

Emotion Recognition in Arabic Tweets

Mariam ElOraby¹, Rawan Reda², and Salma El-Shafey³

¹46–3294

²46–13489

³46–17002

May 25, 2023

Abstract

While sentiment analysis of Arabic tweets has been studied, emotion recognition, which identifies specific emotions expressed in the text, has received less attention. This project aims to explore the field of emotion recognition of Arabic tweets. The proposed approach will leverage BERT-based models, to detect and classify eight emotions: joy, sadness, anger, fear, love, sympathy, and surprise, or none. However, working with Arabic tweets presents several challenges due to the complexity and uniqueness of the Arabic language, including the lack of standardization in spelling and punctuation, the presence of dialects and regional variations, and the rich vocabulary and complex syntax. By preparing the data in an attempt to these gaps and challenges, we use the EmoTone_ar dataset to fine-tune several BERT-based models and compare their performance.

1 Introduction

Social media platforms have become a rich source of information for studying human behavior and emotions. With millions of users expressing their thoughts and feelings on Twitter, sentiment and emotion analysis of tweets have become a popular area of research. However, most of the existing studies on emotion recognition of tweets have been focused on English tweets, while the analysis of other languages is relatively less explored. Arabic is one of the most widely used languages on social media platforms ¹, and the analysis of Arabic tweets can provide valuable insights into the emotions and attitudes of Arabic-speaking users.

Our project aims to explore the field of emotion recognition of Arabic tweets. Emotion recognition refers to the automatic detection and classification of emotions expressed in the text. The difference between sentiment analysis and emotion recognition is that sentiment analysis focuses on identifying the polarity of a text, whether it is positive, negative, or neutral, while emotion recognition aims to identify the specific emotion expressed in the text. We are focusing on the recognition of seven emotions: joy, sadness, anger, fear, love, sympathy, and surprise, and none.

Despite the growing interest in emotion recognition of tweets, there are still several gaps in the existing literature, particularly in the analysis of Arabic tweets. Most of the existing studies on Arabic tweets have focused on sentiment analysis rather than emotion recognition.

Our motivation is to address these gaps and contribute to the field of emotion recognition of Arabic tweets by proposing an automatic approach that can detect and classify emotions expressed in Arabic tweets. Our proposed approach will leverage deep learning techniques, including several transformer-based models (BERT-based models in particular) to capture the semantic and contextual information of the text. The models will be evaluated on an Arabic tweet dataset to assess their effectiveness and robustness.

Working with Arabic tweets datasets in general has several challenges due to the complexity and uniqueness of the Arabic language. One of the significant challenges is the lack of standardization in spelling and punctuation in Arabic tweets, specially in dialectal Arabic. It is a highly inflected language, meaning that words can have multiple forms depending on their position in the sentence and

¹<https://www.extradigital.co.uk/articles/social-media/arabic-social-media.html>

their relationship with other words. Therefore, Arabic tweets may contain misspellings, abbreviations, or variations in spelling and punctuation, making it challenging to preprocess and normalize the data. Additionally, Arabic language has a rich vocabulary and complex syntax, making it challenging to develop effective models for emotion recognition. Another challenge is the presence of dialects and regional variations in Arabic tweets, which can lead to variations in meaning and expression of emotions.

2 Background

Transformers are a state-of-the-art model in natural language processing (NLP) that have shown remarkable success in various tasks, including sentiment analysis and emotion recognition. These models, such as BERT, use self-attention mechanisms to capture contextual information and have been shown to outperform traditional NLP models on a range of NLP tasks [DCLT18]. One of the main advantages of transformers is their ability to learn from large amounts of data and transfer knowledge across different tasks, making them a powerful tool for emotion recognition in Arabic tweets. By leveraging pre-trained transformer-based models, we can achieve high accuracy in recognizing emotions in Arabic tweets, even when dealing with challenges such as non-standard spelling and regional variations. Overall, transformers represent a promising approach for advancing Arabic NLP research and improving our understanding of people’s emotions and attitudes in the Arab world.

