Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 63–72

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates

AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets

Nora Al-Twairesh,Hend Al-Khalifa,AbdulMalik Al-Salman,Yousef Al-Ohali*

College of Computer and Information Sciences, King Saud University,Riyadh, Saudi Arabia

Abstract

Arabic Sentiment Analysis is an active research area these days. However, the Arabic language still lacks sufficient language resources to enable the tasks of sentiment analysis. In this paper, we present the details of collecting and constructing a large dataset of Arabic tweets. The techniques used in cleaning and pre-processing the collected dataset are explained. A corpus of Arabic tweets annotated for sentiment analysis was extracted from this dataset. The corpus consists mainly of tweets written in Modern Standard Arabic and the Saudi dialect. The corpus was manually annotated for sentiment. The annotation process is explained in detail and the challenges during the annotation are highlighted. The corpus contains 17,573 tweets labelled with four labels for sentiment: positive, negative, neutral and mixed. Baseline experiments were conducted to provide benchmark results for future work.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: Sentiment Analysis; Arabic NLP; Corpus Sentiment Annotation; Arabic tweets; Saudi Dialect

1. Introduction

Sentiment Analysis (SA) is one of the most vital research fields of Natural Language Processing (NLP) nowadays. It has emerged as an active research field with the proliferation of textual data on the Web especially in social media websites. It can be accomplished through both supervised and unsupervised learning techniques. In both techniques, labelled data is required for training and testing in supervised learning and for testing in

* Corresponding author. E-mail address: twairesh@ksu.edu.sa

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.094

unsupervised learning. Most NLP tasks need an annotated corpus for training machine learning classifiers, the corpus has to be in machine-readable form i.e. it has to be annotated for the machine to understand it.

SA of English has been thoroughly researched; however, research on SA of Arabic has just flourished. Arabic is ranked fourth among languages on the web although it is the fastest growing language on the web among other languages. Arabic is a morphologically rich language where one lemma could have hundreds of surface forms; this complicates the tasks of SA. Moreover, Arabic language is in a state of diglossia where the formal language used in written form differs radically from the one used in every-day spoken language [1]. The formal language is called Modern Standard Arabic (MSA) and the spoken language differs in different Arabic countries producing numerous Arabic dialects sometimes called informal Arabic or colloquial Arabic. The language used in social media is known to be highly dialectal [2]. Previous research on SA of Arabic was merely for MSA, but recently researchers started addressing Dialectal Arabic (DA). Dialects differ from MSA phonologically, morphologically and syntactically [1]. Moreover, dialects do not have standard orthographies. Most Arabic NLP solutions are designed for MSA and perform poorly on DA [3]. As Farghaly and Shaalan [4] point out, it is very difficult and almost impossible for one NLP solution to process all the variants of Arabic. As such, an Arabic NLP solution has to specify the Arabic variant it can process beforehand.

Twitter is considered a powerful tool for disseminating information and a rich resource for opinionated text containing views on many different topics: politics, business, economic, social etc. [5]. This has stimulated the interest of the NLP research community to study this rich language resource. In particular, the Saudi community has witnessed an increased use of this social media platform as confirmed in a study by Semiocast, that the number of twitter users in Saudi Arabia almost doubled in the span of 6 months in 2012 and that Riyadh (the capital city of Saudi Arabia) is now the 10th most active city on Twitter [6]. In a recent study, Mubarak and Darwish [7] used Twitter to collect a multi-dialect corpus of Arabic; a dataset of 175 M Arabic tweets was collected. Then after filtering on tweet user location, a subset of 6.5 M tweets was classified according to the tweet's dialectal language, they found that 61% of the tweets were in Saudi dialect followed by 13% Egyptian and 11% Kuwaiti. This demonstrates the enormous presence of the Saudi community on Twitter.

Accordingly, a possible solution for Arabic sentiment analysis requires language resources that are dialect specific. It had been demonstrated in [8] that the lack of Arabic corpora is one of the challenges that face Arabic sentiment analysis. In this paper, a dataset that contains around 2.2 million Arabic tweets was collected during a span of three months. A corpus of Saudi tweets was extracted from the dataset. Then, the corpus was manually annotated by three annotators and kappa statistics were calculated to ensure the reliability of the annotations. The corpus was used in the development and testing of several sentiment analysis classifiers for Saudi tweets.

