# Saudi Twitter Corpus for Sentiment Analysis

Adel Assiri, Ahmed Emam, Hmood Al-Dossari

***Abstract*****—**Sentiment analysis (SA) has received growing attention in Arabic language research. However, few studies have yet to directly apply SA to Arabic due to lack of a publicly available dataset for this language. This paper partially bridges this gap due to its focus on one of the Arabic dialects which is the Saudi dialect. This paper presents annotated data set of 4700 for Saudi dialect sentiment analysis with (K= 0.807). Our next work is to extend this corpus and creation a large-scale lexicon for Saudi dialect from the corpus.

***Keywords*****—**Arabic, Sentiment Analysis, Twitter, annotation.

## I. Introduction

THE Internet contains a growing amount of useful information that can be mined and, in turn, made accessible back to its users in creative ways [1]. Users can add reviews or opinions on web content. They can also share their ideas and opinions via social media such as Twitter, Facebook, personal blogs, and forums [2] as long as the web technology supports these features. Through social media, Arabic users tend to communicate with each other by using unstructured and ungrammatical slang Arabic language [3]. Sentiment analysis (SA) is the determination of text polarity as positive or negative [4]. In spite of the recent strong interest in SA, few studies have applied it to Arabic language analysis due to a lack of publicly available annotated data [4]. As a result, the focus of this research is on one Arabic dialect – Saudi. The purpose of this work is to present the first Saudi annotated corpus. This will be achieved by reporting a procedure of manual corpus annotation. This corpus includes data from Twitter and covers several domains such as sport, economy, and politics. The intention of this paper is to create the first reliably annotated Twitter data for the Saudi dialect which will be subsequently released to the LREC community as part of this submission.

## II. Arabic Language Challenges

As an important player in international politics and the global economy, the Arab world is the focus of many multi-national interest groups and analysts who endeavour daily to decipher sentiments on issues like oil and gas prices, stock market movements, politics and foreign policy, emanating from this part of the world. The resulting chatter being in the Arabic language, there is a great need for natural language analysis of large amounts of Arabic language text and documents to support the required sentiment extraction. As described in the foregoing, the relative importance of the Arabic language in global communications demands a proportional amount of interest and research for natural-language processing of large amounts of Arabic language text and documents to facilitate sentiment extraction for industrial use [5]-[7]. The reality, however, is that there is relatively little available support for Arabic-language sentiment analysis, majorly for the following reasons: (1) relatively limited scholarly work and research funding in this area, when compared to other-language studies, especially English. (2) Morphological complexities and dialectal varieties of the Arabic language which require advanced pre-processing and lexicon-building steps beyond what is applicable for the English language domain [6]-[8]. This limits the potential applications of current tools and custom tools for Arabic SA may not be easy to come by, may be limited in current functionality, or may not be freely available. Farra et al. [7] illustrated the challenges of Arabic-language sentiment analysis: the existence of many inflectional and derivation forms - where words have transitional meanings depending on position within a sentence, and the type of sentence (verbal or nominal). Multiple word prefixing, suffixing, affixing, and diacritical forms add high-order dimensionality for words, where the same three-letter root can generate different words in each case.

## III. Related Work

A recent study by [9] used Twitter to analyze a variety of Arabic dialects. By cross-referencing the geographical information from user profiles, the researchers used dialectical word n-grams in users tweets to identify their country.

A project called COLABA (Cross Lingual Arabic Blog Alerts) [10] uses multiple systems to develop NLP resources for Arabic dialects. To date, these dialects have included Levantine, Egyptian, Moroccan, and Iraqi. The system use in the current study also utilized MAGEAD (Morphological Analyzer and Generator for the Arabic Dialect) [11] and the Buckwalter morphological analyzer and generator (BAMA) [12]. COLABA's capacity to process Arabic dialects was evaluated through its information retrieval system. Importantly, COLABA does not draw upon other dialects during its operation. Meanwhile, DIWA (Dialectical Word Annotation Tool for Arabic) [13] operates as a desktop application and can be used offline. DIWAN captures the annotation of morphological features in context. In addition, DIWAN can assist in training taggers because it can help create corpora. Unlike COLABA, DIWAN can use resources

Adel Assiri is with Information System Department, King Saud University, Saudi Arabia, Riyadh, currently with Abha Technology College, Saudi Arabia, Abha, TVTC (corresponding author, phone: 00966500101053; e-mail: aadel_3@hotmail.com).