In this project we use BERT-based models to recognize emotions in Arabic tweets. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained transformer model that has achieved state-of-the-art results on a variety of natural language processing tasks. Unlike traditional NLP models that process text in a linear manner, BERT uses a multi-layer bidirectional transformer encoder that captures contextual information from both left and right contexts of the input text. Additionally, BERT utilizes a self-attention mechanism, which allows it to focus more on important parts of the input text and less on irrelevant information, leading to improved performance on tasks that require understanding long-range dependencies in the input text. The aforementioned properties allow BERT to learn rich semantic representations of text that can be fine-tuned for various downstream tasks, including sentiment analysis, question answering, and language translation. BERT-based models are pre-trained on massive amounts of data, and can be fine-tuned on specific tasks with much smaller labeled datasets. By leveraging BERT’s pre-trained knowledge, we can achieve higher accuracy on several Arabic NLP tasks such as emotion recognition or classification.

3 Dataset

We use EmoTone_ar, a publicly available dataset on HuggingFace². The collection of the dataset went through several phases to make it balanced because emotions such as fear and love were not very common in text unlike happiness and sadness [AKEB18]. The authors of the dataset conducted some standard preprocessing steps on the tweets, such as normalizing Arabic characters, removing diacritics, and eliminating links, mentions, and Retweet indicators (“RT”). Then the processed dataset was made available publicly for research purposes.

3.1 Analysis

The dataset consists of 10,065 tweets, mostly tweets written in the Egyptian dialect. The collected tweets spanned 8 emotions: neutral/none, joy, sadness, anger, sympathy, fear, love and surprise, with counts as shown in Table 1

3.2 Limitations

BERT is a large language model with over 100M parameters. Fine-tuning it needs a large dataset. However, the dataset we had found is considered very small with 10K tweets only. Furthermore, the

²https://huggingface.co/datasets/emotone_ar

dataset mostly consists of tweets about the olympics, which may make it harder to classify the emotions of tweets on other topics. The dataset also consists mostly of tweets in Egyptian Arabic. Classifying emotions of other tweets in other Arabic dialects may result in lower accuracy.

Emotion Label	Count
none	1550
anger	1444
joy	1281
sadness	1256
love	1220
sympathy	1062
surprise	1045
fear	1207

Table 1: Breakdown of emotions in the dataset. From: [AKEB18]

We wanted to analyse the most common words of each emotion category, but the dataset was noisy as expected. Thus, we first focused on cleaning and preprocessing it. The cleaning process began with removing any repeated characters like the “ي” in “إيبيه”, this is to reduce input size and an attempt to normalize the dataset further. We further removed stopwords, punctuations, and replaced some emojis with words that conveys the same mood of the emoji, and removed the rest. The complete list of emojis and emoticons considered in our project are shown in table 2. We also removed any other weird patterns through their uni-codes. After this cleaning process, we tallied the most frequently occurring words within each emotion category. This is presented in Figure 1. The counts in more details are documented in Table 5 in the Appendix.

It is worth noting that the dataset authors pointed out that a section of the dataset was obtained by filtering tweets from Egypt between July 31st, 2016 and August 20th, 2016 using the search term “أولمبياد” (Olympics). This was because during that period, people were anticipated to tweet about the Olympics using a wide range of emotions, as explained in the dataset paper. Therefore, the frequent appearance of the term in the word clouds can be attributed to this.

4 Solution Approach

Our proposed method begins with some slight preprocessing and cleaning on the tweets in the dataset to eliminate any noise or special characters. We won’t need to remove the stopwords here because they provide valuable contextual information, such as negation words.

We then proceed to tokenize the cleaned tweets using BERT’s tokenizer and fine-tune three BERT-based models for the task of emotion classification. After training, we evaluate the performance of the classification models and compare their results. Additionally, we experiment with different preprocessing techniques to observe their impact on the accuracy of the models.













































قلب	ضحك	سعادة	حزن	مفاجأة	مخرج	ملل	غضب
 <3	  :P:- P	           :D	            :(:(:'(  	  	     	     

Table 2: Emojis and the words they were replaced with.

