was tackled by defining a minimum daily target for each annotator. This was set at a minimum of 350 tweets per day, which encouraged annotators to stay on track. Additionally, the use of the annotation interface was considered to be time-consuming due to the switching between typing on the keyboard and using the mouse to select.

Other than the abovementioned operational challenges of annotating such a large corpus, the main challenge lies in the fact that the Gulf dialect is non-standardized. Hence, there were many obscure words and much jargon that annotators were not familiar with. This led to several linguistic challenges that complicated their decision making. These challenges are explained in the following points, while examples of each challenge are presented in Table 13.

1. The first challenge was maintaining objectivity, especially when annotating tweets in political and sports domains. This confused some annotators when categorizing tweets into positive or negative because they found themselves supporting one view over another. The annotators were asked to adopt the stance of the tweet's author and to judge the tweet accordingly.
2. The second challenge was use of jargon. Examples of this challenge are words in the sports domain. These terms were initially unknown to annotators and looking them up extended their decision-making process.

TABLE 13. Examples of linguistic challenges.

Challenge	Example
Maintaining objectivity	الهلال_الشباب لو يسجل مهاجموا الزعيم ربع الفرص الذهبيه لخسرت الانديه بالاربعه والخمسه والسته يالله العوض بالجاي
Jargon words	دفع رباعي، هاترك
Obscure dialectal words	اهب، كمخه، ملطشه، خنفشاري، داهيه
New nomenclature	جحفله
Non-Arabic language	هاردلك، قود لك، برافو
Dual tweets	أسعد يوم لما مات الحثالة
Multi-subject tweets	العساف منحه الراتبين لموظفي الدوله من اهم اسباب زياده مصروفات العام الحالي رفع_اسعار_البنزين
Spelling and grammatical mistakes	تعليق_الدراسه لازم نت عاطف م ع اخوان نا في مكه وج ده م

3. The third challenge was use of obscure dialectal words that are infrequent in certain regions. They also had to be looked up.
4. The fourth challenge was new nomenclatures, especially ones that were created and extensively used to indicate sentimental references. The word *jahfalah*, for example, was created in 2015 and is based on the name of a football player who scored a goal seconds before the end of a match, surprising the opposing team and winning the game. Since then, the word has been used as both a noun and a verb to express shocking and unexpected victories.
5. The fifth challenge was use of non-Arabic words—but written in Arabic script—to express meaning. These have no standard spelling and can be ambiguous.
6. The sixth challenge was dual sentiment, meaning that a tweet holds two polarities. This was the motivation behind creating a new label, called *both*.
7. The seventh challenge was multi-subject tweets, which refers to the fact that a tweet contains reference to more than one topic. Specifying a topic is important for expressing sentiment in a domain. Nevertheless, such tweets were rare.
8. The eighth challenge was spelling and grammatical errors, which can change the meaning of a tweet or make it ambiguous.

V. BENCHMARK EXPERIMENTS

In this section, we present the results of training and testing a classifier on datasets that were created from the corpus. The aim is establish a benchmark for researchers who wish to use them in their research. We performed three-way classification on the datasets so they include only the tweets annotated as positive, negative, or neutral. We selected these three labels out of the five labels to report the classification results to provide researchers with comparable measures to existing sentiment analysis datasets.

We created two datasets. The first is an unbalanced one, which has a different number of tweets for each label and

TABLE 14. Unbalanced dataset statistics.

	Training				Development				Testing			
	Total	Neg	Neut	Pos	Total	Neg	Neut	Pos	Total	Neg	Neut	Pos
Tech	4,054	2,727	971	356	496	341	122	43	495	342	121	42
Sport	21,197	6,608	4,824	9,765	2,647	825	602	1,220	2,645	825	600	1,220
Social	12,478	4,187	5,976	2,315	1,559	523	746	290	1,559	523	746	290
Political	7,707	3,065	3,214	1,428	963	382	403	178	962	382	401	179
Total	45,436	16,587	14,985	13,864	5,675	2,071	1,873	1,731	5,671	2,072	1,868	1,731

TABLE 15. Balanced dataset statistics.

