https://doi.org/10.34120/jgaps.v50i193.325

TweetSentKW: A corpus of multi-label emotion analysis for Kuwaiti Arabic tweets

Eiman T. Alsharhan(1)
Eisa N. Alnashmi(2)
Allan M. Ramsay(3)

Abstract

Objectives: Arabic language is primarily represented in two varieties: Modern Standard Arabic (MSA) and Dialectal Arabic (DA). With the advent of social media, there has been a shift from the predominant use of MSA in writing to the incorporation of DA, thereby generating extensive resources for dialectal text studies. Kuwaiti Arabic (KA), a sub-variety of the Gulf dialect and one of the five principal Arabic dialects, differs significantly from MSA in all linguistic aspects. KA is an under-resourced language with a notable deficiency in language resources. The development of emotion classification tools relies heavily on the availability of resources such as annotated corpora. This study introduces TweetSentKW, a multi-label emotion annotated corpus for KA. **Method**: TweetSentKW was developed by collecting tweets and selecting relevant emotional classes for the annotation process. Each tweet was annotated by three independent annotators. **Results**: The TweetSentKW corpus comprises 40,000 manually labeled tweets across various topics. Besides constructing the corpus, this study provides a comprehensive analysis of annotator behavior and the co-occurrences of emotions. The corpus is anticipated to significantly

(1) Associate Professor, Department of Arabic Language and Literature, Kuwait University.
E-mail: Eiman.alsharhan@ku.edu.kw

(2) Associate professor, Department of Mass Communication and Journalism, Kuwait University.
E-mail: Eisa.alnashmi@ku.edu.kw

(3) Professor, Department of Computer Science, University of Manchester.
E-mail: Allan.Ramsay@manchester.ac.uk

- Submitted: 31/8/2023, Revised: 9/1/2024, Accepted: 14/1/2024.

To cite p. 391

contribute to sentiment analysis research, a crucial method for gauging public opinion. **Conclusion**: The widespread use of social media platforms, such as Twitter, has led to continuous and uninhibited public expression of opinions on diverse issues. The public and archived nature of these opinions presents a rich opportunity for researchers to analyze and understand public sentiment and perspectives.

Keywords: natural language processing, sentiment analysis, emotion classification, deep learning, communication

مدونة لغوية متعددة الأوسمة لتحليل المشاعر في التغريدات الكويتية

إيمان توفيق الشرهان[1]
عيسى نشمي النشمي[2]
آلان مكغريغور رامزي[3]

ملخص

الأهداف: هدفت الدراسة لبناء مدونة لغوية للتغريدات المكتوبة باللهجة الكويتية، إذ تشكل الكتابة باللهجة المحلية تحديا لتطوير أنظمة حاسوبية لتحليل اللغة المستخدمة في مواقع التواصل الاجتماعي ومساعدة الحكومات والمؤسسات في التعرف على التوجه العام حول القضايا المختلفة ورسم السياسات المستقبلية؛ وذلك بسبب نقص الموارد اللغوية الخاصة بهذه اللهجات. **المنهج:** اتبعت الدراسة المنهج الوصفي الإحصائي وذلك في تقديم وصف تفصيلي للمعلومات والإحصائيات المتعلقة بالمدونة التي تم بناؤها، وقياس مدى الإتفاق بين المرمِّزين، ونوع المشاعر المرصودة، كل تغريدة تم ترميزها من قبل 3 مرمزين، وبمشاركة إجمالية من 468 مرمز مدرب في إنجاز هذه المهمة. **النتائج:** نتج عن الدراسة بناء مدونة لغوية خاصة بالتغريدات المكتوبة باللهجة الكويتية تحتوي على نوع المشاعر التي تحملها كل تغريدة ، مثل: حب، غضب، تأييد، اعتراض، تفاؤل، تشاؤم...، إضافة إلى بناء المدونة، قدمت الدراسة دراسة لمصداقية المرمزين كما بحث عن مدى التوافق بينهم، باستخدام نموذج جاكارد المعدل، وعلى أساسها تم استبعاد المرمزين غير الموثوقين، والوصول إلى فهم أفضل لأسباب ضعف الإتفاق في بعض المواضع. **الخاتمة:** ستكون المدونة اللغوية متوفرة بشكل كامل للباحثين المهتمين في مجال التصنيف الآلي للمشاعر للقيام بأبحاثهم المهتمة باللهجة الكويتية.

الكلمات المفتاحية: المعالجة الآلية للغة، قياس الرأي العام، اللهجة الكويتية، وسائل التواصل الاجتماعي، التعرف الآلي على المشاعر

(1) أستاذ مشارك، قسم اللغة العربية وآدابها، جامعة الكويت. الإيميل: Eiman.alsharhan@ku.edu.kw

(1) أستاذ مشارك، قسم الإعلام، جامعة الكويت. الإيميل: Eisa.alnashmi@ku.edu.kw

(1) أستاذ، قسم علوم الكمبيوتر، جامعة مانشستر. الإيميل: Allan.Ramsay@manchester.ac.uk

- تُسلّم البحث في: 2023/8/31، عُدّل في: 2024/1/9، أجيز للنشر في: 2024/1/14.

للاستشهار بهذه المقالة انظر ص. 391

Introduction

Sentiment analysis (SA) is the field of study that analyses people's opinions, sentiments, attitudes, and emotions towards entities given a piece of text (also known as opinion mining) (Liu 2012). It is a highly challenging research topic and one of the most important sub-areas in natural language processing (NLP) research. The wide spread of social media networks that contain an enormous number of opinionated texts and the increasing concern in identifying the peoples' emotions make SA an active research area.

With the rapid growth of this field of study, SA gained more names with slightly different tasks, such as emotion analysis (EA). Some researchers find SA to be different from EA (Badarneh et al., 2018). The focus of SA is in identifying polarity (negative or positive) of a given text, whilst EA focuses on identifying and analyzing the emotions conveyed in the text such as joy, fear, anger, etc. From this perspective, we can think about EA as a finer-granularity sentiment analysis and a top layer of the (relatively) simple sentiment classification.

Furthermore, a single piece of text can carry one or more than one emotion. Single- label emotion analyzing task aims at predicting only one emotion of a given text, whilst the multi-label emotion analyzing task captures all possible emotions conveyed in a given text. The main disadvantage of the single-label emotion analysis is that it neglects the other emotions that can be found in the text which makes it difficult to completely understand the author's emotional status. This disadvantage is overcome in the multi-label emotion classification.

Creating an annotated dataset is a crucial step for EA. The size of the dataset and quality of the data and the annotations have a direct relation to the accuracy of the classifier that will be developed. The manual effort involved in creating a dataset and annotating a large

number of training examples is one of the bottlenecks in developing an emotion classifier as researchers confirm (Liu , 2012).