Ahmed Emam (Associate Professor) is with Information System Department, King Saud University, Saudi Arabia, Riyadh, was with Menoufia University, Menoufia, Egypt (e-mail: aemam@ksu.edu.sa).

Hmood Al-Dossari (Assistant Professor) is with Information System Department, King Saud University, Saudi Arabia (e-mail: hzaldossari@ksu.edu.sa).

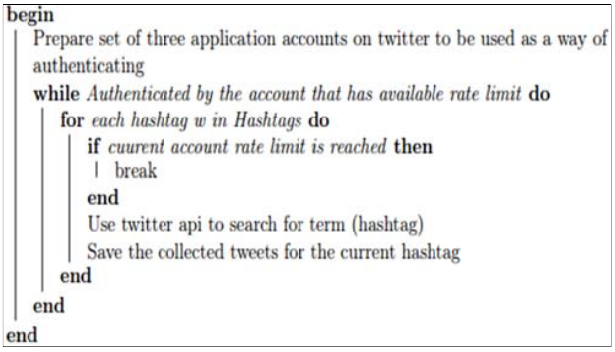
from other dialects. In a study on annotated Arabic via Twitter, researchers presented a public corpus for Arabic based on a data set of nearly 8.868 feeds [14]. This corpus was subsequently assessed for subjectivity using SA. Another study on annotation produced a dataset of documents derived from online Arabic forums [15]. However, the process by which the data was annotated is unavailable because the dataset was not made available to the public. This lack of access reveals an important gap in the existing literature. In another instance of Arabic annotation [4], reveal a corpus of MSA (Modern Standard Arabic). Focused on the news as its domain, this corpus was built through manual annotation of subjectivity at the sentence level. Soon after, this corpus was extended using AWATIF, a multi-genre corpus of MSA [16]. The genres used included Wikipedia, web forums, and the Penn Arabic Treebank. In fact, the MSA corpus broke a barrier as the first dataset about newswires. Furthermore, [17] conducted an Arabic social sentiment analysis of Twitter and classified data as either objective, subjective positive, subjective negative, or subjective mixed. In addition, [17] used 4-way sentiment classification and two stage classification. This research also generated a seed sentiment lexicon.

## IV. Data Collection

SA is often performed on data from genres such as movies or on product reviews and blogs in which the opinions of users are openly expressed and are often highly subjective [18]. In spite of the important role Twitter plays in the daily lives of millions of people, Twitter has received much very little attention as a medium for SA. This lack of research created the impetus to build a comprehensive SA application and collect an extensive dataset for this platform. This application needed to be capable of dynamically generating a random dataset from the required data using specific algorithms. As a result, the application used for this search was a web-based system running on the open-source Ruby-on-Rails (RoR) technology. The use of well-developed technology such as RoR was instrumental for creating an efficient development cycle.

The application used the Twitter API to collect data. It was hosted on "Heroku," a powerful web-hosting platform. In essence, Twitter became the data source tweets originating from Saudi Arabia were collected using the Twitter API. A set of Twitter hashtags originating in KSA were used as the initial dataset for searching Twitter.

Of note, this data collection protocol encountered several challenges. As a result, several adaptations were required. The first challenge pertained to API restrictions. Twitter sets some rules and limitations regarding the use of its API that prevented us from collecting enough data. In particular, Twitter sets a rate limit that sets the amount of requests for information at 180 per 15-min interval. A second challenge included the need for extensive text cleaning. A lot of Arabic tweets contain links, hashtags, and other non-Arabic words, that were unneeded in for the dataset. Therefore, tweets needed to be cleaned to remove non-Arabic elements. Finally, duplicated tweets (e.g., retweets) were a problem for developing the dataset and, as a result, required deletion. In order to overcome these challenges, a custom search engine was developed to search for hashtags in Twitter using the Twitter API. The Twitter rate limit obstacle was addressed by creating a multi-account stack, and serially implemented a batch process with the logic that a new account will come into effect when the rate limit is reached for the current account. Additionally, the search engine removes retweeted tweets and duplicate text before and after the cleaning process. A description of the algorithm is available in Fig. 1.



```
begin
  Prepare set of three application accounts on twitter to be used as a way of
  authenticating
  while Authenticated by the account that has available rate limit do
    for each hashtag w in Hashtags do
      if cuurent account rate limit is reached then
        break
      end
      Use twitter api to search for term (hashtag)
      Save the collected tweets for the current hashtag
    end
  end
end
```

Fig. 1 Pseudo Code for collecting tweets using twitter API

In total, 4.700 randomly selected tweets from the collected data were annotated to create the final dataset. This was accomplished by using the application's user interface (UI). This UI allowed for the marking of a tweet as positive, neutral, or negative as the interface also allowed for the filtering of the tweets by polarity. For example, this provided the option to only display the positive tweets. Tweets were also filterable by hashtags and user handles used to collect them. Beyond its filtration capabilities, the user interface allowed for the addition of new text, the editing of existing text, or the deleting of text.