The contributions of this paper can be summarized as follows:

1. A large dataset of Arabic tweets was collected (2.2 million tweets) which was utilized to construct data resources for Arabic SA.
2. A corpus of Saudi tweets was extracted from the larger dataset, the size of the corpus after manual annotation was 17,573. The corpus was annotated by four labels (positive, negative, neutral and mixed). To the best of our knowledge this is the largest manually annotated corpus of Saudi tweets.
3. The annotation guidelines and the challenges during annotation are highlighted to provide insights for future annotation projects of Arabic SA.
4. A set of benchmark experiments were provided to establish a baseline for future benchmarking of Arabic SA.
5. We make the corpus publically available for the research community.

The rest of this article is organized as follows: Section 2 reviews the related work on Arabic sentiment corpora. Section 3 describes the details of the datasets used to construct the corpus and how they were collected. In section 4, we present the AraSenTi-Tweet corpus and the annotation process is explained in detail. In section 5, the benchmark experiments are illustrated. Section 6 highlights the challenges faced during annotation. Finally, we conclude the paper in section 7.

2. Related Work

Research on SA of Arabic emerged in 2008 with the publication of [9] which presented a supervised approach to sentiment analysis of both English and Arabic in web forums. A survey on SA of Arabic was first presented in [10],

a recent survey on SA of Arabic can be found in [8].We review here the corpora and datasets publically available for Arabic SA while highlighting the construction methods of each. Also, for a comprehensive review of the resources available for Arabic SA including corpora, we refer the reader to [11]. One of the earliest corpora for Arabic SA is the Opinion Corpus for Arabic (OCA) [12] which has served as a benchmark for several Arabic SA studies. It was constructed manually through the extraction of 500 Arabic movie reviews from the Web. It contains 250 positive reviews and 250 negative reviews. It was used by numerous research papers e.g. [13]–[16] to name a few. Another corpus in the domain of reviews is the Large Arabic Book Review Corpus (LABR) [17]. It contains a dataset of 63,257 book reviews from the book readers' social network www.goodreads.com. The reviews were already rated from 1-5. They considered as positive reviews those with ratings 4 or 5, and negative reviews those with ratings 1 or 2. Reviews with rating 3 are considered neutral and not included in the polarity classification. This dataset was used in the following papers: [18]–[21]. AWATIF is a multi-genre corpus for MSA SA [22]. It consists of sentences from the Penn Arabic Tree Bank (PATB), Wikipedia Talk pages and Web forums. However, it hasn't been released to the public.

In the expanse of Twitter datasets, [23] constructed and released a corpus of Arabic tweets annotated for subjectivity and sentiment analysis available on the LREC repository of shared resources[3]. 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. The authors in [24] presented the Arabic Sentiment Tweets Dataset (ASTD) which is a dataset of 10,000 Egyptian tweets. It consists of 799 positive, 1,684 negative, 832 mixed and 6,691 neutral tweets. The authors also conducted a set of benchmark experiments of four-way sentiment classification and two stage classifications.

MIKA is a SA corpus of MSA and Egyptian dialect [25]. The corpus is extracted from tweets, comments on hotel reservations and TV programs and product reviews annotated at sentence level. It consists of 2154 positive, 1648 negative and 198 neutral texts. The text was also annotated with polarity strength (-10 to 10) following a set of rules that handle negation, contextual intensifiers and mixed polarity cases.

In [19], the authors constructed a dataset of reviews: hotels, restaurants, movies and products. The total was 33,116 reviews. The sentiment of the reviews was determined through the ratings, so no manual annotation was required. A small lexicon was extracted that contained around 2,000 words classified as per review type. Several experiments were conducted to determine the best features and classifiers to be used with the dataset. This work serves as a good benchmark for the reviews genre.

Most of the work mentioned above does not give any details on the annotation process and what annotation guidelines were used. In the field of Arabic SA, [26] outlined the annotation guidelines that were used in annotating Arabic text from the newswire genre. The annotation guidelines were linguistically motivated and provided novel insights that we adopted in our work when preparing the guidelines. Moreover, [27] presents a practical guide to annotation of sentiment while outlining the challenges of annotating sentiment; which also helped us in shaping the guidelines in our work.

3. Data Collection

We followed the approaches in previous work on SA of English Twitter to collect the datasets. As in [28], [29] we utilized distant supervision through emoticons as noisy labels to construct the first dataset EMO-TWEET. In [30], [31] the authors used hashtags of sentiment words such as #good and #bad to create corpora of positive and negative tweets, we adopted a similar approach to theirs. Initially, we tried collecting Arabic sentiment words with hashtags such as #سعادة #مؤسف but the search results were too low. Accordingly, we opted to use the sentiment words as keywords without the hashtag sign and the number of search results was substantial. These results constitute our second dataset KEY-TWEET. The third dataset is the SAUDI-TWEET dataset, where we extracted from the previous datasets all tweets that had their location set to a Saudi location. In total, we collected around 6.3 million Arabic tweets in a time span of three months. Certain filtration and cleaning was applied on the tweets and the remaining tweets were 2.2 million tweets.