Creating an annotated dataset for Arabic is rather difficult. This is mainly caused by the dialectal varieties that are found in Arabic. The Arabic language has a diversity of forms, with a significant split between Modern Standard Arabic (MSA), which is restricted to formal use, and dialectal Arabic (DA), which is used in daily communication. In addition, Arabic dialects are extremely diverse to the point that it can be argued that they are distinct languages instead of dialects of the same language. Despite the fact that all Arabic varieties are considered to be one language, substantial differences can be observed between Arabic dialects at all linguistic levels, including phonetics and phonology, morphology, syntax, and lexicon. The vast difference between Arabic dialects makes it unpractical to use a dataset that is created for a specific dialect to develop a language processing tool for another. This means that it is important to create a dialect-specific dataset prior to developing an emotion classifier.

This study is concerned with creating an annotated dataset for Kuwaiti Arabic (KA). KA is the local colloquial version of MSA spoken in Kuwait. It is a sub-variety of the Gulf dialect which is one of the main five Arabic dialects. It differs from MSA in all linguistic levels. It is an under resourced language for which there is a real lack of all kinds of language resources.

Twitter is a rich source for opinionated text, which makes it the perfect data source for emotion classification tasks. The access of data in twitter is easy and free for academic researchers with a complete archive of historical public tweets, thanks to the academic research product track released by twitter. Twitter is one of the top social media applications used in Kuwait with a total of 1.45 million

active users in early 2022[1]. This huge volume of data produced daily provides a great opportunity for researchers in the field of sentiment analysis to observe the public emotions, opinions, and attitude towards products, policies, or entities.

The contribution of this study can be highlighted in two main points: a) creating a large resource for KA and b) evaluating the resource. This resource is a public dataset that is intended to accelerate research in KA text mining. To our knowledge, this is the only dataset annotated with emotions available for KA. This resource is big in terms of size (40K tweets) and rich in terms of diversity of emotions (nine emotional labels).

In addition, the fact that annotators do not always agree in assigning an emotion to a tweet requires further investigations to find the reasons behind the disagreement. A thorough investigation is carried out in the study to evaluate the annotators' behavior and to investigate which emotions tend to occur together.

The field of multi-label emotion classification has attracted researchers in recent years due to its potential applications in various domains. For instance, it has a wide range of applications in e-learning, policy making, health care, marketing, etc.

This study begins by reviewing the work done in the literature in the field of SA focusing on the works that target Arabic language. It will then go on to highlighting the challenges in processing tweets. The fourth section presents the process of creating the multi-label emotion corpus starting with the data collection step, then introducing the emotional classes used in the annotation process, and finally explaining how the annotation is performed. The fifth section is

(1) This is the number published in Twitter's advertising resources.

concerned with finding the reasons for disagreement between annotators. It is also concerned with the evaluation of the annotators' performance. Many tactics are presented and discussed in this section to reduce disagreement and remove unreliable annotators. In the sixth section the study tried to interpret the reasons behind the disagreement between annotators. A thorough investigation is carried out in the seventh section to find out what emotions tend to occur together and why. The distribution of the labels among the nine emotions is given in the eighth section.

Related work

Research in the field of SA requires access to an annotated corpus that is large in terms of size and good in terms of quality. In recent years, there has been an increasing amount of literature on creating datasets that focus on annotating Arabic text with sentiments. However, those datasets have many limitations in size, the restriction to a specific dialect or MSA. In addition, most of the works focused on polarity classification or single-label emotion classification tasks. Moreover, most of these datasets are in-house datasets that are not available to the public. Below we give a short review of the recent works toward building annotated Arabic corpus.

We can categorize the works found in the literature into two main groups: the first group, which is the dominant group, focuses on sentiments annotation (positive, negative, or neutral) and the second group focuses on the emotion's classification (love, anger, fear, etc...). Both categories are conducted either automatically or manually.

For sentiment analysis, Abdellaoui & Zrigui (2018) and Kwaik et al. (2020) applied an automatic approach to collect a dataset with sentiments. Abdellaoui & Zrigui (2018) presented a large dataset for Arabic sentiment analysis, which is annotated automatically using

emojis in tweets and sentiment lexicons (TEAD). The researchers adopted a distant supervision algorithm for automatically collecting and labeling more than 6 million tweets labeled as positive, negative, or neutral. The presented algorithm is used to deal with mixed-content tweets MSA and DA). In a similar manner, Kwaik et al. (2020) collected an Arabic Tweets Sentiment Analysis Dataset (AT- SAD). Researchers adopted an automatic annotation approach using emojis for 36K tweets labelled positive and negative by employing distant supervision and self- training approaches. In addition, 8K tweets are manually annotated as a gold standard. The corpus is evaluated by comparing the emoji-based annotation with the human annotation and 77.2% agreement is reported.

On the other hand, most works found in the literature adopted the manual approach in collecting the dataset for sentiments. For instance, we have two small datasets reported in the work of Refaee & Rieser (2014) and Nabil et al. (2015). Refaee & Rieser (2014) collected 8,868 multi-dialect annotated Arabic twitter feeds. The study focuses on subjectivity and the overall contextual polarity of the tweet. The tweets are manually annotated by two annotators and the annotations indicate good inter-annotator agreement. The reported dataset is available to the public. Similarly, Nabil et al. (2015) presented the Arabic Sentiment Tweets Dataset (ASTD). It contains 10K tweets written in Egyptian Arabic annotated with the four-way sentiment classification.

More recently, Al-Twairesh et al. (2017) developed an annotated corpus which contains 17,573 of tweets that are written in MSA or in Saudi Arabic. The annotation process was done manually with four sentiments labels: positive, negative, neutral, and mixed. Another work for Saudi Arabic is presented by Al-Thubaity et al. (2018) it is a small dataset

that comprises 5,400 tweets of Saudi dialects and Modern Standard Arabic classified for both sentiment analysis and emotion analysis. They employed three annotators to classify each tweet according to its polarity and the emotion it carries using Ekman basic emotions. Their annotations show moderate agreement as the average agreement among any two annotators is 65%, the average kappa for any two annotators is 0.55, and Fleiss' kappa for the three annotators is 0.55.

Two small corpora of manually annotated texts are reported by Atoum & Nouman (2019) and Qwaider et al. (2019). The first one contains only 3,550 Jordanian dialect tweets with sentiments distribution as follows: 616 positive tweets, 1,313 negative tweets, and 1,621 neutral tweets. The second one targets Levantine Arabic with ap- proximately 2,500 posts from social media sites in general topics. Those feeds are annotated manually with positive, negative, and neutral labels.