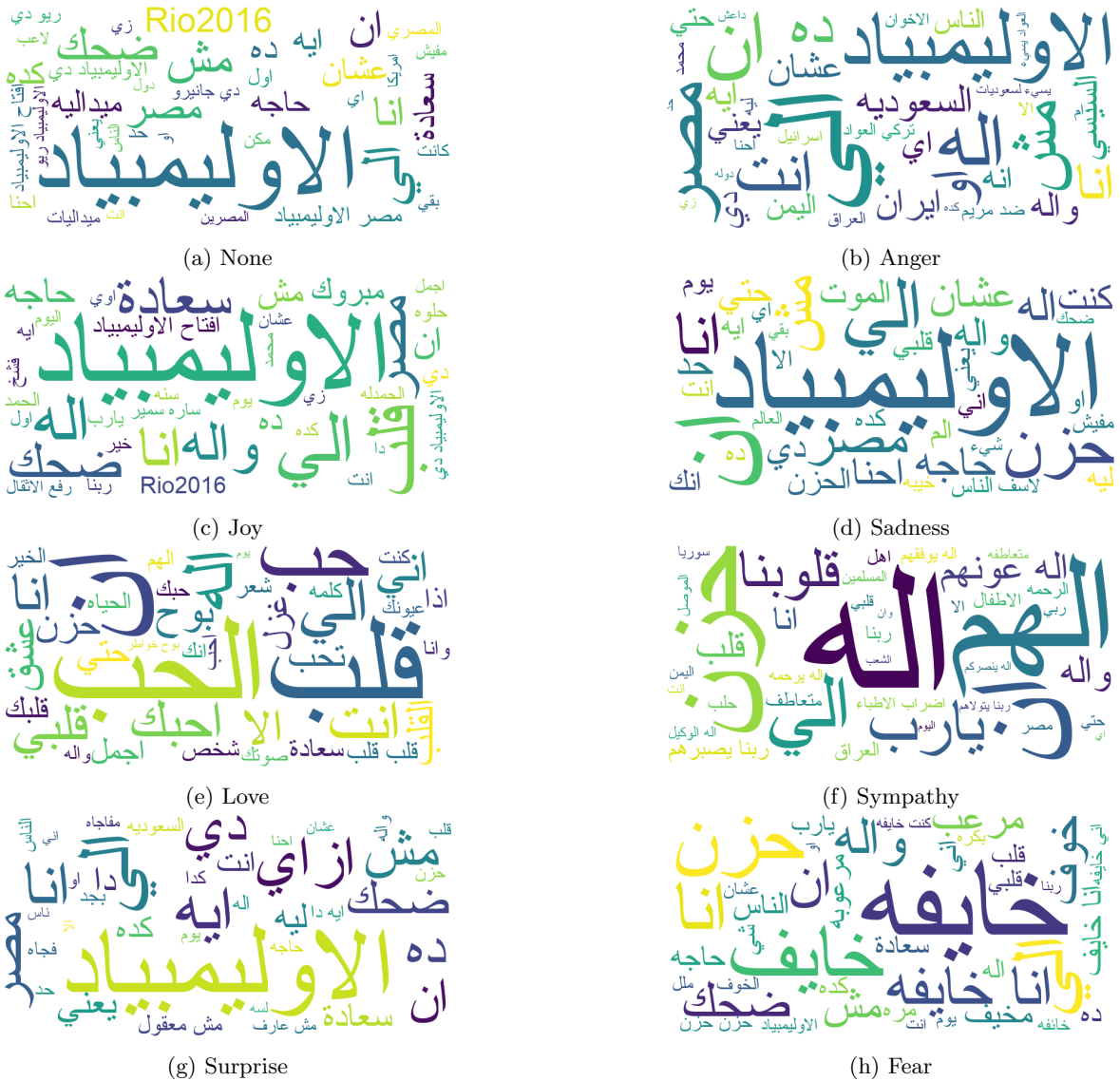


Figure 1: Visualizing emotions: wordclouds of the most frequently used words.