Tweets containing the happy emoticon ":)" and the sad emoticon ":(" and the rule "lang:ar" (to retrieve Arabic tweets only) were collected during November and December 2015. The total number of Tweets collected is shown in Table 1. Retweets, tweets containing URLs or media and tweets containing non-Arabic words were all excluded

from the dataset. The reason for excluding tweets with URLs and media is that we found that many tweets containing URLs and media were spam. The remaining tweets after filtering and cleaning constitute our first data set which we will name EMO-TWEET. We further decompose it to: EMO-TWEET-POS for tweets containing the smiley emoticon which we consider as positive and EMO-TWEET-NEG for tweets containing the sad emoticon which we consider as negative.

In addition, tweets containing 10 Arabic words having positive prior polarity and 10 Arabic words having negative prior polarity were collected during January 2016. The same pre-processing steps described above were applied on these tweets. We name the resulting datasets KEY-TWEET-POS for tweets containing the Arabic words with positive prior polarity and KEY-TWEET-NEG for tweets containing the Arabic words with negative prior polarity.

Previous studies have shown that the majority of daily Arabic tweets (60%) are from Saudi Arabia [6], [7]. Our approach to identify Saudi tweets was to filter on user location. The reason for this was that there are many common dialect words that are used by different countries at the same time and filtering on these words could cause false positives [7]. We found that most tweets do not have the location field set in the user profiles and as such the number of Saudi tweets we found was very low compared to the number of collected tweets in all. We used a list of Saudi locations provided by [7] to filter on the location field of the tweets. Moreover, we also added tweets that had the time-zone field set to Riyadh. The number of Saudi Tweets is shown in Table 1. We will call this data set SAUDI-TWEET and further decompose it to SD-EMO-TWEET-POS, SD-EMO-TWEET-NEG, SD-KEY-TWEET-POS, SD-KEY-TWEET-NEG.

Table 1. Statistics of Collected Tweets.

	Positive Emoticon :)	Negative Emoticon :(	Positive Keywords	Negative Keywords	Total
Collected tweets	2,245,054	1,272,352	1,823,517	1,000,212	6,341,135
Cleaned and filtered tweets	1,033,393	407,828	447,170	337,535	2,225,926
Saudi tweets	51,393	29,933	32,188	29,349	142,863

3.1. Dataset Cleaning and Preprocessing

The text of tweets is known to be noisy and should be cleaned and pre-processed in order for the analysis to be performed. Tweets that are retweets and tweets that contained URLs or media were already excluded from the dataset in the collection phase. In addition, user mentions (@user) were also removed from the tweets. Then the tweets were processed through MADAMIRA [32] for normalization and tokenization. Normalization is the process of unifying the shape of some Arabic letters that have different shapes. The Arabic letters (أ, ة, ي, و) are normalized

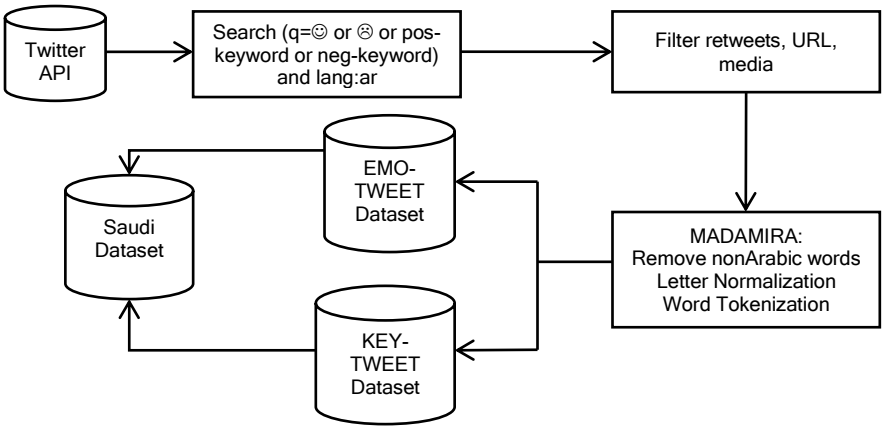

Figure 1. Dataset Cleaning and Preprocessing.

to convert multiple shapes of the letter to one shape, the different forms of "alif" (أ,آ,إ) are converted into (ا), the letter "ta'a" (ة) is converted to (ه), the different forms of "ya'a " (ي,ى) are converted into (ي), and the letters (ؤ, ئ) are converted to (ء). The different steps of data collection, filtering, cleaning and pre-processing are illustrated in Figure 1.