Mohammed & Kora (2019) reported a manually annotated corpus of 40K tweets with two sentiment polarities: positive and negative polarity. The corpus includes tweets written in MSA or Egyptian Arabic form in several domains. The validation of the annotations was done manually by two different experts who were asked to check the annotations. Both experts were consistent with the annotations with 100% ac- curacy. The main throwback of this corpus is that with the extensive cleaning and pre-processing stage, it turns to be a very hard crafted corpus where the tweets are different than the real tweets. The manual selection of tweets is constrained to those tweets that are clearly interpreted into positive or negative, excluding all tweets with multiple emotions and neutral.

ASAD is another new and large dataset for Arabic sentiment Analysis (Alharbi et al., 2021). The dataset contains 95K manually

annotated multiple-dialects tweets with three-class sentiment labels (positive, negative, and neutral). Each tweet is an- notated by 3 or 4 annotators. The overall Fleiss Kappa coefficient of ASAD is = 0.56 which indicated moderate agreement among the annotators. The dataset is publicly available for the research community.

In another effort towards the development of a benchmark dataset for sentiment analysis, Alowisheq et al. (2021) presented MARSA a large sentiment annotated corpus for Gulf dialect Arabic that consists of 61,353 manually labeled tweets in multi domains. One of the main advantages of MARSA is that it is a multi-domain corpus covering the following domains: sports, politics, technology, and social issues. This makes it possible to create domain-specific classifiers which leads to enhance performance of sentiment analysis. The tweets were annotated by eleven annotators who work on classifying the tweets into five labels: positive, negative, neutral, sarcasm, and both.

Going beyond the task of mainly classifying tweets into positive or negative categories, several works are reported that focus on detecting the emotions found in the text as will be reviewed below.

The work presented by Hussien et al. (2016) follows an automatic approach to deal with the lack of resources for Arabic emotions classification. The reported approach is based on the use of emojis to annotate data with emotions. Only four emotion classes (joy, sadness, anger, and disgust) were used to categorize a total of 134,194 Arabic tweets. Tweets were labeled based on the type of emojis used in the tweet. The automatic annotation approach resulted in 10,467 joy tweets, 7878 sadness tweets, 2874 anger tweets, and 1533 disgust tweets.

Few more works are found in the literature that aim at introducing Arabic datasets for emotions detection using a manual approach. El

Gohary et al. (2013) employed child's stories to build their dataset. Their data is composed of 2,514 sequential sentences which were later annotated using six emotion classes: surprise, disgust, anger, fear, sadness, and happiness, in addition to neutral class.

Rabie & Sturm (2014) worked on Egyptian Arabic tweets to developed both: a corpus of emotion annotated tweets as well as a sample word-emotion lexicon. Ekman's basic emotions are considered in annotating the corpus. The total number of tweets used in this study is 1,776, each of which annotated by an average of 15 human annotators. These tweets were filtered to exclude those with less than 50% agreement, which leads to reducing the number of tweets to 1,605 tweets. In addition, the researchers extracted the sets of words which are highly correlated with each emotion aiming at presenting a seed for building emotional lexicons in the future Abdul-Mageed et al. (2016) created DINA, a multi-dialect dataset for Arabic emotion classification. The dataset contains about 3k tweets annotated with the six main emotions by two annotators. In addition to annotation the data with the emotional classes, each tweet is assigned another label to describe the emotion intensity as low, medium, or high. Following the annotation process, the inter-annotator agreement was measured for emotion labels. It was found that the annotators' agreement was higher in certain classes compared to others; the agreement level was 0.71 in happiness, and 0.23 in fear, but on average it was 0.51, which indicates the difficulty inherent of emotion classification.

Another effort towards building an Egyptian Arabic dataset for emotion classification is presented by Sayed et al. (2016). The authors used Ekman's set of emotions in analyzing the tweets' emotions with some modifications. For instance, they merged anger and disgust in one class and the same for sadness and fear. Additionally, they

added sarcasm as an emotion class. The manual annotation process resulted in building a corpus of 10,177 tweets each annotated by three specialized annotators. In addition to building an annotated emotion corpus, the authors constructed a lexicon of 563 emotional Arabic words from twitter.

Al-Khatib & El-Beltagy (2017) presented the work of building Egyptian Arabic dataset that contains 10,065 tweets. The emotion classes applied in this dataset are sadness, anger, joy, surprise, love, sympathy and fear in addition to the "no emotion" class.

In their 2018 study, Badarneh et al. constructed an additional dataset comprised of Egyptian Arabic tweets. This dataset was specifically annotated to facilitate the analysis of emotions, addressing the task as both a multi-label and multi-target problem. They considered Ekman's basic emotions and presented their efforts towards building and annotating a dataset of 11,503 tweets. The dataset is annotated from two perspectives: the writer perspective and the reader perspective, which leads to generating a particular dataset for each perspective. Two specialized human annotators participated in the annotation process and Cohen's Kappa measure was calculated to determine their concordance. The study found that in the writer dataset, the most agreement value is on the joy class (0.616) while the least agreement is on the sadness class (0.414). In the reader dataset, the most agreement is on the fear class (0.667) while the least agreement is on the surprise class (0.400).

As obvious from the previous discussion, the available Arabic corpora or datasets reported in the literature lacked some aspects. They have limitations in the size, or they target a specific dialect. In addition, most of the work focuses on identifying polarity rather than emotions. Even the works that focus on emotion classification, most of them are found to follow a single-label emotion classification.

The work reported in this study is the first work, to the best of our knowledge, that targets KA to build a large and fine-grained dataset for multi-label emotion classification.

Challenges of processing tweets

Processing texts must adapt to the type of text. Researchers confirm that text found in social media, especially tweets, shows numerous challenges when compared to formally structured text such as the one found in newspapers and scientific journals (Alwakid et al., 2017). These unique features of tweets must be understood before processing them. Following is the main characteristics of tweets:

- The short text length is due to the constraints on the number of characters in the tweet (280 characters). This leads users to employ abbreviations in the tweets to make room for other words.
- Tweets are an example of user-generated content, which is hard to handle because of the unstructured language found in its contents, spelling mistakes, use of abbreviations, a lot of ironic and sarcastic sentences, and slang words.
- Most of the tweets are written in dialectal Arabic, which lacks standard orthography. This results in having texts with lots of spelling inconsistencies, which sometimes leads to difficulties in understanding the tweet.
- Most people write tweets as they speak. Intonation can help understand the meaning while speaking. Some twitter users are not aware of this fact which leads to misunderstanding of their tweets.