TABLE I
Agreement for SA Sentences

| | POS | NEG | NEU | Total |
|---|---|---|---|---|
| POS | 1740 | 1 | 89 | 1830 |
| NEG | 0 | 1946 | 45 | 1991 |
| NEU | 360 | 69 | 450 | 904 |
| Total | 2100 | 2016 | 584 | 4700 |

## V. Data Annotation

To perform the annotation, two postgraduate students, native speakers of Arabic, annotated 4.700 sentences extracted from Twitter. Cohen's Kappa [19] was used to assess the reliability of the annotations. Cohen's Kappa coefficient measures interpreter reliability for qualitative (categorical) items and represents the probability of agreement occurring by chance. Table I shows the agreement table for the two annotators. The overall observed agreement is 88% (Kappa = 0.807), indicating the strength of agreement is considered to be 'very good'. [20].

The annotation procedure required the annotators to follow six specific instructions:

1- If annotators believed the sentence under review was an objective report of a news item, they were directed to avoid labeling good news as positive or bad news as negative (i.e., absent of any positive or negative sentiment). By definition, a bad news item or a good news item can be labeled as neutral. The following are two examples of neutrally labeled items:

Example 1: "أجواء اليوم حارة وغبار"

**Translation:** "Today is too hot and dusty"

Example 2: "سوق الاسهم طايح"

**Translation:** " The stock market is decreasing sharply "

2- Because the point-of-view from which a sentence is composed influences whether a positive or negative sentiment label is assigned, annotators were asked to assign sentences based on the authors who wrote them. For example, the statement, “النصر يهزم الهلال”, "Al-Nasser defeated Al-Hilal" the act of winning is deemed a positive from the perspective of Al-nassr’ followers but, contrarily, the same item can be deemed negative when taken from the perspective of Al-Hilal’s audience.

3- Annotators were asked to consider epistemic modality during their evaluation of Twitter sentences. Epistemic modality, a modality that deals with the speaker’s judgment of the knowledge of their statements, has been shown to influence trust in the truth of a proposition [21]. For example, confidence can be influenced by the use by hedges, such as “maybe”, “perhaps”, or “somewhat”, or strengthened with boosters or intensifiers like “of course” or “certainly” [22]. Furthermore, epistemic modality can also modify the subjectivity and polarity of a sentence. For example, a statement like “car crashes has killed a lot of people” lacks sentiment and, therefore, can be annotated as Neutral. By contrast, the sentence “Regrettably, car crashes have killed a lot of people” can be labeled strong Negative. “Strong” is applied because of the use of “regrettably.”

4- Each annotator was ordered not let any personal background knowledge influence the interpretation of content. Examples of background knowledge include exposure to influential social, cultural, or religious forces. Although challenging to suspend, such background knowledge can shape decisions about polarity. For example, the statement, “conservatism should not be taught in public schools” would be negative to a person with conservative political beliefs but positive to someone who does not.

5- During annotation process there will be what so called " Do'aa " which means " supplicating to God ". Such a case of Do'aa is commonly used in Twitter. Do'aa can be devided into two main streams: (a) Do'aa for the sake of someone and (b) Do'aa against the sake of someone. By consulting semantic people we came a conclusion that Do'aa denotes two attitudes: positive and negative.

Example 3: For positive Do'aa:

"ربي يوفقك في الدنيا والآخرة"

**Translation:** "My lord guide you in this life and in the hereafter".

Example 4: For negative Do'aa:

"حسبي الله عليهم"

**Translation:** "Allah is enough for me upon them"

6- The final direction for each annotator for this study, each annotator was to assign one of three labels:

### *A. Positive (POS)*

Example 5:

" الله يسعدك خبر يفتح النفس "

**Translation:** "May Allah make you happy for the news that makes soul happy"

### *B. Negative (NEG)*

Example 6:

"حتى موبايلي مخيس"

**Translation:** "Mobily is as well as rubbish"

### *C. Neutral (NEU)*

" عندي هذا الجهاز جديد الي يحتاجه يكلمني في جده "

**Translation:** "For whom he needs it, call me in Jeddah"

## VI. Conclusion

The purpose of this research was to present an annotated corpus that applied SA to Twitter content. The preceding application of this annotation first of its kind for the Saudi dialect. As a result, this corpus will be the first publicly released corpus of its kind. In future work we will extend this corpus, then use it to generate a large scale lexicon for Saudi dialect this lexicon we help us to build a comprehensive SA system for Saudi dialect using big data technique.