Training				Development				Testing			
Total	Neg	Neut	Pos	Total	Neg	Neut	Pos	Total	Neg	Neut	Pos
1068	356	356	356	129	43	43	43	129	43	43	43

TABLE 16. Classification results on the development partition for the unbalanced dataset.

	Negative			Neutral			Positive			F1 Avg	Acc
	P	R	F1	P	R	F1	P	R	F1		
Tech.	**0.87**	**0.98**	**0.92**	**0.88**	0.72	**0.79**	0.87	0.47	0.61	**0.86**	**0.87**
Sports	0.75	0.80	0.77	0.64	0.55	0.59	0.85	**0.87**	**0.86**	0.77	0.77
Social	0.61	0.68	0.64	0.71	**0.78**	0.74	0.81	0.43	0.56	0.67	0.68
Political	0.67	0.81	0.73	0.78	0.72	0.75	**0.94**	0.66	0.78	0.75	0.75
All	0.73	0.80	0.76	0.72	0.70	0.71	0.86	0.77	0.81	0.76	0.76

TABLE 17. Classification results on the testing partition for the unbalanced dataset.

	Negative			Neutral			Positive			F1 Avg	Acc
	P	R	F1	P	R	F1	P	R	F1		
Tech.	**0.83**	**0.97**	**0.90**	**0.88**	0.63	0.73	0.89	0.40	0.56	**0.83**	**0.84**
Sports	0.78	0.74	0.76	0.66	0.62	0.64	0.81	**0.86**	**0.83**	0.77	0.77
Social	0.69	0.70	0.70	0.72	**0.84**	**0.78**	0.81	0.44	0.57	0.71	0.72
Political	0.71	0.82	0.76	0.73	0.77	0.75	**0.90**	0.51	0.65	0.75	0.74
All	0.75	0.79	0.77	0.71	0.74	0.73	0.83	0.75	0.79	0.76	0.76

TABLE 18. Classification results on the 80:20 unbalanced dataset.

	Negative			Neutral			Positive			F1 Avg	Acc
	P	R	F1	P	R	F1	P	R	F1		
Tech.	0.69	0.82	0.75	0.76	0.74	0.75	**0.92**	0.59	0.72	0.74	0.74
Sports	0.65	0.69	0.67	0.72	**0.81**	0.76	0.81	0.44	0.57	0.69	0.70
Social	0.76	0.77	0.77	0.65	0.59	0.62	0.83	**0.86**	**0.85**	0.77	0.77
Political	**0.85**	**0.98**	**0.91**	**0.88**	0.68	**0.77**	0.88	0.44	0.58	**0.86**	**0.85**
All	0.74	0.80	0.77	0.71	0.72	0.72	0.85	0.76	0.80	0.76	0.76

domain. It comprises a total of 56,782 tweets and its detailed statistics are shown in Table 14. The second dataset is balanced, to reduce the bias towards larger classes. The dataset contains a total of 6,630 tweets, and has the same number of tweets for each domain and label. Its statistics are shown in Table 15.

The experiments were implemented in Python using the SVM classifier from Scikit Learn. TF-IDF was used to represent the text. Results are presented in the following sub-sections, and for all the datasets we trained and tested five classifiers, one classifier for each of the four domains and a general classifier on the whole dataset.

A. UNBALANCED DATASET BENCHMARK EXPERIMENTS

As shown in Table 14, the unbalanced dataset was partitioned into an 80:10:10 ratio—for training, development, and testing, respectively. Table 16 and Table 17 show the classification results, where the highest F1 and accuracy were achieved in the technology domain.

Furthermore, we provided an alternative partition with an 80:20 ratio for training and testing, where the testing partition combines the development and testing partitions. The results are given in Table 18, where the highest results are in the political domain.

