Arabic Tweets Emotion Recognition

Linah Hossam Toka Ossama Nerimane ElRefai

May 8, 2023

1 Motivation

NLP, or Natural Language Processing, is a field of artificial intelligence that focuses on the interaction between human language and computers. some NLP applications involve language translation, chatbots and virtual assistants, Information Retrieval, text summarization, and sentiment analysis which is the topic of our project. It involves analyzing text to determine whether it expresses a positive or negative sentiment. Understanding the underlying attitudes, feelings, and views stated in a text is the motivation behind sentiment analysis. The importance of sentiment analysis has developed as a tool for businesses and organizations to analyze consumer feedback, social media sentiment, and other types of data as the amount of data available on the internet keeps growing.

Sentiment analysis can offer insightful information about how people feel about a particular product, service, brand, or subject. Businesses can spot repeated patterns, sentiment trends, and potential problems or concerns by examining social media posts, customer reviews, and other types of data. Social media is one of the best places to collect useful data for sentiment analysis since the new generation is writing posts, tweets, and comments to express all types of emotions, which is why our proposal focuses on the idea of sentiment analysis of Arabic tweets to address any of the per-mentioned applications. This data is useful because it is realistic, and diverse since it comes from people of different backgrounds, genders, and age categories. As shown in Figure 1, our project aims to classify tweets as a number of positive emotions, negative emotions, and neutral ones which are referred to as none.

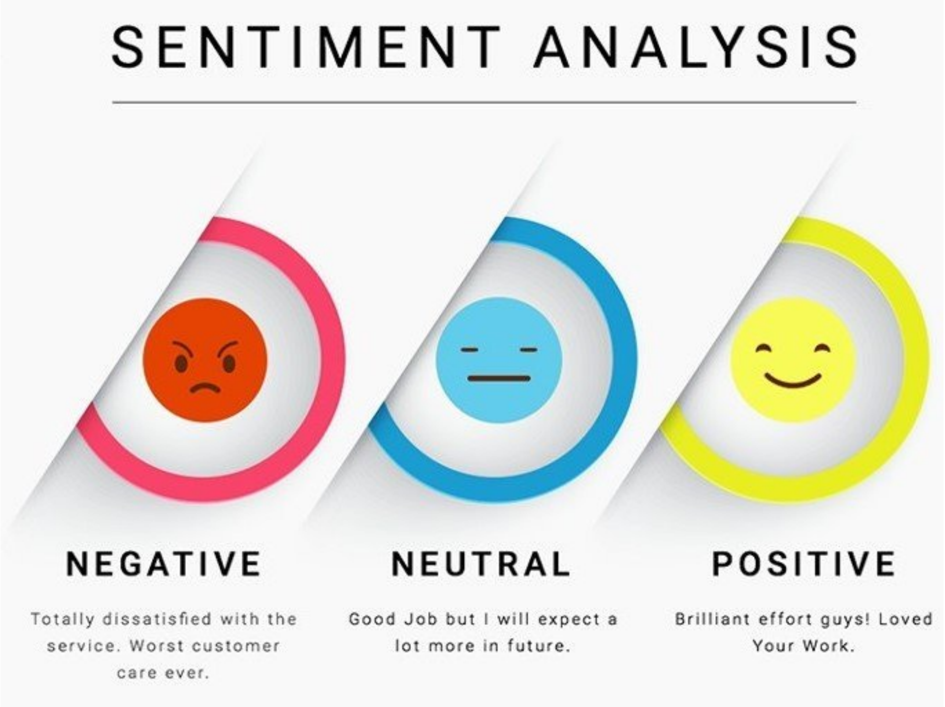

SENTIMENT ANALYSIS
NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.