Creating TweetSentKW

This section gives an overview of the creation of TweetSentKW the corpus with multi- label emotion analysis for KA.

Data collection

The research used twitter search Application Programming Interface (API) (academic track) for corpus collection. Only tweets that are published from Kuwait are collected. The extracted tweets cover several topics such as social topics, sports, political issues, health, sarcastic jokes, poetry, prayers, and individuals' opinions concerning different topics. Tweets are randomly retrieved over the period from January 2021 to April 2021.

After collecting the tweets, preliminary data cleaning stage is conducted to ensure minimum data quality level. This includes the following:

- removing duplicate tweets
- removing irrelevant content, such as user mentions and URLs.
- removing empty and meaningless tweets such as those contain only URLs or non-Arabic tweets.

Unlike some previous works (such as MARSA), we did not delete emojis when we cleaned the data due to their important role in understanding the author's emotions. The total number of tweets in the dataset is 40k tweets.

Emotion classes

Most work on emotion detection uses Plutchik (2001)'s emotional model. Plutchik's model has two obvious attractions: (i) because it has become a de facto standard, using it makes work based on it re-usable by other people and comparable with other people's work; and (ii) the 'wheel of emotions' (Figure 1) is extremely easy to understand and work with. We have taken five of the primary emotions ("anger", "fear",

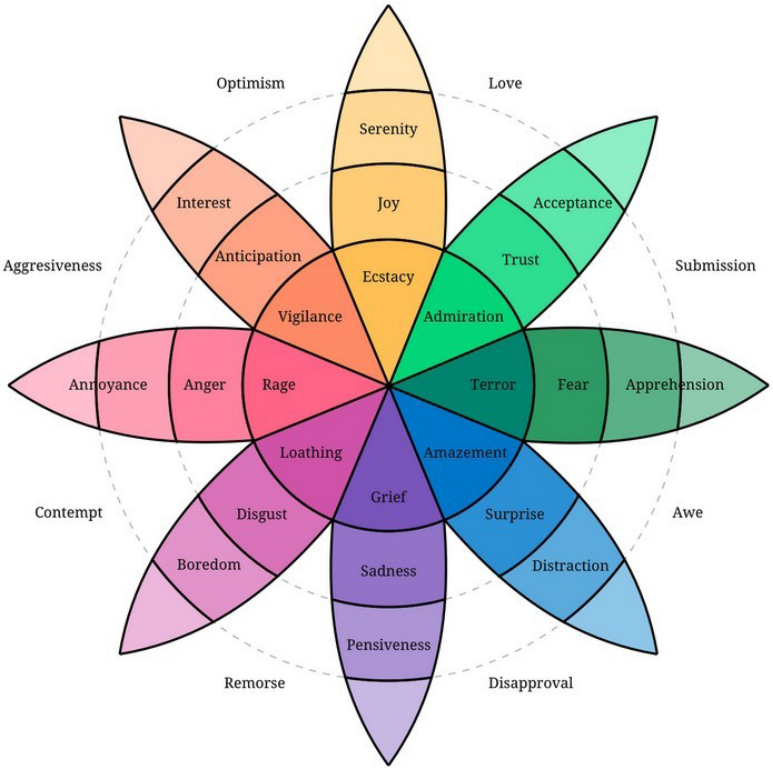
"joy", "trust", "disgust" and two of Plutchik's dyads from adjacent pairs ("love", "optimism"). Given that opposite petals in Plutchik's wheel are supposed to represent opposite emotions, we also added "pessimism" (as the opposite of optimism) and "dissatisfaction", which seems a more obvious opposite to "trust" than "disgust" as given in the original wheel. As Plutchik himself notes:

"Over the centuries, from Descartes to the present, philosophers and psychologists have proposed anywhere from 3 to 11 emotions as primary or basic. All the lists include fear, anger and sadness, most include joy, love and surprise. **There is no unequivocal way to settle on a precise number** (our emphasis), although factor-analytic studies, similarity- scaling studies, child-development studies and cross-cultural studies are useful" (Plutchik 2001)

As Plutchik notes, there are hundreds of words for emotions in English, and the labels in Figure 1 are not necessarily the definitive terms for the positions in the wheel they denote. Given that our annotators were Arabic speakers annotating Arabic texts, we chose a set of Arabic terms for labels, namely: غضب for "anger", استياء for "dissatisfaction" تشاؤم for "pessimism", خوف for "fear", رفض for "disgust", حب for "love" سعادة for "joy", تأييد for "trust" and تفاؤل for "optimism".

Figure 1

Plutchik's Wheel (picture from www.simonwhatley.co.uk/writing)



We allowed annotators to assign multiple labels (or none) to a tweet, to cope with the fact that some tweets express more than one emotion (a surprising number expressed "fear" and "love" at the same time: we will return to this below), and some are completely neutral and do not express any emotions. Given that we allowed annotators to assign zero or more emotions to a tweet, there was no point in explicitly including a label for "neutral", and indeed doing so would have led to confusion with some annotators possibly assigning zero emotions to a tweet and others assigning it the label "neutral".

Annotations

Annotations were performed by three annotators (A, B, and C). All the annotators are native Kuwaiti Arabic speakers who are undergraduate or postgraduate students at Kuwait university. Annotators were asked to label a minimum of 250 tweets.

We developed a user-friendly interface for the annotators to label the tweets. A screenshot for the platform of our project is given in figure 2. The first page of the interface includes explicit guidelines to help annotators understand the assigned job.

Figure 2

A Screen Shot for the Annotation Interface

Annotators were asked to answer three questions under each tweet:

- specify the language of the tweet (KA- MSA-another dialect-non-Arabic)
- specify the type of the tweet (textual tweet- poetry- Quranic verse and prays, advert).
- pick all the emotions found in the tweet from the author's perspective.

All emotions are defined, and we gave an example for each emotion to help annotators in making their choice. A total of 129,695 emotional labels were gathered by the end of annotation work. Table 1 gives details for the total number of assigned labels distributed over the 9 emotions. It can be noted that the distribution of emotion is unbalanced. However, we do not want it to be balanced. We want data that carries the natural distribution.

Table 1

The Distribution of the Emotional Labels in the Developed Corpus

Anger	Disgust	Dissatisfaction	Fear	Joy	Love	Optimism	Pessimism	Trust
9460	5857	25401	6061	21834	34040	21913	7807	17735

Inter-Annotator agreement

Given that assigning emotions to tweets is highly subjective, it is extremely hard to tell whether disagreements between annotators arise because one or more of the annotators is not taking the task seriously or because the tweet in question is only faintly indicative of some emotion(s). Consider the following case:

دواء البندول من أهم الادويه لأعراض التطعيمات ، فقد من السوق الكويتي، وزارة الصحه عجزت عن توفير هذا العلاج، من صاحب المصلحه في انقطاعه، ولماذا وزارة الصحه عاجزه عن توفيره مع انها صرفت مئات الملايين منذ الجائحه الموضوع بحاجه الي تدخل الوزاره لتوفيره.