5 Experiments and Results

5.1 Experiments

We fine-tune seven BERT-based models on the aforementioned dataset and compare their results. The dataset was partitioned into 80% training, 10% validation and 10% test using a seed of 2023. The models we considered for this project are:

- AraBERTv2-base
- AraBERTv0.2-Twitter-base
- MARBERTv2
- CaMeLBERT-MSA (2 sizes)
- CaMeLBERT-DA
- mBERT

Details about each model’s training data and number of parameters are available in Table 6 in the appendix.

The EmoTone dataset was used to fine-tune each model twice. In the first round, the emojis were replaced as previously mentioned, while in the second round, all emojis were removed from the data.

All of the models were developed using the HuggingFace transformers library³ and fine-tuned on GPUs provided by Google, using a batch size of 32. All models were fine-tuned using AdamW with the default learning rate of $5e-5$, and epsilon value of $1e-8$. However, the number of epochs were varied to avoid over-fitting. Regarding the input length, we used 65 as a maximum length as we found that this length accounts for all tweets in the dataset after pre-processing.

5.2 Results

Model	Emojis Replaced	Emojis Removed
AraBERTv2-base	0.74 (3 epochs)	0.74 (3 epochs)
AraBERTv0.2-Twitter-base	0.78 (2 epochs)	0.78 (2 epochs)
MARBERTv2	0.81 (2 epochs)	0.81 (2 epochs)
CaMeLBERT-MSA	0.74 (3 epochs)	0.75 (3 epochs)
CaMeLBERT-MSA-16th	0.75 (3 epochs)	0.74 (3 epochs)
CaMeLBERT-DA	0.78 (2 epochs)	0.78 (2 epochs)
mBERT	0.70 (3 epochs)	0.70 (3 epochs)

Table 3: A comparison of the accuracy of each BERT model and the corresponding number of epochs used for training.

According to Table 3, it is clear that MARBERTv2 had the best accuracy compared to all the other models used. On the other hand, the lowest performance was shown with the mBERT model. The difference in model accuracy between MARBERTv2 and mBERT is 0.11. MARBERT is a large-scale pre-trained masked language model focused on both Dialectal Arabic (DA) and MSA, which makes it a good candidate for training tweets used in this project. Contrastingly, mBERT is a model that is usable with 102 languages, meaning that is not focused on Arabic semantics, which explains its low performance. Another clear observation is that the replacement/removal of emojis had no effect on the accuracy obtained after training the model. In this project, we used two versions of AraBERT and three versions of CaMeLBERT. Among the AraBERT models, AraBERTv0.2-Twitter-base leads to an accuracy that is 0.04 higher than AraBERTv2-base, one possible reason for this is that AraBERTv0.2-Twitter-base is trained explicitly for Arabic dialects and tweets, trained by continuing the pre-training using the Masked Language Model (MLM) task on 60M Arabic tweets. Among the CaMeLBERT models, we can deduce that CaMeLBERT-DA does better than both CaMeLBERT-MSA and CaMeLBERT-MSA-16th by a difference of 0.03 accuracy. CaMeLBERT-DA model was trained on Dialectal Arabic, while CaMeLBERT-MSA was trained on Model Standard Arabic. Since we are training on tweets, which mostly use dialectal language, it makes sense for CaMeLBERT-DA to have a better performance. Both versions of CaMeLBERT-MSA have similar accuracy, indicating that the size of the Modern Standard Arabic corpus used does not have a significant effect on the model performance.

We tested the MARBERT on some tweets or common expressions, some of which are sarcastic or hold no emotions, or could hold multiple emotions at once. We compared the model’s predicted output to our expected output and compiled the results in Table 4. We found that MARBERT achieved a good accuracy in correctly classifying the tweets overall, while its ability to classify sarcastic text may be limited.