## References

[1] M.T. Diab, L. Levin, T. Mitamura, O. Rambow, V. Prabhakaran, and W. Guo. 2009. Committed belief annotation and tagging. In Proceedings of the Third Linguistic Annotation Workshop, pages 68–73. Association for Computational Linguistics.

[2] M. N. Al-Kabi, A. H. Gigieh, I. M. Alsmadi, H. A. Wahsheh, and M. M. Haidar,2014. Opinion Mining and Analysis for Arabic Language. *Int. J. Adv. Comput. Sci. Appl.,* 5(5).

[3] Nawaf A. Abdulla, Nizar A. Ahmed, Mohammed A. Shehab and Mahmoud Al-Ayyoub, 2013. Arabic Sentiment Analysis: Lexicon-based and Corpus-based. *IEEE Conference on Applied Electrical Engineering and Computing Technologies, Jordan,* pp.1-6.

[4] M. Abdul-Mageed and M. T. Diab. Subjectivity and sentiment annotation of modern standard arabic newswire. In Proceedings of the 5th Linguistic Annotation Workshop, LAW V ’11, pages 110–118, 2011.

[5] Al-kabi MN, Abdulla NA and Al-ayyoub M. An analytical study of arabic sentiments: maktoob case study. In: 8th international conference for internet technology and secured transactions, IEEE, London, UK, pp. 89-94, 2013.

[6] Farra N, Challita E, Abou-assi R and Hajj H. Sentence-level and document-level sentiment mining for arabic texts. In: International conference on data mining workshops, IEEE, pp. 1114-1119, 2010.

[7] Korayem M, Crandall D and Abdul-mageed M. Subjectivity and sentiment analysis of arabic: a survey. In Advanced Machine Learning Technologies and Applications, 128-139, 2012.

[8] Sarah O. Alhumoud, Mawaheb I. Altuwaijri, Tarfa M. Albuhairi, Wejdan M. Alohaideb. Survey on Arabic Sentiment Analysis in Twitter. International Science Index, 9 (1), pp. 364-368, 2015.

[9] Mubarak, H., & Darwish, K. (2014). Using twitter to collect a multi-dialectal corpus of arabic. ANLP 2014, 1.

[10] Diab, M., Habash, N., Rambow, O., Altantawy, M., & Benajiba, Y. (2010). Colaba: Arabic dialect annotation and processing. In Lrec workshop on semitic language processing (pp. 66–74).

[11] Habash, N., & Rambow, O. (2006). Magead: a morphological analyzer and generator for the arabic dialects. In Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics.

[12] Buckwalter, T. (2004). Buckwalter arabic morphological analyzer version 2.0. ldc catalog number ldc2004l02 (Tech. Rep.). ISBN 1-58563-3-0.

[13] Faisal Al-Shargi, Owen Rambow. DIWAN: A Dialectal Word Annotation Tool for Arabic. Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 49–58, Beijing, China, July 26-31, 2015. c 2014 Association for Computational Linguistics.

[14] Eshrag Refaee and Verena Rieser. 2014. An Arabic twitter corpus for subjectivity and sentiment analysis. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14), Reykjavik, Iceland, may. European Language Resources Association (ELRA).

[15] A. Abbasi, H. Chen, and A. Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Trans. Inf. Syst., 26:1–34.

[16] M. Abdul-Mageed and M. Diab. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), 2012.

[17] Mahmoud Nabil, Mohamed Aly, Amir F. Atiya. ASTD: Arabic Sentiment Tweets Dataset. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2515–2519, Lisbon, Portugal, 17-21 September 2015. c 2015 Association for Computational Linguistics.

[18] A. Balahur and R. Steinberger. 2009. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. Proceeding of WOMSA.

[19] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1):37–46.

[20] Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, 22(2):249–254.

[21] F. Palmer. 1986. Mood and Modality. 1986. Cambridge: Cambridge University Press.

[22] L. Polanyi and A. Zaenen. 2006. Contextual valence shifters. Computing attitude and affect in text: Theory and applications, pages 1–10.