MARSA: Multi-domain Arabic Resources for Sentiment Analysis

AREEB ALOWISHEQ¹,³, NORA AL-TWAIRESH²,³, MAWAHEB ALTUWAIJRI⁴, AFNAN ALMOAMMAR¹, ALHANOUF ALSUWAILEM¹, TARFA ALBUHAIRI⁵, WEJDAN ALAHAIDEB⁴ AND, SARAH ALHUMOUD¹
¹College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh 13318, Saudi Arabia.
²STC’s Artificial Intelligence Chair, Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh 11495, Saudi Arabia.
³National Center for Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA), Riyadh 12391, Saudi Arabia.
⁴Princess Nourah Bint Abdulrahman University, Riyadh 11564, Saudi Arabia.
⁵Elm, Riyadh 12333, Saudi Arabia.

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

The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University for funding and supporting this work.

ABSTRACT

The Arabic language has many spoken dialects. However, until recently, it was primarily written in Modern Standard Arabic (MSA), which is considered to be the formal variant of Arabic. Social media platforms have changed the face of written Arabic because their 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.

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 convey 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 I 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 onAWN 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.

| 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 | 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 |

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 creation 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 subsections explain these four different stages in detail.

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 II. 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,إ) were converted into (ا), the letter ta (ت) was converted to (ت).
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 II: CORPUS DESCRIPTION BEFORE PREPROCESSING

| Domain   | Date       | Keyword/Hashtag | Description                      | Count  |
|----------|------------|-----------------|----------------------------------|--------|
| Tech     | Nov, STC, MBC |                | about internet connection weakness | 34,242 |
| Sport    | Nov, Feb   |                 | Hashtags created and active during football matches. | 244,221 |
| Social   | Nov, Dec   |                 | Royal orders affecting budget, income, prices. | 251,000 |
| Political| Nov, Jun, Feb|                | Political events that included news about terrorism or military activities. It also included hashtags initiated by Saudi government opponents. | 131,860 |

### TABLE III: EXAMPLES OF TWEETS FROM DIFFERENT LABELS

| Label | Example |
|-------|---------|
| Positive | Thank God congratulations Al-Zaeem’s (Al-Hilal) fans, wishing the cup God willing Al-Hilal_Al-Taawoun |
| Neutral | Al-Taawoun gives and should thank luck refereeing in my opinion is the main reason in not winning hard luck sukari Al-Qassim (Al-Taawoun) Congratulations Al-Hilal-Al-Hilal_Qualifying |
| Sarcasm | A simple annotation interface was created. The interface is shown in Figure I. 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 IV. |

A simple annotation interface was created. The interface is shown in Figure I. 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 IV.
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 V. There was a total of 11 annotators and each tweet was annotated by 2 annotators. This process took approximately four months.

### Table V: 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 VI. The table shows the number of tweets classified by domain and the six labels from

Table III. 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 VI, 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.

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 #أعدام_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 I, 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 VII 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 VI.

Table VIII, 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 IX. 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]:

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

The overlap ratios in Table IX 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 X. Examples of these words are explained below and shown in Table XI.

In Table XI, 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.

The word ملوح (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 XI.

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 “مالح” (spoil) was used positively in the first tweet and negatively in the second, as shown in Example 3 in Table XI.
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, $\kappa$, 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]:

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

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 $\kappa = 0$, where negative values indicate that agreement is less than random. Equation 3 depicts the calculation of $p_e$ as follows:

$$p_e = \frac{1}{N^2} \sum_i n_i1 \cdot n_i2, \tag{3}$$

where $N$ is the total number of tweets, $i$ are the labels, $n_i1$ is the number of times that the first annotator assigned label $i$ to tweets, and $n_i2$ is the number of times that the second annotator assigned label $i$ to tweets.

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

| $\kappa$ | 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        |

### 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 XIII.

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.

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.
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 domain. It comprises a total of 56,782 tweets and its detailed statistics are shown in Table XIV. 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 XV.

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 subsections, 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 XIV, the unbalanced dataset was partitioned into an 80:10:10 ratio—for training, development, and testing, respectively. Table XVI and Table XVII 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 XVIII, where the highest results are in the political domain.

B. BALANCED DATASET BENCHMARK EXPERIMENTS

As mentioned above, Table XV 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 XIX and Table XX 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 XXI displays the results for the 80:20 partition. They are consistent, meaning that the highest results are in the technology domain.