³<https://huggingface.co/docs/transformers/>

Tweet Text	Prediction	What We Expected
أنا مبسوط	joy	joy
احنا هانتخرج	sadness	none
ادا احنا هنتخرج!	surprise	surprise
احنا هنتخرج!	surprise	joy or surprise
مش عارف احل الاساينمنت	sadness	sadness or anger
الأكل حلو أوي	joy	joy
الأكل مش حلو أوي	sadness	sadness
الامتحان بكرا و انا مش عارف حاجة	fear	fear
الامتحان بكرا	fear	fear
تراني تاثر	sympathy	sympathy
يعيني سهران طول الليل	sadness	sympathy
♡♡♡	love	love
!مش معقول اللي بيحصل ده	surprise	anger
وحشتيني جدا	joy	love
بمووت فيها بجد	joy	love or joy
فاكر لما كان عندك إهتمامات وشخصية قبل ما الرأسمالية تعرفك؟	surprise	sadness or surprise

Table 4: Texts, the prediction output and our expected result.

6 Suggestions for Other Approaches

So far, in this project, we have used a learning-based approach that essentially uses a trained transformer classifier to categorize tweets into their corresponding emotion classes. This approach is faster and easier to adapt to changes since it can quickly learn new features from a large training set. Other approaches include Keyword based approach, the Rule-based approach, and the Hybrid approach.

The key-based approach utilizes knowledge of important features that are paired with emotional labels through the use of a lexicon, such as Word-Net Affect or SentiwordNet. The approach applies linguistic rules and sentence structures to analyze the data. Pre-processing of the text is also necessary, which involves removing stopwords, tokenizing, and lemmatizing. Additionally, the approach evaluates keyword spotting and emotion intensity, while also checking for negation. Ultimately, this approach determines the emotional label for each sentence.

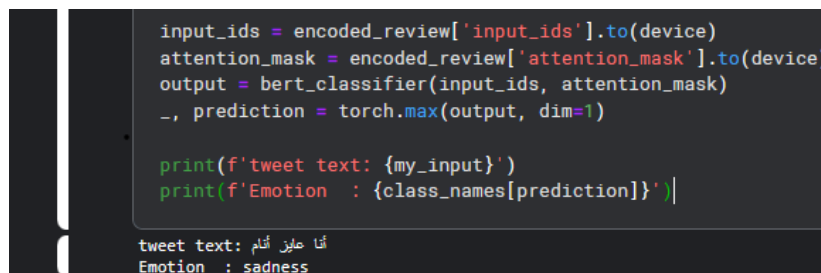
The rule-based approach is employed to manipulate knowledge and perceive information in a beneficial manner. The approach starts with text preprocessing, which involves tasks such as stop-word elimination, POS tagging, and tokenization. Emotion rules are then derived based on concepts from statistics, linguistics, and computation. The best rules are selected and subsequently applied to emotion datasets to determine emotion labels. Furthermore, appropriate rules are chosen and applied to the emotion dataset to determine the corresponding emotion labels.

Finally, the hybrid approach involves combining several approaches together for emotion prediction and categorization.

Another suggestion is to take an approach of Neurosymbolic AI for more explainable predictions. While transformers have already achieved high accuracies for sentiment analysis tasks, they are still considered a black box; it is not always clear how they arrive at their predictions. For example, when passing a text like “أنا عايز أنام” (I want to sleep) through the fine-tuned MARBERT model, the predicted emotion is “Sadness”, as illustrated in Figure 2. It’s not clear why this prediction was made. Thus, *Cambaria et al.* [CLD+22] propose the SenticNet framework; an approach that combines neural approaches including language models and kernel methods to create symbolic representations of text and build a hierarchical commonsense knowledge graph. This is a hierarchical structure where named entities or multi-word expressions are generalized in terms of concepts using linguistic patterns.

Concepts are generalized using deep learning into more specific primitives, and the primitives are generalized using logic into higher-level primitives or superprimitives. The superprimitives are then assigned a polarity. The main goal of this generalization is to allow the framework to figure out the polarity of text based on the building blocks of meaning, bringing it one step closer to natural language understanding.