4. AraSenTi-Tweet Corpus

Sentiment annotation is the process of annotating text for sentiment. The labels used in annotation depend on the classification output. In this paper, we propose a four-way classification of sentiment into (positive, negative, neutral, and mixed). As such the labels used for annotation are positive, negative, neutral, mixed and we added one more label: indeterminate to accommodate the cases where the annotator cannot predict the sentiment of the text as suggested by [33]. The meaning and examples of each label are illustrated in Table 2. Then we recruited three annotators that are CS graduates and Arabic/Saudi native speakers. We chose three not two annotators so we can resolve the conflicts in annotation through majority voting i.e. we would choose the label that at least two annotators chose. This seems as a rational choice, since the annotation of three is more confident than two.

Table 2. Labels used in annotation and examples of each.

Labels	Example in Arabic	English Translation
Positive: if there is a clear indicator that the opinion is positive even if it is not strong.	مطار. الملك . خالد تغير إيجابي ملحوظ	King Khaled Airport a remarkable positive change.
Negative: if there is a clear indicator that the opinion is negative even if it is not strong.	للأسف أثبت أنه مذيع فاشل جدا	Unfortunately he prove that he is a failure interviewer.
Mixed: if the tweet contains both a positive and negative opinion.	قارئ جرير رائع لكن الاسعار غالية	The Jarir reader is fabulous but the prices are expensive.
Neutral: if there is no opinion in the tweet.	ممكن ترشح لي برنامج قارئ باركود ممتاز	Can you please suggest an excellent barcode reader program.
Indeterminate: if there is an opinion but it is not clear if it is positive or negative, or if you are not sure if it contains an opinion or not.		

4.1. Annotation Guidelines

The annotators were presented with annotation guidelines and received a one-hour session of training. We outline here the guidelines and present the rationale behind each one.

(1) *News*: News are not considered subjective. So they should be annotated as neutral, even if they are conveying good or bad news.
(2) *Perspective*: The sentiment should be considered from the author's perspective not the annotator's.
(3) *Context*: The choice of label should be according to the context of the text.
(4) *Ambiguity*: If the opinion is not clear, please do not try to guess, just choose (indeterminate).
(5) *Mixed*: The mixed label should be chosen with careful consideration.
(6) If the tweet contains a smiley emoticon but the content of the tweet is negative or the opposite, you should choose mixed.
(7) The subject of tweet could not be clear sometimes because the mentions and hashtag sign were removed, but the sentiment can still be determined.

In the guidelines 1 and 2, we adopted what was presented in [26] for annotating news and determining the perspective that the annotation should rely on; since it is common that the annotator could have an opposite opinion towards a certain topic than the author has. In guideline 3, the annotator was asked to choose the sentiment that is understood from the context of the tweet, this is to ensure that the annotator does not rely on her background knowledge of the topic of the tweet when choosing the label. Since it is common that the text of the tweet might be not clear due to the ambiguous nature of the Arabic language and the length of the tweet being short, the annotators

were asked not to guess the sentiment if it is not clear but to choose the label (indeterminate). This ensures that we do not present the classifier with training examples that are ambiguous and thus affect the classification results. Mixed sentiment could be zconfusing; as such the annotators were directed to choose it with consideration. The reason for this is that a tweet could contain positive and negative words but still convey one sentiment. Also a tweet could convey both positive and negative sentiments but towards different topics or entities. Moreover, considering that the tweets contain emoticons that convey sentiment, there could be cases where a positive emoticon appears with text conveying negative sentiment or vice versa; annotators were asked to label these cases as mixed. Finally, since during the pre-processing of tweets, the mentions and hashtag signs were removed, the annotators could be confused as to what the subject of the tweet is. However, during the manual selection of tweets to be included in the corpus by the authors, we made sure that this does not affect determining the sentiment.

Although the guidelines were designed to be as simple as possible and the annotation process was explained to the annotators in the training session, the annotators did face some challenges during annotation. We highlight these challenges in a following section.

4.2. Data Annotation

Our intention was to annotate 20,000 tweets by extracting 5,000 tweets from each of the four subsets of the Saudi dataset: SD-EMO-TWEET-POS, SD-EMO-TWEET-NEG, SD-KEY-TWEET-POS, and SD-KEY-TWEET-NEG. The process of extracting tweets from the larger datasets was done in two stages: first we made a manual inspection of the datasets and found that we cannot choose randomly and had to extract the tweets that did contain sentiment and would not be confusing for the annotators, this was done by one of the authors. After manually reviewing 50,500 tweets, only 13,226 tweets were found valid to be included in the dataset i.e. around 26%.