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 XX and XXI it was similar, however in Table XIX 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 the tweets were about, affected a majority of the community who belonged to diverse demographic groups with varied interests, and as a result expressed their opinions in different manners.
TABLE XIV: UNBALANCED DATASET STATISTICS

|        | Training |          |          | Testing |          |          |
|--------|----------|----------|----------|---------|----------|----------|
|        | Total    | Neg      | Neut     | Pos     | Total    | Neg      | Neut     | Pos     |
| Tech   | 4,054    | 2,727    | 971      | 356     | 496      | 341      | 122      | 43      |
| Sport  | 21,197   | 6,608    | 4,824    | 9,765   | 2,647    | 825      | 602      | 1,220   |
| Social | 12,478   | 4,187    | 5,976    | 2,315   | 1,559    | 523      | 746      | 290     |
| Political | 7,707 | 3,063    | 3,214    | 1,428   | 963      | 382      | 403      | 178     |
| Total  | 45,436   | 16,587   | 14,985   | 13,864  | 5,675    | 2,071    | 1,873    | 1,731   |

TABLE XV: BALANCED DATASET STATISTICS

|        | Training |          |          | Testing |          |          |
|--------|----------|----------|----------|---------|----------|----------|
|        | Total    | Neg      | Neut     | Pos     | Total    | Neg      | Neut     | Pos     |
|        | 1,068    | 356      | 356      | 356     | 129      | 43       | 43       | 43      |

TABLE XVI: CLASSIFICATION RESULTS ON THE DEVELOPMENT PARTITION FOR THE UNBALANCED DATASET

|        | Negative | Neutral | Positive | F1 Avg | Acc  |
|--------|----------|---------|----------|--------|------|
| P      | R        | F1      | P        | R      | F1   |
| Tech   | 0.87     | 0.98    | 0.92     | 0.88   | 0.72 | 0.79 |
| Sports | 0.75     | 0.80    | 0.77     | 0.64   | 0.55 | 0.59 |
| Social | 0.61     | 0.68    | 0.64     | 0.71   | 0.78 | 0.74 |
| Political | 0.67 | 0.81    | 0.73     | 0.78   | 0.72 | 0.75 |
| All    | 0.73     | 0.80    | 0.76     | 0.72   | 0.70 | 0.71 |

TABLE XVII: CLASSIFICATION RESULTS ON THE TESTING PARTITION FOR THE UNBALANCED DATASET

|        | Negative | Neutral | Positive | F1 Avg | Acc  |
|--------|----------|---------|----------|--------|------|
| P      | R        | F1      | P        | R      | F1   |
| Tech   | 0.83     | 0.97    | 0.90     | 0.88   | 0.63 | 0.73 |
| Sports | 0.78     | 0.74    | 0.76     | 0.66   | 0.62 | 0.64 |
| Social | 0.69     | 0.70    | 0.70     | 0.72   | 0.84 | 0.78 |
| Political | 0.71 | 0.82    | 0.76     | 0.73   | 0.77 | 0.75 |
| All    | 0.75     | 0.79    | 0.77     | 0.71   | 0.74 | 0.73 |

TABLE XVIII: CLASSIFICATION RESULTS ON THE 80:20 UNBALANCED DATASET

|        | Negative | Neutral | Positive | F1 Avg | Acc  |
|--------|----------|---------|----------|--------|------|
| P      | R        | F1      | P        | R      | F1   |
| Tech   | 0.69     | 0.82    | 0.75     | 0.76   | 0.74 | 0.75 |
| Sports | 0.65     | 0.69    | 0.67     | 0.72   | 0.81 | 0.76 |
| Social | 0.76     | 0.77    | 0.77     | 0.65   | 0.59 | 0.62 |
| Political | 0.85 | 0.98    | 0.91     | 0.88   | 0.68 | 0.77 |
| All    | 0.74     | 0.80    | 0.77     | 0.71   | 0.72 | 0.72 |

TABLE XIX: CLASSIFICATION RESULTS ON THE DEVELOPMENT PARTITION FOR THE BALANCED DATASET

|        | Negative | Neutral | Positive | F1 Avg | Acc  |
|--------|----------|---------|----------|--------|------|
| P      | R        | F1      | P        | R      | F1   |
| Tech   | 0.73     | 0.77    | 0.75     | 0.82   | 0.77 | 0.80 |
| Sports | 0.55     | 0.63    | 0.59     | 0.40   | 0.44 | 0.42 |
| Social | 0.47     | 0.53    | 0.50     | 0.55   | 0.63 | 0.59 |
| Political | 0.55 | 0.77    | 0.64     | 0.72   | 0.65 | 0.68 |
| All    | 0.50     | 0.74    | 0.60     | 0.67   | 0.56 | 0.61 |