B. BALANCED DATASET BENCHMARK EXPERIMENTS

As mentioned above, Table 15 shows the statistics for the balanced dataset, which was first partitioned into an 80:10:10 ratio and then into an 80:20 ratio. Table 19 and Table 20 show the results for the 80:10:10 partitions on the development and testing partitions, respectively. Similar to the results for the unbalanced dataset, the F1 and accuracy measure results are the highest in the technology domain. Table 21 displays the results for the 80:20 partition. They are consistent, meaning that the highest results are in the technology domain.

TABLE 19. Classification results on the development partition for the balanced dataset.

	Negative			Neutral			Positive			F1	Acc
	P	R	F1	P	R	F1	P	R	F1	Avg	
Tech.	**0.73**	**0.77**	**0.75**	**0.82**	**0.77**	**0.80**	0.75	**0.77**	**0.76**	**0.77**	**0.77**
Sports	0.55	0.63	0.59	0.40	0.44	0.42	0.85	0.65	0.74	0.58	0.57
Social	0.47	0.53	0.50	0.55	0.63	0.59	0.61	0.44	0.51	0.53	0.53
Political	0.55	**0.77**	0.64	0.72	0.65	0.68	**0.90**	0.63	0.74	0.69	0.68
All	0.50	0.74	0.60	0.67	0.56	0.61	0.86	0.58	0.69	0.63	0.63

TABLE 20. Classification results on the testing partition for the balanced dataset.

	Negative			Neutral			Positive			F1	Acc
	P	R	F1	P	R	F1	P	R	F1	Avg	
Tech.	**0.67**	0.72	**0.70**	**0.86**	0.72	**0.78**	0.72	**0.79**	**0.75**	**0.74**	**0.74**
Sports	**0.67**	0.65	0.66	0.59	0.74	0.66	0.73	0.56	0.63	0.65	0.65
Social	0.58	0.51	0.54	0.50	**0.77**	0.61	0.73	0.44	0.55	0.57	0.57
Political	0.55	**0.77**	0.64	0.68	0.65	0.67	**0.93**	0.60	0.73	0.68	0.67
All	0.55	0.60	0.58	0.55	0.49	0.52	0.61	0.63	0.62	0.57	0.57

TABLE 21. Classification results on the 80:20 balanced dataset.

	Negative			Neutral			Positive			F1	Acc
	P	R	F1	P	R	F1	P	R	F1	Avg	
Tech.	**0.70**	0.74	**0.72**	**0.83**	**0.74**	**0.79**	0.73	**0.76**	**0.75**	**0.75**	**0.75**
Sports	0.60	0.64	0.62	0.50	0.59	0.55	0.79	0.60	0.68	0.61	0.62
Social	0.53	0.55	0.54	0.53	0.70	0.60	0.68	0.44	0.54	0.56	0.56
Political	0.55	**0.77**	0.64	0.70	0.66	0.68	**0.91**	0.60	0.73	0.68	0.68
All	0.54	0.69	0.60	0.62	0.55	0.58	0.72	0.60	0.66	0.61	0.61

C. REFLECTING ON THE RESULTS

From the perspective of domain-dependent sentiment analysis, it is important to study the performance of domain specific classifiers compared to a general classifier. In the experiments preformed we can see that in the unbalanced datasets the domain specific classifiers for both the sport and technical domain outperformed the general classifier, for the 80:10:10 partition. In the 80:20 classifier the political classifier outperformed the general classifier while the technical and sports were close.

In the balanced dataset the technical and political classifiers consistently performed better than the general classifier. However, the performance of the sport classifier varied compared to the general classifier, in Table 20 and 21 it was similar, however in Table 19 it was slightly worse.