Consequently, we had to decrease the number of extracted tweets from the datasets to 13,226 and add 6,090 newly collected tweets. The reason for this was twofold. First, we found that the tweets in the emoticon datasets were of a chatting nature and rarely constitute any sentiment or even a complete sentence. Out of the randomly chosen tweets from the emoticons datasets which were 18,142 tweets, only 3,209 tweets were found valid to be included in the dataset i.e. 17% only. The second reason was that we found that most of the annotated tweets in the first stage were positive or negative and we needed to augment the dataset with more neutral tweets. Therefore, we collected 4,000 tweets from two Saudi news accounts: The Saudi Press Agency @spagov and Sabq online newspaper @sabqorg. These tweets were not manually annotated but labelled neutral without annotation as all the tweets were news. However, a manual inspection of these tweets was performed to guarantee they are all neutral.

We also needed a set of tweets that was not from the larger dataset for testing purposes. This will ensure fair evaluation of the classification methods to be developed. Therefore, we collected 2090 tweets from three trending Saudi hashtags during February 2016. These tweets were also filtered and cleaned as in section 3. The resulting number of tweets after cleaning was 1580. A total of 14,806 tweets were manually annotated by the recruited annotators. The annotation process took around two months. To ensure the reliability of the annotations, reliability measures were calculated as presented in the following sub-section.

4.3. Inter-Annotator Agreement

We need to prove that the annotations are reliable. If the annotators assign similar labels, then we can deduce that they have a similar understanding of the annotation guidelines and that their performance was consistent. Reliability measures test the trustworthiness of annotation guidelines and scheme. Inter Annotator Agreement (IAA) can also be used to know how difficult a task is. It can be argued that if two or more annotators cannot reach an agreement on a specific annotation then it is rational to say that the machine learning classifier will not be able to classify the instance correctly also.

In the case of having more than two annotators, Fleiss's Kappa [34] is used for measuring IAA. We calculated Fleiss's Kappa for the 14,806 tweets that were annotated by three annotators with the following classes: positive, negative, mixed, neutral, and indeterminate. The Kappa was 0.60 which is considered moderate according to [35]. Moreover, studies show that detecting sentiment in text is a hard task for humans, and annotator agreement in

sentiment analysis is significantly lower than other NLP tasks such as POS tagging [36]. Accordingly, we attempted to identify the challenges that appeared during the annotation in section 6.

We found that 62% of the tweets were classified by the three annotators with the same class, while 31% of the tweets were classified by two annotators with the same class and 7% were not agreed upon by all three annotators. We excluded this last set from the corpus since they were confusing for the annotators; we expect them to be confusing and misclassified also by the machine learning classifiers. Moreover, the annotators classified 151 tweets as indeterminate and these were also excluded. We named the corpus AraSenTi-Tweet. Statistics of the AraSenTi-Tweet corpus are illustrated in Table 3.

Table 3. Statistics of the AraSenTi-Tweet corpus.

Class	No. of Tweets	No. of Tokens
Positive	4957	93,601
Negative	6155	127,182
Neutral	4639	71,492
Mixed	1822	39,883
Total	17573	332,158

5. Benchmark Experiments

The AraSenti-Tweet corpus is divided into a training set and test set. The split was not based on a percentage as what is done when there is one dataset. As we illustrated in section 4.2 when constructing the AraSenTi-Tweet corpus, the training and test sets were constructed separately since we needed a test set that was not extracted from the larger dataset that the training set was extracted from. This ensures fair evaluation and also eliminates the possibility of overfitting. The splits of the dataset are illustrated in Table 4.

Table 4. Dataset splits.

Class	Training set	Test Set	Total
Positive	4235	722	4957
Negative	5515	640	6155
Neutral	4065	574	4639
Mixed	1777	45	1822
Total	15592	1981	17573

To establish a baseline that can be used for benchmarking on the AraSenTi-Tweet corpus, we conducted several experiments for multi-way sentiment classification. For two-way classification we use only the positive and negative tweets, for three-way classification we use only the positive, negative and neutral tweets, and for four-way classification we use all the classes. For evaluation we report the F1-score [37] of the classification. All the reported results are on the test set.

For classification, we used SVM with a linear kernel. For the term feature we test the term-presence, term-frequency, and TF-IDF(Term Frequency- Inverse Document Frequency) features. The results for all classification models are illustrated in Table 5. We observe that the term presence is the best performing feature for two-way and three-way classification. This result was expected since tweets are short and the possibility of a term to be repeated in one tweet is very low. However, for four-way classification TF-IDF showed better performance.