Figure 1: Basic Sentiment Analysis Classes

2 Challenges

In this project, we face multiple challenges such as preprocessing the data, since we are using a dataset collected from Twitter social media platform as we will elaborate later in section 3, data samples are spontaneous and random sentences with lots of slang words and emojis. This sort of unstructured data will need much preprocessing to be useful. The second challenge is finding a suitable classification base model that we can modify and enhance its accuracy. The third challenge is dealing with Modern Standard Arabic (MSA) and the Egyptian dialect. These are the three main challenges among others.

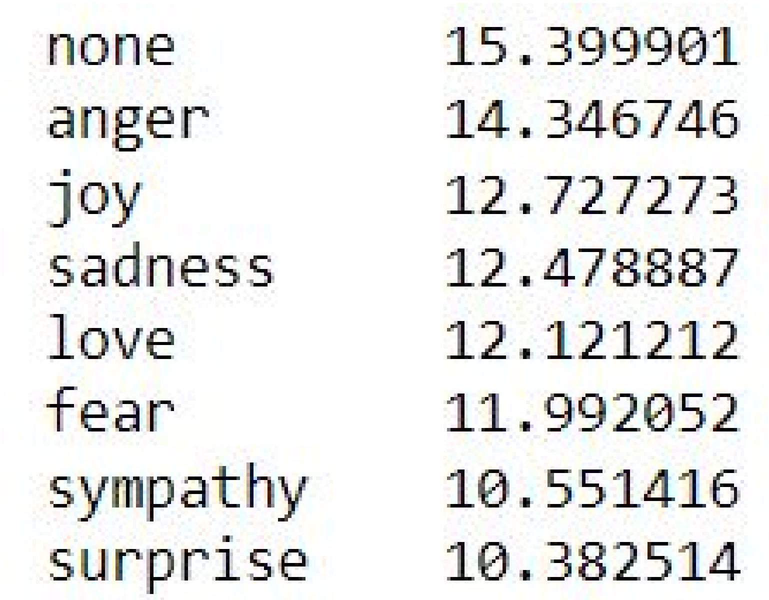
Our plan of how to solve these challenges is discussed in the following sections where we discuss the dataset used, the analysis of the data, and its preprocessing in section 3. We present an overview of our system architecture and some initial details about it in section 4.

3 Dataset

To find a suitable dataset, we searched for papers with contributions in arabic sentiment analysis. We found a paper titled "Emotional Tone Detection in Arabic Tweets"[1] we started using their dataset in order to use the results as a benchmark to evaluate our model. According to the paper, The dataset has 10,065 tweets and its goal was to cover the most frequently used emotions in Arabic tweets.It was collected from multiple resources : The first source was a corpus made up of 1167 tweets that had previously been gathered and classified as good, negative, or neutral by the Nile University (NU) text mining research group[2, 3]. A team of graduate students at NU then re-annotated this corpus to include emotional content. The students employed Ekman's emotional model and its six emotional classes—happiness/joy, sadness, anger, disgust, fear, and surprise to annotate this dataset. The second resource was made up of 2807 tweets that the same students had collected using Twitter's search API and annotated with the same set of feelings. Egypt's geolocation was used to filter the collected tweets between July 31, 2016, and August 20, 2016. Since many were watching the Olympics at the time, the word أولمبيادْ was used as a search keyword to download the tweets. The third resource was a dataset that was compiled following a search of NileULex sentiment lexicon words added to the Twitter API [4]. More than 500,000 tweets were collected as a consequence of the search. A randomly chosen portion of them was used for labelling. For a non-skewed dataset , A category oriented search was conducted for the under-represented categories (sadness, surprise, love, sympathy, and fear). Towards this end, the Twitter API was queried using hashtags that were expected to return tweets with the desired emotions.