On the other hand, the social classifier performed worse than the general classifier on all the datasets. This could be a consequence of the social domain containing less domain specific phrases, and therefore may contain several subdomains. Moreover, the social issues discussed in the tweets affected diverse demographic groups in the community with varied interests, and as a result expressed their opinions in different manners.

VI. CONCLUSION

There is a lack of corpora provided for the study of dialectal Arabic, even more so is the lack of resources to study domain dependent sentiment analysis. This research provides a gold-standard sentiment-annotated multi-domain Arabic corpus in the Gulf dialect. It contains a total of 61,353 tweets, with a total of 840,702 tokens. Each tweet was manually annotated by two annotators, resulting in substantial agreement as indicated by a kappa coefficient of 0.65. The tweets were collected from four domains: political, social, sports, and technology. As a result, the corpus is a collection of four domain specific corpora, thus providing an essential resource for domain dependent sentiment analysis. Furthermore, four sentiment lexicons were manually created from these domains. In this paper, we presented the statistics about the overlap in the lexicons' entries, providing evidence for contextual polarity of certain words.

We also observed the prevalence of negative tweets in our corpus and in other corpora presented in the literature. This raises interesting questions. For instance, could this be explained by the negativity effect? Is this observed in other languages? Does the platform (Twitter in our case) facilitate this trend? And, in a wider sense, how do social media platforms compare in facilitating the negativity effect?

Furthermore, to establish a baseline for interested researchers in the field, this study provides the results of the sentiment classification that was performed on the corpus.

REFERENCES

[1] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *What Top 200 Most Spoken Languages*. Dallas, TX, USA: SIL International, 2021. Accessed: Mar. 22, 2021. [Online]. Available: https://www.ethnologue.com/statistics/size

[2] I. Guellil and F. Azouaou, "Arabic dialect identification with an unsupervised learning (based on a lexicon) application case: Algerian dialect," in *Proc. IEEE Int. Conf. Comput. Sci. Eng. (CSE)*, Aug. 2016, pp. 724–731, doi: 10.1109/CSE-EUC-DCABES.2016.268.

[3] *Most Popular Social Networks Worldwide as of July 2020, Ranked by Number of Active Users*. Accessed: Mar. 22, 2021. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[4] *Most Tweeted Language by World Leaders as of June 2014*. Accessed: Mar. 22, 2021. [Online]. Available: https://www.statista.com/statistics/348508/most-tweeted-language-world-leaders/

[5] S. O. Alhumoud, M. I. Altuwaijri, T. M. Albuhairi, and W. M. Alohaideb, "Survey on Arabic sentiment analysis in Twitter," *Int. Sci. Index*, vol. 9, no. 1, pp. 364–368, Feb. 2015.

[6] *MARSA Dataset*. Accessed: Mar. 22, 2021. [Online]. Available: https://github.com/imamu-asa/ASA

[7] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh, and M. N. Al-Kabi, "A comprehensive survey of Arabic sentiment analysis," *Inf. Process. Manag.*, vol. 56, no. 2, pp. 320–342, Mar. 2019.

[8] G. Badaro, R. Baly, H. Hajj, W. El-Hajj, K. B. Shaban, N. Habash, A. Al-Sallab, and A. Hamdi, "A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 18, no. 3, p. 27, May 2019, doi: 10.1145/3295662.

[9] O. Oueslati, E. Cambria, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," *Future Gener. Comput. Syst.*, vol. 112, pp. 408–430, Nov. 2020, doi: 10.1016/j.future.2020.05.034.

[10] M. Abdul-Mageed and M. T. Diab, "AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis," in *Proc. 8th Int. Conf. Lang. Resour. Eval. Conf. (LREC)*, Istanbul, Turkey, 2012, pp. 3907–3914. Accessed: Jun. 16, 2014. [Online]. Available: http://www.seas.gwu.edu/~mtdiab/files/publications/refereed/13.pdf

[11] M. A. Aly and A. F. Atiya, "LABR: A large scale Arabic book reviews dataset," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, Sofia, Bulgaria, Aug. 2013, pp. 494–498.