English: Panadol medicine is one of the most important medicines for the symptoms of vaccinations. It was discontinued from the Kuwaiti market. The Ministry of Health was unable to provide this treatment. Who is responsible for this interruption, and why is the Ministry of Health unable to provide it even though it has spent hundreds of millions since the pandemic started. This situation needs the intervention of the Ministry to provide it.

Annotator 2211132850 says dissatisfaction., Annotator 2211116770 says dissatisfaction + anger. Annotator 2211120164 says anger.

It would seem reasonable to accept that this tweet expresses both “anger” and “dissatisfaction” – these two emotions are not all that different, they are each suggested by two of the three annotators, taking the majority view seems like a sensible thing to do.

There are, however, plenty of much less clear-cut examples, e.g.

13104: أنا قايل لكم الوزير ما يدري وين الله قاطه هههههه

English: I told you that this minister is clueless 😂

Annotator 2211132850 says joy, Annotator 2211129012 says joy,

Annotator 2211129416 says dissatisfaction.

13112: اللهم سهل على هذا اليوم و أزيل مشقته

English: Oh god, make this day easy for me.

Annotator 2211132850 says fear, Annotator 2211113293 says optimism, Annotator 2211129416 says optimism.

The majority vote for 13104 is for “joy”, the majority vote for 13112 is optimism, but in each case, there is a clear dissenting voice. Should we accept the majority, or should we only accept an emotion if all three annotators agree on it? There are also some cases where one annotator has clearly said something odd:

12970: في احد يدرس تربيه خاصه رياضيات بالتطبيقي عندي سوال

English: Does anyone teach mathematics in the institution? I have a question.

Annotator 2201114637 says no emotion, Annotator 2211117115 says no emotion, Annotator 2211129416 says:anger+dissatisfaction+ pessimism+fear+disgust+love+joy+trust+optimism.

It seems very unlikely that Annotator 2211129416 really thought that this tweet expressed all those emotions at the same time. This seems much more likely to be either a case of the annotator getting fed up with the task and just pressing all the buttons at once or of them making a mistake on this example. Inspection of the data suggests that most of the time people who produce a set like this behave perfectly reasonably the rest of the time, so we cannot just take this as evidence that they are not treating the task as a whole seriously.

We thus have two problems: some people seem to be basically unreliable, and some people are usually reliable but sometimes do rather odd things. While it is true that assigning emotions to a text is a subjective task, and hence there will inevitably be a certain amount of disagreement, we would like to weed out the most egregious cases. We therefore took the following steps:

1 - We removed the annotators who had done the fewest tweets. We asked our annotators to do 250 tweets each, and whilst not everyone did that many, we assume that people who did not do very many did not really take the task seriously.

2 - We removed any annotations that assigned more than six emotions to a single tweet. As noted above, inspection of tweets with more than six emotions suggests that such assignments are generally nonsensical, and while the people who make them are often reasonable in their other tweets these ones should be removed.

3 - We removed people who seemed not to agree with their co-annotators. This was slightly tricky, since we do not have an absolute measure of how much someone can disagree with other people while still being a reliable annotator. There are three issues to be considered here:

a) We have taken care to make sure that annotators do not fall into cliques, i.e., that each annotator has a range of co-annotators. This means that the usual measure of inter-annotator agreement, Fleiss (1971)'s κ, is not the best thing to use. We therefore used Krippendorff's α measure (Krippendorff 1970), which is said to be more reliable when you have a lot of missing datapoints, which is what happens if each annotator has a mixture of co-annotators (if A, B, C, …X, Y, Z are annotators and some tweet has A, B and Z as annotators then the opinions of B, C, …, X, Y on that tweet are missing)2.

b) Annotators are allowed to assign zero or more labels to a tweet. κ and α are both usually used with data where each observation has been assigned a single label. We follow Krippendorff (1970) by using a distance measure to evaluate how dissimilar two labels are. We use a variant of Jaccard distance in which the distance between an empty assignment and anything else is 0: the point here is that an empty assignment will not lead to any incorrect assignments when we take the majority view – if A and B assigned "anger"

to some tweet and C did not assign anything then it is safe enough to assume that it expressed anger to some degree.

c) There is no standard table of estimates of how good (or bad) the agreement between one annotator and their co-annotators is. For standard κ there are some rather vague statements such as 'a score of 0.41 – 0.60 denotes moderate agreement'. Once we introduce distance measures and multiple annotators, even vague statements of this kind are unavailable. What we can say is that a low value of α is worse than a high one, so we simply discard the 10% of annotators with the lowest α score when compared to their co-annotators.

Table 2 shows the effect of these filters. We start with 38578 tweets and 468 annotators. Each tweet is annotated by three annotators, with 129695 annotations in total, i.e. the average number of labels assigned to a tweet by an annotator is 129695/(3 38578) = 1.12 (recall that an annotator may assign no labels at all or more than 1 label to a tweet, so there is no reason to assume that the average will be 1). We considered two ways of making use of the labels that were assigned to a tweet, saying either that the tweet expressed a given sentiment if all the annotators had assigned it that label or that we simply required most of them to have done so. A high proportion of the majority vote cases come where the third annotator says nothing at all. We have included the scores for these as majority+ – it seems reasonable to suppose that having A and B say "love" and C say nothing at all is better evidence for a tweet to evoke love than having A and B say "love" and C say "hate". We include the overall α-score for the set of annotations being used – the higher the value α the more the set of annotators agrees.

4 - We use the implementation of α from the NLTK agreement package by Tom Lipincott, which is in turn based on Artstein & Poesio (2008)'s discussion of inter-annotator agreement measures.

The first line in Table 2 shows that if we use all the annotations provided by any of the annotators, we get 10711 emotions assigned by all three annotators, 36045 if we accept a majority vote and 24685 if we accept the majority in cases where the third annotator said nothing (majority+). This set of annotations has an overall α-score of 0.28. As noted above, it is unclear what a good α-score using modified Jaccard distance would be, but this provides a baseline for comparing with the reduced sets. majority+ seems like the best option, with considerably more labels assigned than using unanimity as a threshold but at no great risk of introducing errors (all the extras are cases where two of the annotators said the same thing and the third said nothing at all).