From Table 5 we notice the following observations: the number of classes highly affects the performance; the more the classes the less the performance. Albeit, this is expected in classification problems. This suggests that three-way and four-way classification need a sophisticated feature extraction model to enhance the performance. Also, we notice from Table 5 that the number of training and testing instances for the mixed class are much less than

the other classes. As such, adding more instances of this class could enhance the performance of the four-way classification model.

Table 5. F-score of classification experiments.

Classification model	Term Presence	Term Frequency	TF-IDF
Two-way Classification	**62.27**	61.5	60.05
Three-way Classification	**58.17**	58.09	58.15
Four-way Classification	54.23	54.6	**54.69**

6. Challenges in Annotation of Arabic Sentiment

To identify the challenges of the annotation task of Arabic sentiment, we presented the annotators with a questionnaire after the completion of the annotation. In response to the clarity of the annotation guidelines, two annotators stated that the guidelines were very clear, while one annotator stated that they were kind of clear. The annotators were asked if the annotation of tweets was clear (always, sometimes, a little bit); all three annotators chose sometimes. The annotators were also asked which label was the hardest to determine, two annotators said (mixed) and one annotator said (neutral). Then the annotators were asked to state the challenges they faced. We summarize the challenges and give examples when appropriate.

- *Supplications*: annotators expressed difficulty in determining the sentiment of supplications since they can contain positive or negative words but whether they convey sentiment is unclear. We suggest for future annotation tasks of Arabic sentiment that an explicit guideline is added to show how to label supplications.
  - *Ex*: اللهم اجعل لنا في القلب نور وفي المال بركة وفي الناس محبة وفي الدنيا سعادة وفي الاخرة نجاة مساء الخير
    Translation: Oh God, give us light in our hearts and blessing in money and love in people and in this world happiness and in the hereafter survival Good evening
    Annotation: This tweet was labelled neutral by two annotators and positive by one annotator.
  - *Ex*: يا رب فرج هم كل من كان في ضيق . . و إبدله سعادة لا تنتهي .
    Translation: O Lord, relief those who are in oppression, and replace him with happiness that is endless.
    Annotation: This tweet was labelled neutral by two annotators and indeterminate by one annotator.
- *Advice*: this comes in the form of advice given as if it is good or bad to do or not to do something.
  - *Ex*: أي شيء مستور لا تحاول أن تكشفه. سترنا الله و اياكم في الدنيا والآخرة لا تحاول أن تبحث عن الوجه الثاني لأي شخص حتى لو كنت متأكدا انه سيء
    Translation: Anything that is hidden, do not try to expose it. May God shield us in life and hereafter, do not try to look for the other face of any person even if you are sure he is bad
    Annotation: This tweet was labelled differently by all three annotators: positive, negative, indeterminate.
- *Quotes*: these come in the form of inspirational quotes that convey usually positive meaning, but they are not an explicit opinion about a specific target. We suggest that this type of text is considered neutral since it is not conveying sentiment.
  - *Ex:* على المرء أن يحاول إعداد نفسه للحياة دون سعادة سواء جاءت أم لم تجئ. - جورج إليوت درر - الكلام
    Translation: Whether happiness may come or not, one should try and prepare one's self to do without it ..George Eliot
    Annotation: This tweet was labelled differently by all three annotators: positive, neutral, indeterminate.
- *Rhetorical questions*: these could be queries about a certain topic or requests and therefore should be considered neutral. However, sometimes questions may be due to frustration and thus convey negative sentiment. We suggest a clear guideline about questions that shows the annotator how to differentiate between the two cases.
  - *Ex*: اللي يعرفلي اخصائي اسنان ممتاز في المدينة يتواصل معايا ضروري لاهان .
    Translation: If you know an excellent dentist in Madina please contact me necessary.
    Annotation: This tweet was labelled neutral by two annotators and indeterminate by one annotator.
- *Author's perspective*: the annotators stated that some tweets could be positive for a group of people but negative for another group and since the author of the tweet is unknown it is sometimes hard to determine which sentiment the author is trying to express. We suggest that this case should be labelled as indeterminate.

- *Ex*: أما قضية المقاومة الفلسطينية فهي قصة أخرى استغلتها ايران بشكل رائع وبسبب ذلك الدعم الرمزي حصلت طهران على مكاسب سياسية ضخمة
  Translation: As for the Palestinian resistance case, it is another story that Iran took advantage of in a fabulous way, and due to that figure support, Tehran got huge political advantages
  Annotation: This tweet was labelled differently by all three annotators: positive, neutral, indeterminate. We can see in this example that it is hard to determine the sentiment of the tweet without knowing if the author is a supporter of Iran or against it.