[12] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel Arabic-reviews dataset construction for sentiment analysis applications," in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds. Cham, Switzerland: Springer, 2018, pp. 35–52.

[13] A. Elnagar and O. Einea, "BRAD 1.0: Book reviews in Arabic dataset," in *Proc. IEEE/ACS 13th Int. Conf. Comput. Syst. Appl. (AICCSA)*, New York, NY, USA, Nov. 2016, pp. 1–8.

[14] H. ElSahar and S. R. El-Beltagy, "Building large Arabic multi-domain resources for sentiment analysis," in *Proc. Comput. Linguistics Intell. Text Process.*, Cham, Switzerland, 2015, pp. 23–34, doi: 10.1007/978-3-319-18117-2_2.

[15] M. Rushdi-saleh, M. T. Martín-valdivia, L. A. Ureña-lópez, and J. M. Perea-ortega, "OCA: Opinion corpus for Arabic," *J. Assoc. Inf. Sci. Technol.*, vol. 62, no. 10, pp. 2045–2054, 2011, doi: 10.1002/asi.21598.

[16] E. Refaee and V. Rieser, "An Arabic Twitter corpus for subjectivity and sentiment analysis," in *Proc. LREC*, Reykjavik, Iceland, 2014, pp. 2268–2273. Accessed: Jun. 16, 2014. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/317_Paper.pdf

[17] M. Nabil, M. Aly, and A. Atiya, "ASTD: Arabic sentiment tweets dataset," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2515–2519.

[18] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "AraSenTi-tweet: A corpus for Arabic sentiment analysis of Saudi tweets," *Proc. Comput. Sci.*, vol. 117, pp. 63–72, Jan. 2017, doi: 10.1016/j.procs.2017.10.094.

[19] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proc. 11th Int. Workshop Semantic Eval.*, Vancouver, BC, Canada, Aug. 2017, pp. 502–518, Accessed: May 19, 2018. [Online]. Available: http://www.aclweb.org/anthology/S17-2088

[20] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proc. 12th Int. Workshop Semantic Eval.*, New Orleans, LA, USA, 2018, pp. 1–17, doi: 10.18653/v1/S18-1001.

[21] A. A. Elmadany, H. Mubarak, and W. Magdy, "ArSAS?: An Arabic speech-act and sentiment corpus of tweets," *Proc. 11th Int. Conf. Lang. Resour. Eval.*, 2018, p. 20.

[22] C. Qwaider, S. Chatzikyriakidis, and S. Dobnik, "Can modern standard Arabic approaches be used for Arabic dialects? Sentiment analysis as a case study," in *Proc. 3rd Workshop on Arabic Corpus Linguistics*, Cardiff, U.K., Jul. 2019, pp. 40–50, Accessed: Jul. 18, 2020. [Online]. Available: https://www.aclweb.org/anthology/W19-5606

[23] K. A. Kwaik, S. Chatzikyriakidis, S. Dobnik, M. Saad, and R. Johansson, "An Arabic tweets sentiment analysis dataset (ATSAD) using distant supervision and self training," in *Proc. 4th Workshop Open-Source Arabic Corpora Process. Tools, Shared Task Offensive Lang. Detection*, Marseille, France, May 2020, pp. 1–8. Accessed: Jul. 6, 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.osact-1.1

[24] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in *Proc. Appl. Electr. Eng. Comput. Technol. (AEECT)*, 2013, pp. 1–6, Accessed: Jun. 16, 2014. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6716448

[25] M. Abdul-Mageed, M. T. Diab, and M. Korayem, "Subjectivity and sentiment analysis of modern standard Arabic," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2, 2011, pp. 587–591, Accessed: Jun. 16, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002851

[26] M. Abdul-Mageed and M. Diab, "Toward building a large-scale Arabic sentiment lexicon," in *Proc. 6th Int. Global WordNet Conf.*, 2012, pp. 18–22.