The next two lines (lazy10, lazy50) were obtained by ignoring anything done by the 10% (50%) of annotators who did the fewest annotations. This corresponds to case (1) above. Removing the worst 10% improved the overall α-score slightly, which suggests that we had indeed removed some people who were unreliable annotators, without decreasing the number of labels assigned by very much. Removing the worst 50% produces no further improvement in α, but does lose quite a lot of cases.

Table 2

The Effects of Removing Annotators and Annotations

Filter	#tweets	Annotators	Annotations	α	Unanimity	Majority	Majority+
none	38578	468	129695	0.28	10711	36045	24685
lazy10	38577	422	129695	0.29	10706	36030	24673
Lazy50	38577	234	125752	0.29	9720	34632	23319
α10	38577	422	129691	0.29	10711	36045	24685
α50	38577	234	119819	0.30	8675	32587	21622
both	38577	397	128491	0.29	10569	35683	24438

The next two (α10, α50) were obtained by deleting the annotators for whom the α score over all the tweets that they annotated was in the bottom 10% (bottom 50%) (case 2 above). It is difficult to distinguish between an annotator who is unreliable themselves and one for whom the set of other people who co-annotated their tweets happen to be unreliable, but taking the α-score for the tweets that they were involved in and comparing that to the same score for other people seems like a reasonable surrogate. Deleting people with low personal α-scores does improve the overall α- score, but that is more or less inevitable. The numbers of labels assigned are very similar to those assigned using the lazyn tactic.

Given that each of the main tactics produces some improvement in the overall α-score without decreasing the numbers of labels found dramatically, we tried using both together. There was some overlap between the annotators who did not annotate many tweets and those with a poor individual α-score, with 25 of the 46 people from the 10% of people who did the fewest annotations also falling in the group of 46 people with the worst α-scores. There was no further improvement in the overall α-score, but there was a further decrease in the number of labels assigned.

It is, as noted earlier, not possible to say that one of these tactics produces more accurate results than another, because there is no way of knowing what the 'right' answer is – asking yet another annotator will not work, since there is no reason to believe that their judgement is any better than that of the existing ones. Removing annotators who we suspect of being unreliable does seem to slightly increase the overall α-score, which is the best way we have of measuring the level of agreement among the annotators, without dramatically decreasing the number of emotions found. Using majority+, i.e., demanding that either all three annotators agreed or that two did and the third said

nothing at all, gets a substantially higher number of emotions than simple unanimity based on reasonably persuasive evidence, and we would recommend this strategy overall, possibly with one or other (or possibly both) of lazy10 and α10.

Understanding the annotators' behavior

Generally speaking, the agreement among our annotators is considered to be good. However, there are many cases where annotators disagree. In this section, we try to interpret the reasons behind the disagreement between the annotators. For this purpose, we extracted all the cases where the three annotators disagree, and we try to study these cases and find the reasons behind the disagreement in determining the emotions. In general, it is well known that the task of predicting the emotions found the in text is a non-trivial task. A previous study shows that the annotators could not identify the emotions of 39.70% of the tweets. The researchers refer to the low agreement to the difficulty of emotion classification task (Al-Thubaity et al., 2018).

Another reason for disagreement between annotators is the natural subjectivity of this task. In fact, the text can be viewed from different points of view depending on the background knowledge of the annotator and the contextual information that the annotator has in hand. This subjectivity, in turn, translates the low agreement between annotators in some tweets.

In addition to the difficulty and subjectivity of the task, the following reasons were found to have an important role behind the low inter-annotator scores in some cases:

- Some tweets are poetic verses, which are hard for annotators to understand the hidden meanings due to the indirect expression nature of poetic. In addition, the emotions are sometimes

shaded in poetic which makes it a hard job to find the emotions. Poetic verses comprise 19.5% of the dataset.

- Some tweets have many sentences, each carries a different meaning with different emotions. Some annotators are found to assign only the prominent emotion, whereas some others are found to assign all possible emotions that found in the tweet. Example:

سنة 2021 على الصعيد النفسي والعاطفي مؤلمة بالنسبة لي فقدت ناس أحبهم ورحيلهم موجع عسا الله يغفرلهم ويرحمهم وعلى الصعيد العلمي صرت عضو بالجمعية الملكية للصحة العامة البريطانية ووقعت عقود مع جامعات خليجية لإلقاء المحاضرات والدورات بصراحة ٦ أشهر كانت راحة و ٦ أشهر كانت سفر وشغل.

English: The year 2021, on the psychological and emotional level, is painful for me. I have lost loved ones and their passing is painful. May God forgive them and have mercy on them. On the scientific level, I became a member of the British Royal Society for Public Health and signed contracts with Gulf universities to give lectures and courses. Honestly, 6 months were rest and 6 months were travel and work.

- Irony or sarcasm expressions are a common way to express opinions and emotions in social media. In sarcasm expressions, there are two semantic levels: the actual meaning and the literal meaning and they are often the opposite. This makes sarcasm tweets complex and hard to interpret. This ambiguity makes it hard for annotators to detect the meaning and the emotions and leads to disagreement between annotators. Example:

أهنيك بصراحة وأصفقلك بحرارة.. بيوم واحد عفستنا ... برافو عليك

English: I congratulate you; you missed us up in one day, Bravo.

- Another source for disagreement in assigning the emotions to tweet is the short or incomplete tweets. Such as replies to tweets where the sentences are not complete, and the meaning of the tweet is vague. Example:

ايي والله و نقفل بيبان

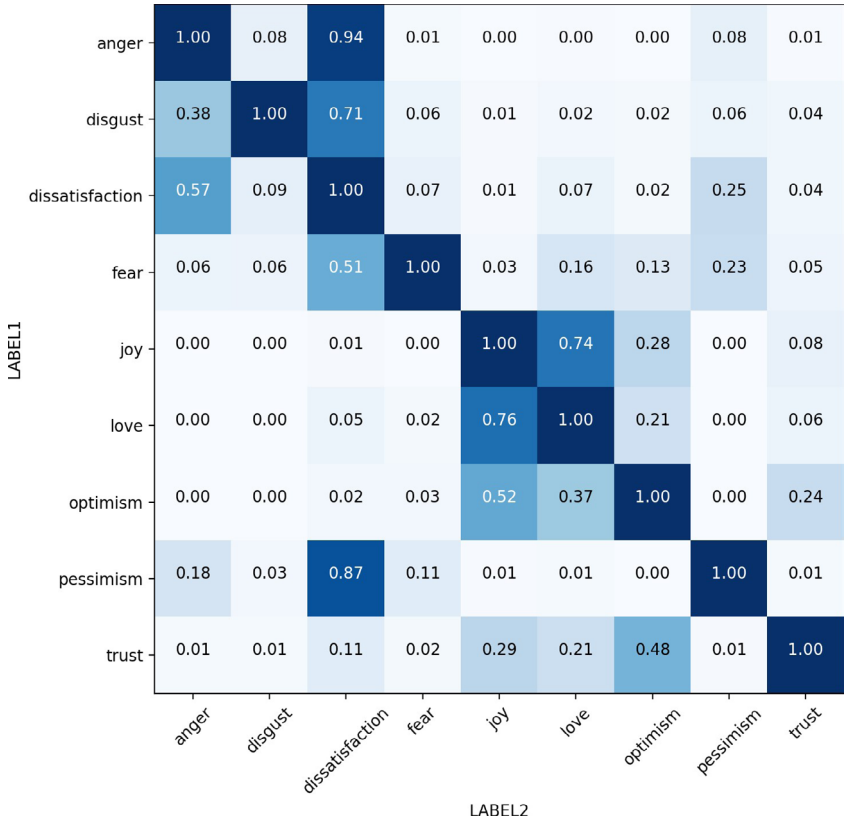
English: yes, and we lock the doors.