SenticNet achieves higher accuracy for sentiment analysis of English text when compared to other lexicons, and the authors claim that the approach is reproducible. Thus, it would be interesting to test a similar approach for Arabic sentiment analysis or emotion classification in particular.



```

input_ids = encoded_review['input_ids'].to(device)
attention_mask = encoded_review['attention_mask'].to(device)
output = bert_classifier(input_ids, attention_mask)
_, prediction = torch.max(output, dim=1)

print(f'tweet text: {my_input}')
print(f'Emotion : {class_names[prediction]}')

```

tweet text: أنا حزين ألام
Emotion : sadness

Figure 2: Screenshot of a code passing a tweet to MARBERT fine-tuned model and printing the predicted emotion.

References

- [AKEB18] Amr Al-Khatib and Samhaa R. El-Beltagy. Emotional tone detection in arabic tweets. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 105–114. Springer International Publishing, 2018.
- [CLD⁺22] Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839, Marseille, France, June 2022. European Language Resources Association.
- [DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [Ek16] Ibrahim Abu El-khair. 1.5 billion words arabic corpus, 2016.
- [IAB⁺21] Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. The interplay of variant, size, and task type in arabic pre-trained language models, 2021.
- [ZGEL19] Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy, August 2019. Association for Computational Linguistics.

Appendix

Emotion Label	Most Common Words and Their Counts
None	الاولمبياد : 1512 الي : 293 مصر : 190 مش : 166 دي : 156 ضحك : 148 ان : 132 انا : 123 ده : 84 سعادة : 83
Anger	الي : 207 اله : 156 الاولمبياد : 144 ان : 131 مصر : 118 مش : 110 انت : 81 ده : 72 السعوديه : 72 انا : 68
Joy	الاولمبياد : 392 اله : 124 الي : 114 قلب : 98 سعادة : 86 مصر : 78 انا : 77 دي : 73 مبروك : 69 واله : 64
Sadness	الاولمبياد : 317 ان : 200 الي : 161 حزن : 159 انا : 127 مش : 113 مصر : 103 اله : 69 حاجه : 55

	عشان : 55
Love	قلب : 352 الحب : 296 ان : 280 حب : 131 اله : 112 الي : 93 بوح : 79 انت : 75 الا : 72 انا : 72
Sympathy	اله : 485 الهم : 230 حزن : 125 ان : 106 الي : 93 ربنا : 91 قلب : 83 يارب : 80 انا : 79 قلوبنا : 77
Surprise	الاولمبياد : 302 مش : 171 الي : 142 ايه : 136 انا : 128 ازاي : 89 ده : 82 دي : 81 دا : 78 مصر : 73
Fear	خايفه : 675 انا : 318 خايف : 236 حزن : 228 الي : 179 ان : 106 واله : 94 خوف : 88 مش : 86 اني : 84

Table 5: Most frequent 10 words per each emotion category, and their corresponding counts.

Model	Size	Training Data
AraBERTv2	136M	Unshuffled OSCAR corpus, Arabic Wikipedia dump ⁴ , Abu El-Khair Corpus [Ek16], the OSIAN Corpus [ZGEL19] and the Assafir news articles.
AraBERTv0.2-Twitter-base	136M	Same as v2 + 60M Multi-Dialect tweets
MARBERTv2	160M	Random sample of 1B Arabic tweets.
CAMELBERT-MSA	137M	The Arabic Gigaword Fifth Edition ⁵ , Abu El-Khair Corpus [Ek16], OSIAN corpus [ZGEL19], Arabic Wikipedia, unshuffled OSCAR corpus.
CAMELBERT-MSA-16th	137M/16	Same as CAMELBERT-MSA.
CAMELBERT-DA	137M	A collection of dialectical corpra mentioned in their paper [IAB+21]
mBERT	110M	Wikipedia data in 102 languages including Arabic

Table 6: Comparison between BERT models