[27] M. Al-Kabi, A. Gigieh, I. Alsmadi, H. Wahsheh, and M. Haidar, "An opinion analysis tool for colloquial and standard Arabic," in *Proc. 4th Int. Conf. Inf. Commun. Syst. (ICICS)*, 2013, pp. 23–25.

[28] K. Al-Rowaily, M. Abulaish, N. Al-Hasan Haldar, and M. Al-Rubaian, "BiSAL—A bilingual sentiment analysis lexicon to analyze Dark web forums for cyber security," *Digit. Invest.*, vol. 14, pp. 53–62, Sep. 2015.

[29] S. R. El-Beltagy, "NileULex: A phrase and word level sentiment lexicon for Egyptian and modern standard Arabic," in *Proc. Tenth Int. Conf. Lang. Resour. Eval. (LREC)*, Portoro, Slovenia, May 2016, pp. 2900–2905. Accessed: Jul. 4, 2020. [Online]. Available: https://www.aclweb.org/anthology/L16-1463

[30] H. S. Ibrahim, S. M. Abdou, and M. Gheith, "Idioms-proverbs lexicon for modern standard Arabic and colloquial sentiment analysis," *Int. J. Comput. Appl.*, vol. 118, no. 11, pp. 26–31, May 2015, doi: 10.5120/20790-3435.

[31] A. Mourad and K. Darwish, "Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs," in *Proc. WASSA*, 2013, p. 55.

[32] G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj, "A large scale Arabic sentiment lexicon for Arabic opinion mining," in *Proc. ANLP*, 2014, p. 165.

[33] F. H. H. Mahyoub, M. A. Siddiqui, and M. Y. Dahab, "Building an Arabic sentiment lexicon using semi-supervised learning," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 26, no. 4, pp. 417–424, Dec. 2014.

[34] R. Eskander and O. Rambow, "SLSA: A sentiment lexicon for standard Arabic," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 2545–2550.

[35] T. Al-Moslmi, M. Albared, A. Al-Shabi, N. Omar, and S. Abdullah, "Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis," *J. Inf. Sci.*, vol. 44, no. 3, pp. 345–362, 2017.

[36] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman, "AraSenTi: Large-scale Twitter-specific Arabic sentiment lexicons," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Berlin, Germany, Aug. 2016, pp. 697–705, doi: 10.18653/v1/P16-1066.

[37] N. Al Twairesh, M. Al Tuwaijri, A. Al Moammar, and S. Al Humoud, "Arabic spam detection in Twitter," in *Proc. 2nd Workshop Arabic Corpora Process. Tools*, 2016, p. 38.

[38] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs, "Bad is stronger than good," *Rev. Gen. Psychol.*, vol. 5, no. 4, pp. 323–370, 2001.

[39] E. H. R. Rho and M. Mazmanian, "Political hashtags & the lost art of democratic discourse," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, Apr. 2020, pp. 1–13, doi: 10.1145/3313831.3376542.

[40] M. Jenders, G. Kasneci, and F. Naumann, "Analyzing and predicting viral tweets," in *Proc. 22nd Int. Conf. World Wide Web*, New York, NY, USA, 2013, pp. 657–664, doi: 10.1145/2487788.2488017.

[41] S. Alhumoud, "Twitter analysis for intelligent transportation," *Comput. J.*, vol. 62, no. 11, pp. 1547–1556, Nov. 2019, doi: 10.1093/comjnl/bxy129.

[42] C. Collins, S. Hasan, and S. Ukkusuri, "A novel transit rider satisfaction metric: Rider sentiments measured from online social media data," *J. Public Transp.*, vol. 16, no. 2, pp. 21–45, Jun. 2013.