Cooccurrences

Given that annotators are allowed to assign multiple labels to each tweet, it is worth investigating which emotions tend to occur together. We therefore constructed a cooccurrence matrix for tweets that end up with multiple emotions assigned to them to see whether people assign conflicting emotions, and to investigate how the assignments map onto Plutchik's wheel (Figure 3). Note that this is not a confusion matrix, in the sense that the vertical axis depicts the correct label, and the horizontal axis depicts the predicted label. Indeed, as discussed above, it is not possible for us to construct a confusion matrix of this kind, since we do not (and cannot) have an independent gold standard to compare the labels assigned by our annotators to. Instead, the vertical axis, LABEL1, just enumerates the possible emotions and the horizontal one, LABEL2, specifies how often one emotion cooccurs with another in those cases where more than one emotion has been assigned to a tweet, e.g., "dissatisfaction" occurs with "anger" in 8% of cases where "anger" occurs with something else, "dissatisfaction" occurs in 94% of such cases and so on.

Figure 3

Cooccurrence Matrix for Majority Voting



The extent to which two emotions cooccur fairly accurately reflects the distance between them in Plutchik's wheel, suggesting that perhaps rather than expressing multiple emotions some tweets express a point somewhere between them, e.g., given that "love" appears as a dyad between "joy" and "trust", tweets that are labelled as expressing "joy" and "love" actually express a mixture of "joy" and "trust", with "joy" predominating. We did not find substantial numbers of tweets that expressed clearly unrelated emotions, apart

from “fear”, which cooccurs a non-trivial number of times with “love” and “optimism”, as well as with the more obviously related “dissatisfaction” and “pessimism”. The cooccurrence of “fear” with “love” and “optimism” is more likely to be seen in prayers. For instance, positive emotions are found to cooccur with the “fear” emotion in 79.3% of times in prayers tweets, compared to 6.4% in multi-topic textual tweets. This can be explained to the nature of the worship prayers which requires the presence of all three qualities: “love” of God, “hope” in his mercy and “fear” of his punishment. Take the following examples:

جدتي الحبيبة توفت... يارب رحمتك فيها وانت ارحم الراحمين

English: My beloved grandmother passed away... O God, have mercy on her, and you are the most merciful of those who show mercy.

Another surprising result is the existence of “trust” with “dissatisfaction” in many cases. We tried to investigate this further by looking at this case and we noticed that our annotators look for the emotion in the tweet from their perspective rather than the author’s perspective. In other words, annotators were choosing the “trust” label when they agree with the opinion explained in the tweet. For instance, in the following tweet the author is saying:

غلاء الأسعار مرفوض.. رحيل وزير التجارة مطلب شعبي

English: high prices are unacceptable, the departure of the minister of commerce is demanded by the public.

The assigned emotions for this tweet are:

21000621: anger+dissatisfaction+trust

22078451: anger+disgust+trust

22269833: anger

We can see that two of the annotators are expressing their support to the content of the tweet by choosing "trust" label despite that the tweet does not carry the trust emotion. The "trust" here is the annotators' emotion not the author's emotion.

label distribution

In this section we give details for the label distribution in the annotated dataset. We have 129,695 emotional labels in total. Table 3 shows the distribution of the emotional labels over the tweet's types (Quranic verse or prayer, general topic textual tweets, and poetic tweets). Generally speaking, we can see that 90% of Quranic tweets and prayers convey positive emotions. Meanwhile, the positive emotions labels appeared less often in poetic tweets where the positive emotions make up 73% of the cases. General topic tweets, on the other hand, come in the bottom of the list with only 55% positive emotions.

It can be also noticed from Table 3 that the dominant emotion in Quranic verses tweets is "optimism" which appears in about 59% of the cases, whilst the dominant emotions in textual tweets and poetic tweets are "dissatisfaction" 27% and "love" 55%, respectively.

Table 3

The Distribution of Emotional Labels Over Each Tweet Type

	Anger	Disgust	Dissatisfaction	Fear	Joy	Love	Optimism	Pessimism	Trust
verse	0.00	0.00	0.02	0.03	0.11	0.18	0.59	0.00	0.06
text	0.09	0.02	0.27	0.02	0.16	0.20	0.09	0.05	0.10
poetry	0.02	0.02	0.19	0.02	0.08	0.55	0.06	0.03	0.04

Conclusion

Aiming at extending the limited resources for Kuwaiti Arabic, this study described the creation of the first multi-label emotion dataset for KA tweets. The corpus is composed of approximately 40k tweets with a total of 129,695 tokens. We have taken five of the primary emotions from Plutchik emotional model ("anger", "fear", "joy", "trust", "disgust" and two of Plutchik's dyads from adjacent pairs ("love", "optimism") in addition to "pessimism" and "dissatisfaction". Each tweet is manually annotated by three annotators who were instructed to select all the emotions they found in each tweet. A total of 468 trained native Kuwaiti Arabic annotators participated in this task. The provided annotations were later evaluated and an α score is given for each annotator. Even though it is unclear what a good α-score using modified Jaccard distance would be, we can use it as a baseline for comparing different tactics when eliminating unreliable annotators. The study also provided a study to understand the annotators' behavior and to interpret the reasons behind the low agreement found in some of the cases. This is followed by an investigation for the cooccurrences of the emotions to find what emotions tend to occur together and why.

This corpus is available for researchers who are interested in conducting research in the area of emotion classification of KA. In future work, we plan to use this corpus to build an emotion classification model to understand the public opinion regarding the government decisions.