TABLE XX: CLASSIFICATION RESULTS ON THE TESTING PARTITION FOR THE BALANCED DATASET

|        | Negative | Neutral | Positive | F1 Avg | Acc  |
|--------|----------|---------|----------|--------|------|
| P      | R        | F1      | P        | R      | F1   |
| Tech   | 0.67     | 0.72    | 0.70     | 0.86   | 0.72 | 0.78 |
| Sports | 0.67     | 0.65    | 0.66     | 0.59   | 0.74 | 0.66 |
| Social | 0.58     | 0.51    | 0.54     | 0.50   | 0.77 | 0.61 |
| Political | 0.55 | 0.77    | 0.64     | 0.68   | 0.65 | 0.67 |
| All    | 0.55     | 0.60    | 0.58     | 0.55   | 0.49 | 0.52 |

TABLE XXI: CLASSIFICATION RESULTS ON THE 80:20 BALANCED DATASET

|        | Negative | Neutral | Positive | F1 Avg | Acc  |
|--------|----------|---------|----------|--------|------|
| P      | R        | F1      | P        | R      | F1   |
| Tech   | 0.70     | 0.74    | 0.72     | 0.83   | 0.74 | 0.79 |
| Sports | 0.60     | 0.64    | 0.62     | 0.50   | 0.59 | 0.55 |
| Social | 0.53     | 0.55    | 0.54     | 0.53   | 0.70 | 0.60 |
| Political | 0.55 | 0.77    | 0.64     | 0.70   | 0.66 | 0.68 |
| All    | 0.54     | 0.69    | 0.60     | 0.62   | 0.55 | 0.58 |
V. CONCLUSIONS

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.

VI. REFERENCES

[1] D. M. Eberhard, G. F. Simons, and C. D. Fennig (eds.). “What are the top 200 most spoken languages?” (Ethnologue: Languages of the World. 24th ed.), Dallas, TX, USA: SIL International, 2021. [Online]. Available: https://www.ethnologue.com/statistics/size. [Accessed: 22-March-2021]

[2] I. Guellil and F. Azouaou, “Arabic Dialect Identification with an Unsupervised Learning (Based on a Lexicon). Application Case: ALGERIAN Dialect,” in 2016 IEEE Int. Conf. Computational Science and Engineering (CSE) and IEEE Int. Conf. Embedded and Ubiquitous Computing (EUC) and 15th Int. Symp. Distributed Computing and Applications for Business Engineering (DCABES), 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.” [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. [Accessed: 22-March-2021]

[4] “Most tweeted language by world leaders as of June 2014.” [Online]. Available: https://www.statista.com/statistics/348508/most-tweeted-language-world-leaders/. [Accessed: 22-March-2021]

[5] S. Alhumoud, M. Altuwaijri, T. Albuhairi, and W. Alohaideb, “Survey on Arabic Sentiment Analysis in Twitter,” Int. J. Comput. Inf. Eng., vol. 9, no. 1, pp. 364–368, 2015.

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

[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, 2019.

[8] G. Badaro et al., “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. Proc., vol. 18, no. 3, p. 27:1–27:52, 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. Eighth Int. Conf. Language Resources and Evaluation Conference (LREC), İstanbul, Turkey, 2012, pp. 3907–3914, Accessed: Jun. 16, 2014. [Online]. Available: http://www.seas.gwu.edu/~mitdb/files/publications/referee13.pdf.

[11] M. A. Aly and A. F. Atiya, “LABR: A Large Scale Arabic Book Reviews Dataset.” in Proc. 51st Annual Meeting of the Association for Computational 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: Springer International Publishing, 2018, pp. 35–52.

[13] A. Elnagar and O. Einea, “BRAD 1.0: Book Reviews in Arabic Dataset,” in 2016 IEEE/ACIS 13th Int. Conf. Computer Systems and Applications (ICCSA), New York: IEEE.

[14] H. Elsahar and S. R. El-Beltagy, “Building Large Arabic Multi-domain Resources for Sentiment Analysis,” in Computational Linguistics and Intelligent Text Processing, Cham, 2015, pp. 23–34, doi: 10.1007/978-3-319-18117-2_2.

[15] M. Rushdi-Saleh, M. T. Martin-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, “OCA: Opinion corpus for Arabic,” J. Am. Soc. Inf. Sci. Technol., vol. 62, no. 10, pp. 2045–2054, Oct. 2011, doi: 10.1002/asi.21598.

[16] E. Refae and V. Rieser, “An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis,” presented at the LREC, Reykjavik, Iceland, 2014, Accessed: Jun. 16, 2014. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/317_Paper.pdf.

[17] M. Nabil, M. A. Aly, and A. F. Atiya, “ASTD: Arabic Sentiment Tweets Dataset,” in Proc. 2015 Conf. Empirical Methods in Natural Language Processing, pp. 2515–2519.