[43] M. A. Stefanone, G. D. Saxton, M. J. Egnoto, W. Wei, and Y. Fu, "Image attributes and diffusion via Twitter: The case of #guncontrol," in *Proc. 48th Hawaii Int. Conf. Syst. Sci.*, Jan. 2015, pp. 1788–1797.

[44] R. Pfitzner, A. Garas, and F. Schweitzer, "Emotional divergence influences information spreading in Twitter," in *Proc. ICWSM*, vol. 12, 2012, pp. 1–5.

[45] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytol.*, vol. 11, no. 2, pp. 37–50, 1912.
[46] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.
[47] B. D. Eugenio and M. Glass, "The kappa statistic: A second look," *Comput. Linguistics*, vol. 30, pp. 95–101, Mar. 2004.
[48] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 4, pp. 159–174, Mar. 1977.

AREEB ALOWISHEQ received the B.Sc. and M.Sc. degrees from King Saud University, Riyadh, Saudi Arabia, and the Ph.D. degree from the University of Southampton, U.K. She is currently an Assistant Professor with the Computer Science Department, College of Computer and Information Sciences, Imam Muhammad Ibn Saud Islamic University (IMSIU), Riyadh. She is also an AI Consultant at the National Center for Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA). Her research interests include natural language processing, data science, and linked data.

NORA AL-TWAIRESH received the Ph.D. degree in computer science from King Saud University (KSU). She is currently an Assistant Professor with the Information Technology Department, College of Computer and Information Sciences, KSU. She is also a member of the iWAN Research Group, KSU, and the STC Artificial Intelligence Research Chair at KSU. She is also an AI Consultant at the National Center for Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA). She has published several research articles. Her research interests include natural language processing, data science, and social computing. She has served as a program committee member in many national and international conferences and as a reviewer for several journals.

MAWAHEB ALTUWAIJRI received the B.Sc. degree in computer science from IMSIU, Riyadh, Saudi Arabia. She worked as a System Developer with Tatweer Educational Technologies (TETCO). She currently works as a System Developer at the Data Center, Princess Nourah Bint Abdulrahman University.

AFNAN ALMOAMMAR received the M.Sc. degree in computer science from King Saud University, Riyadh, Saudi Arabia, in 2020. She currently works as a Teacher Assistant at IMSIU. Her research interests include natural language processing, machine learning, and data mining.

ALHANOUF ALSUWAILEM received the B.Sc. degree in information systems from IMSIU, Riyadh, Saudi Arabia, where she is currently pursuing the master's degree in information systems. She is also a Teaching Assistant at IMSIU. She worked as a Data scientist for a period of three years at General Organization for Social Insurance (GOSI) and mainly developed fraud detection system to predict suspicious events. In 2019, she moved to General Authority for Zakat and Tax as a Data Specialist. Her research interests include financial fraud, natural language processing, and machine learning.

TARFA ALBUHAIRI received the B.Sc. degree in computer science from IMSIU, Riyadh, Saudi Arabia, in 2015. Since 2015, she has been working in integration for technical professional services (ITPS) as a Web Developer, where she is currently an Associate Technical Consultant at Elm for Information Security Company. Her research interests include data mining, big data, and web development.

WEJDAN ALAHAIDEB received the B.Sc. degree in computer sciences from IMSIU, Riyadh, Saudi Arabia. She is currently working as an IT Specialist (Full Stack Developer) at the Health Sciences Research Center, Princess Nourah Bint Abdulrahman University.

SARAH ALHUMOUD received the Ph.D. degree in wireless networks from the University of Glasgow, U.K., in 2011. She was appointed a Research Fellow with the Computer Science and Artificial Intelligence Laboratory, MIT, USA, in the field of Arabic NLP. She is currently an Associate Professor with the Department of Computing Science, IMSIU, Riyadh. She has taught several different B.Sc. and M.Sc. courses while being an academic in several universities over the past 18 years. She was the Principal Investigator of the Arabic Sentiment Analysis Research Group, IMSIU. She has more than 30 published papers, articles, and a book.