References

Abdellaoui, H., & Zrigui, M. (2018). Using tweets and emojis to build TEAD: An Arabic dataset for sentiment analysis. *Computación y Sistemas, 22*(3), 777–786.

Abdul-Mageed, M., AlHuzli, H., & Abu Elhija, M. D. (2016). *Dina: A multi-dialect dataset for Arabic emotion analysis*. In The 2nd workshop on Arabic corpora and processing tools (p. 29).

Al-Khatib, A., & El-Beltagy, S. R. (2017). *Emotional tone detection in Arabic tweets. In International Conference on Computational Linguistics and Intelligent Text Processing (pp. 105–114).* Springer.

Al-Thubaity, A., Alharbi, M., Alqahtani, S., & Aljandal, A. (2018). *A Saudi dialect twitter corpus for sentiment and emotion analysis. In 2018 21st Saudi Computer Society National Computer Conference (NCC) (pp. 1–6)*. IEEE.

Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., & Al-Ohali, Y. (2017). *Arasenti-tweet: A corpus for Arabic sentiment analysis of Saudi tweets*. Procedia Computer Science, 117, 63–72.

Alharbi, B., Alamro, H., Alshehri, M., Khayyat, Z., Kalkatawi, M., Jaber, I. I., & Zhang, X. (2021). *ASAD: A twitter-based benchmark Arabic sentiment analysis dataset*. arXiv preprint arXiv:2011.00578.

Alowisheq, A., Al-Twairesh, N., Altuwaijri, M., Almoammar, A., Alsuwailem, A., Albuhairi, T., Alahaideb, W., & Alhumoud, S. (2021). Marsa: Multi-domain Arabic resources for sentiment analysis. *IEEE Access, 9*, 142718–142728.

Alwakid, G., Osman, T., & Hughes-Roberts, T. (2017). Challenges in sentiment analysis for Arabic social networks. *Procedia Computer Science, 117*, 89–100.

Artstein, R., & Poesio, M. (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596. https://aclanthology.org/J08-4004

Atoum, J. O., & Nouman, M. (2019). Sentiment analysis of Arabic Jordanian dialect tweets. *International Journal of Advanced Computer Science and Applications, 10*(2).

Badarneh, O., Al-Ayyoub, M., Alhindawi, N., Jararweh, Y., et al. (2018). *Fine-grained emotion analysis of Arabic tweets: A multi-target multi-label approach*. In 2018 IEEE 12th International Conference on Semantic Computing (ICSC) (pp. 340–345). IEEE.

El Gohary, A. F., Sultan, T. I., Hana, M. A., & El Dosoky, M. (2013). A computational approach for analyzing and detecting emotions in Arabic text. *International Journal of Engineering Research and Applications (IJERA), 3*(3), 100–107.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Hussien, W. A., Tashtoush, Y. M., Al-Ayyoub, M., & Al-Kabi, M. N. (2016). *Are emoticons good enough to train emotion classifiers of Arabic tweets*? In 2016 7th International Conference on Computer Science and Information Technology (CSIT) (pp. 1–6). IEEE.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61–70. https://doi.org/10.1177/001316447003000105

Kwaik, K. A., Chatzikyriakidis, S., Dobnik, S., Saad, M., & Johansson, R. (2020). *An Arabic tweets sentiment analysis dataset (ATSAD) using distant supervision and self-training*. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 1–8).

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*(1), 1–167.

Mohammed, A., & Kora, R. (2019). Deep learning approaches for Arabic sentiment analysis. *Social Network Analysis and Mining, 9*(1), 1–12.

Nabil, M., Aly, M., & Atiya, A. (2015). *ASTD: Arabic sentiment tweets dataset*. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 2515–2519).

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist, 89*(4), 344–350. http://www.jstor.org/stable/27857503

Qwaider, C., Chatzikyriakidis, S., & Dobnik, S. (2019). *Can modern standard Arabic approaches be used for Arabic dialects? Sentiment analysis as a case study*. In Proceedings of the 3rd Workshop on Arabic Corpus Linguistics (pp. 40–50).

Rabie, O., & Sturm, C. (2014). *Feel the heat: Emotion detection in Arabic social media content*. In The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014) (pp. 37–49). Citeseer Kuala Lumpur, Malaysia.

Refaee, E., & Rieser, V. (2014). *An Arabic twitter corpus for subjectivity and sentiment analysis*. LREC (pp. 2268–2273).

Sayed, A. M., AbdelRahman, S., Bahgat, R., & Fahmy, A. (2016). Time emotional analysis of Arabic tweets at multiple levels. *International Journal of Advanced Computer Science and Applications, 7*(10). http://dx.doi.org/10.14569/IJACSA.2016.071045

Eiman Alsharhan, Associate Professor, Department of Arabic Language and Literature at Kuwait University's Faculty of Arts. Ph.D. in Natural Language Processing, University of Manchester, UK, 2014, Span linguistics, and computational linguistics. A notable publication record in prestigious journals including “Computer, Speech, and Language”, “Information processing and management”, “International journal of speech technology”, and “Language resources and Evaluation”. Active participation in international conferences across the United States, Canada, England, Spain, and the UAE. Expertise encompasses linguistics, phonetics and phonology, with a particular focus on Computational Linguistics, Speech Recognition, and Speaker Identification.

E-mail: Eiman.alsharhan@ku.edu.kw

Eisa Al Nashmi, Associate Professor, Digital Media, Department of Mass Communication and Journalism, Kuwait University. Ph.D. in New Media, University of Florida, 2011. Research interests: digital journalism, visual communication, mass communication education, media ethics and political communication. Published in highly revered academic journals relating to media studies, including International Communication Gazette, Journal of International Communication, Journalism and Mass Communication Educator, and Digital Journalism. Served as the Kuwait Visiting Research Fellow at Harvard University’s Middle East Initiative, 2022.

E-mail: Eisa.alnashmi@ku.edu.kw

Allan M. Ramsay, Professor of Formal Linguistics in the Department of Computer Science at the University of Manchester. Ph.D. University of Sussex, 1980. Ramsay's research focuses on Natural language processing, including morphology and syntax. He has published papers on the analysis of free word order languages, particularly morphology of the Arabic language, which poses a number of specific problems.

E-mail: Allan.Ramsay@manchester.ac.uk

To cite:

Alsharhan, E. T., Al Nashmi, E. N., & Ramsay, A. M. (2024). TweetSentKW: A corpus of multi-label emotion analysis for Kuwaiti Arabic tweets. *Journal of the Gulf and Arabian Peninsula Studies*, *50*(193), 257-391. https://doi.org/10.34120/jgaps.v50i193.325