[18] N. Al-Twaireesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, “AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets,” in Procedia Computer Science, Dubai, 2017, vol. 117, pp. 63–72, 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 on Semantic Evaluation (SemEval-2017).
S. R. El-Beltagy, “NileULex: A Phrase and Word Level,” in Proc. 12th Int. Workshop on Semantic Evaluation, New Orleans, Louisiana, 2018, pp. 1–17, doi: 10.18653/v1/S18-1001.

A. A. Elmadany, H. Mubarak, and W. Magdy, “ArSAS: An Arabic Speech-Accent and Sentiment Corpus of Tweets,” Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

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, United Kingdom, Jul. 2019, pp. 40–50, Accessed: Jul. 18, 2020. [Online]. Available: https://www.aclweb.org/anthology/W19-5606

K. Abu 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 on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, May 2020, pp. 1–8, Accessed: Jul. 06, 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.osact-1.1

N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, “Arabic sentiment analysis: Lexicon-based and corpus-based,” in Applied Electrical Engineering and Computing Technologies (AEECT), 2013, pp. 1–6, Accessed: Jun. 16, 2014. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6716448

M. Abdul-Mageed, M. T. Diab, and M. Korayem, “Subjectivity and sentiment analysis of modern standard Arabic,” in Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers—Volume 2, 2011, pp. 587–591, Accessed: Jun. 16, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002851

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.

M. Al-Kabi, A. Gigieh, I. Alsamadi, H. Walsheh, and M. Haidar, “An opinion analysis tool for colloquial and standard Arabic,” in Fourth Int. Conf. on Information and Communication Systems (ICICS 2013), 2013, pp. 23–25.

K. Al-Rowaily, M. Abulaish, N. A.-H. Haldar, and M. Al-Rubaian, “BiSaL—A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security,” Digit. Investig., vol. 14, pp. 53–62, 2015.

S. R. El-Beltagy, “NileULex: A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic,” in Proc. Tenth Int. Conf. on Language Resources and Evaluation (LREC’16), Portoroz, Slovenia, May 2016, pp. 2900–2905, Accessed: Jul. 04, 2020. [Online]. Available: https://www.aclweb.org/anthology/L16-1463

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.

A. Mourad and K. Darwish, “Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs,” WASSA 2013, p. 55, 2013.

G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj, “A large scale Arabic sentiment lexicon for Arabic opinion mining,” ANLP 2014, p. 165, 2014.

F. 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, 2014.

R. Eskander and O. Rambow, “SLSA: A Sentiment Lexicon for Standard Arabic,” in Proc. 2015 Conf. on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 2545–2550.

T. Al-Moslimi, M. Albaraded, A. Al-Shabi, N. Omar, and S. Abdullah, “Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis,” J. Inf. Sci., p. 0165551516683908, 2017.

N. Al-Twaresh, H. Al-Khalifa, and A. Al-Salman, “AraSenTi: Large-Scale Twitter-Specific Arabic Sentiment Lexicons,” in Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Aug. 2016, pp. 697–705, doi: 10.18653/v1/P16-1066.

N. Al-Twaresh, M. Al Tuwaijri, A. Al Moammar, and S. Al Humoud, “Arabic spam detection in Twitter,” in 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media, 2016, p. 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.

E. H. R. Rho and M. Mazmanian, “Political Hashtags & the Lost Art of Democratic Discourse,” in Proc. 2020 CHI Conf. Human Factors in Computing Systems, New York, NY, USA, 2020, pp. 1–13, doi: 10.1145/3313831.3376542.

M. J. Egnoto, W. Wei, and C. Collins, S. Hasan, and S. V. Ukkusuri, “A novel transit rider satisfaction metric: Rider sentiments measured from online social media data,” J. Public Transp., vol. 16, no. 2, p. 2, 2013.

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. 2015 48th Hawaii Int. Conf. on System Sciences, pp. 1788–1797.

R. Pfitzner, A. Garas, and F. Schweitzer, “Emotional divergence influences information spreading in Twitter,” ICWSM, vol. 12, pp. 2–5, 2012.

P. Jaccard, “The distribution of the flora in the alpine zone,” New Phytol., vol. 11, no. 2, pp. 37–50, 1912.

J. Cohen, “A coefficient of agreement for nominal scales,” Psychol. Bull., vol. 43, no. 1, pp. 25–26, 1956.

B. D. Eugenio and M. Glass, “The kappa statistic: A second look,” Comput. Linguist., vol. 30, no. 1, pp. 95–101, 2004.

J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, pp. 159–174, 1977.