- *Determining the topic of the tweet*: the annotators stated that it was sometimes difficult to determine the topic of the tweet and accordingly they couldn't determine the sentiment conveyed. This challenge is correlated with the nature of the language on twitter that is informal and short.
- *Grouping tweets according to topics*: annotators suggested that if tweets were grouped according to the topic they are around, it would make the annotation easier. However, this is not always possible in tweets.

These challenges could be addressed in future annotation projects of Arabic sentiment by having explicit guidelines for each where applicable. Also providing examples of confusing cases could help clarify the annotation process.

7. Conclusion

In this paper, we presented the methodology we followed in collecting and constructing a large dataset of Arabic tweets. The dataset contained around 2.2 million tweets and was used to generate an Arabic corpus of tweets.

A corpus of Saudi tweets was extracted from the datasets and annotated with five labels: positive, negative, mixed, neutral, and indeterminate. The corpus was augmented with more tweets to compensate for the low number of neutral tweets. The corpus was manually annotated by three annotators and kappa statistics were calculated to ensure the reliability of the annotations. The annotation process including the labels used and the guidelines presented to the annotators were illustrated in detail. The size of the corpus has reached 17,573 tweets. To the best of our knowledge, this is the largest annotated corpus of Saudi tweets. Baseline experiments were conducted for two-way, three-way and four-way classification on the AraSenTi-tweet corpus using different term features. These experiments provide a benchmark for future work on SA of Arabic tweets.

As exemplified in previous work, sentiment annotation is a challenging task when compared to annotation of other NLP tasks. Therefore, the challenges that appeared during annotation were highlighted. These could serve as a practical guide for future annotation projects of Arabic sentiment. Moreover, the corpus will be available for the research community.

Acknowledgements

This Project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (GSP-36-332).

References

1. Habash, N.Y. (2010) "Introduction to Arabic natural language processing." *Synthesis Lectures on Human Language Technologies*. **3**: 1–187.
2. Darwish, K., Magdy, W. (2014) "Arabic Information Retrieval." *Foundations and Trends in Information Retrieval*. **7**: 239–342.
3. Habash, N., Eskander, R., Hawwari, A. (2012) "A morphological analyzer for Egyptian Arabic." *In Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*. pp. 1–9. Association for Computational Linguistics.
4. Farghaly, A., Shaalan, K. (2009) "Arabic natural language processing: Challenges and solutions." *ACM Transactions on Asian Language Information Processing (TALIP)*. **8(4)**: 14.
5. MartíNez-CáMara, E., MartíN-Valdivia, M.T., UreñA-LóPez, L.A., Montejo-RáEz, A.R. (2014) "Sentiment analysis in Twitter." *Natural Language Engineering*. **20**:1–28.
6. Semiocast Semiocast — Twitter reaches half a billion accounts — More than 140 millions in the U.S., http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US.
7. Mubarak, H., Darwish, K. (2014) "Using Twitter to collect a multi-dialectal corpus of Arabic." *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*. Doha, Qatar: 1-7.
8. Al-Twairesh, N., Al-Khalifa, H.S., Al-Salman, A. (2014) "Subjectivity and Sentiment Analysis of Arabic : Trends and Challenges." *In:*

The 11th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). Doha, Qatar: 148-155.
9. Abbasi, A., Chen, H., Salem, A. (2008) "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums." *ACM Transactions on Information Systems (TOIS)*. **26(3)**: 12.
10. Korayem, M., Crandall, D., Abdul-Mageed, M. (2012) "Subjectivity and sentiment analysis of arabic: A survey." *In: Advanced Machine Learning Technologies and Applications*. pp. 128–139. Springer.
11. alOwisheq, A., alHumoud, S., alTwairesh, N., alBuhairi, T. (2016) " Arabic Sentiment Analysis Resources: A Survey." *In: Meiselwitz G. (eds) Social Computing and Social Media. SCSM* . pp. 267–278. Springer, Toronto Canada.
12. Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., Perea-Ortega, J.M. (2011) "OCA: Opinion corpus for Arabic." *Journal of the Association for Information Science and Technology*. **62(10)**: 2045–2054.
13. Bayoudhi, A., Belguith, L.H., Ghorbel, H. (2015) "Sentiment Classification of Arabic Documents: Experiments with multi-type features and ensemble algorithms." *In: 29th Pacific Asia Conference on Language, Information and Computation*. , Shanghai,China.
14. Duwairi, R., El-Orfali, M. (2014) "A study of the effects of preprocessing strategies on sentiment analysis for Arabic text." *Journal of Information Science*. **40(4)**: 501–513.
15. Mountassir, A., Benbrahim, H., Berrada, I. (2012) "A cross-study of Sentiment Classification on Arabic corpora." *In: Research and Development in Intelligent Systems XXIX*. pp. 259–272. Springer.
16. Oraby, S., El-Sonbaty, Y., El-Nasr, M.A. (2013) "Finding Opinion Strength Using Rule-Based Parsing for Arabic Sentiment Analysis." I*n: Advances in Soft Computing and Its Applications*. pp. 509–520. Springer.
17. Aly, M.A., Atiya, A.F. LABR: (2013) "A Large Scale Arabic Book Reviews Dataset." *In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pp. 494–498. , Sofia,Bulgaria.
18. Al Shboul, B., Al-Ayyouby, M., Jararwehy, Y. (2015) "Multi-way sentiment classification of arabic reviews." I*n: 6th International Conference on Information and Communication Systems (ICICS)*. pp. 206–211. IEEE.
19. ElSahar, H., El-Beltagy, S.R. (2015) "Building Large Arabic Multi-domain Resources for Sentiment Analysis." *In: Computational Linguistics and Intelligent Text Processing*. pp. 23–34. Springer.
20. Al-Smadi, M., Qawasmeh, O., Talafha, B., Quwaider, M. (2015) "Human annotated Arabic dataset of book reviews for aspect based sentiment analysis." *In: 3rd International Conference on Future Internet of Things and Cloud (FiCloud)*. pp. 726–730. IEEE.
21. Obaidat, I., Mohawesh, R., Al-Ayyoub, M., AL-Smadi, M., Jararweh, Y. (2015) "Enhancing the determination of aspect categories and their polarities in Arabic reviews using lexicon-based approaches." *In: Conference on Applied Electrical Engineering and Computing Technologies (AEECT*). Jordan. pp. 1–6. IEEE.
22. Abdul-Mageed, M., Diab, M.T. (2012) "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis." *In: Proceedings of the eighth international conference on Language Resources and Evaluation Conference (LREC)*. pp. 3907–3914. , Istanbul,Turkey.
23. Refaee, E., Rieser, V. (2014) "An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis." *In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC*. pp. 2268-2273, Reykjavik, Iceland.
24. Nabil, M., Aly, M., Atiya, A.F. (2015) "ASTD: Arabic Sentiment Tweets Dataset." *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pp. 2515–2519. Lisbon, Portugal.
25. Ibrahim, H.S., Abdou, S.M., Gheith, M. (2015) "MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis." *In: 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, pp. 353–358. IEEE.
26. Abdul-Mageed, M., Diab, M.T. (2011) "Subjectivity and sentiment annotation of modern standard Arabic newswire." *In: Proceedings of the 5th Linguistic Annotation Workshop*. pp. 110–118. Association for Computational Linguistics.
27. Mohammad, S.M. (2016) "A practical guide to sentiment annotation: Challenges and solutions." *In: Proceedings of NAACL-HLT*. pp. 174–179.
28. Go, A., Bhayani, R., Huang, L. (2009) "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford. 1–12.
29. Pak, A., Paroubek, P. (2010) "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*. Valleta,Malta.
30. Davidov, D., Tsur, O., Rappoport, A. (2010) " Enhanced sentiment learning using twitter hashtags and smileys." *In: Proceedings of the 23rd International Conference on Computational Linguistics*. pp. 241–249. Association for Computational Linguistics.
31. Kiritchenko, S., Zhu, X., Mohammad, S.M. (2014) "Sentiment analysis of short informal texts." *Journal of Artificial Intelligence Research*. **50**:723–762.
32. Pasha, A., Al-Badrashiny, M., Kholy, A.E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., Roth, R. (2014) "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic." *In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC*. pp. 1094-1101. Reykjavik, Iceland.
33. Mohammad, S., Zhu, X. (2014) Sentiment Analysis of Social Media Texts. *In: Tutorial at the 2014 Conference on Empirical Methods on Natural Language Processing*. Doha, Qatar.
34. Fleiss, J.L. (1971) "Measuring nominal scale agreement among many raters." *Psychological bulletin*. **76** : 378.
35. Landis, J.R., Koch, G.G. (1977) " The measurement of observer agreement for categorical data." *Biometrics*. **33**: 159–174.
36. Mohammad, S.M. (2015) "Sentiment analysis: Detecting valence, emotions, and other affectual states from text." *Emotion Measurement* : 201-238.Springer.
37. Manning, C.D., Schütze, H. (1999) "Foundations of statistical natural language processing." *MIT Press*.