3.1 Data Analysis

The Data Analysis process was mainly about knowing more about the data and making sure that the dataset was downloaded correctly and that no tweets were corrupted. We made sure by exploring some of the tweets appearing at the beginning of the dataset and others appearing at the very end using df.head() and df.tail() as shown in Fig[]. Moreover, as a way of being more aware of the data, a count plot was drawn in order to help maintain a visual representation of the number of tweets falling under every class of emotions represented in (sadness, surprise, love, sympathy, and fear). The count plot was drawn using the Seaborn python library and the output was as shown in Fig [6]. A final representation was conducted in order to make it easier to compare the number of tweets in every class which shows the percentage of every class as follows.



```
none         15.399901
anger        14.346746
joy          12.727273
sadness      12.478887
love         12.121212
fear         11.992052
sympathy     10.551416
surprise     10.382514
```

Figure 2: Percentage of every category

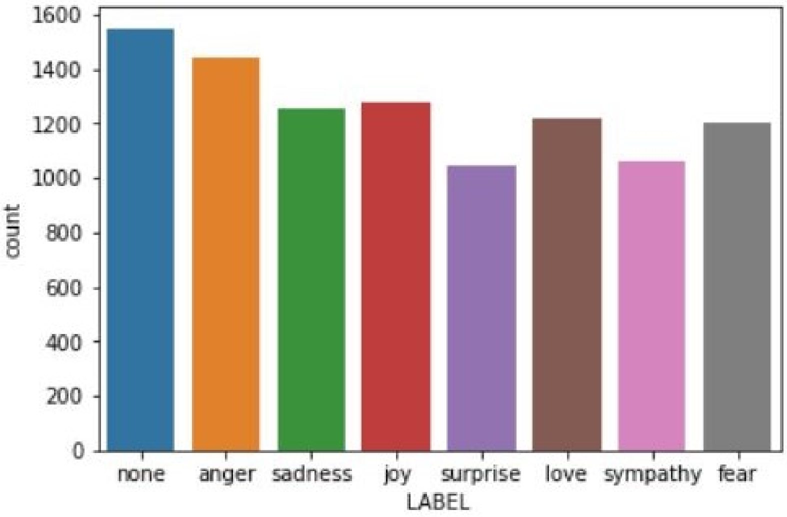
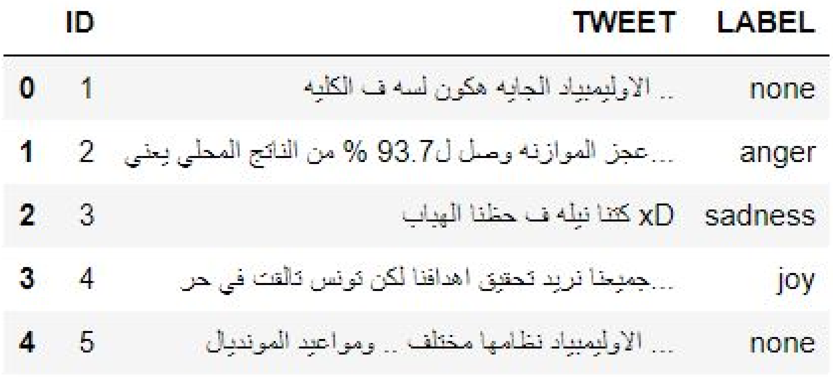


Figure 3: Count plot of categories

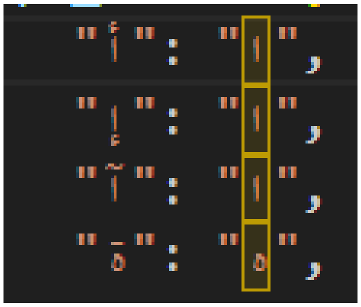


	ID	TWEET	LABEL
0	1	.. الاوليمبياد الجايه هكون لسه ف الكليه	none
1	2	...عجز الموازنه وصل ل93.7 % من الناتج المحلي يعني	anger
2	3	كنا نيله ف حظنا الهباب xD	sadness
3	4	...جميعنا نريد تحقيق اهدافنا لكن تونس تالقت في حر	joy
4	5	... الاوليمبياد نظامها مختلف .. ومواعيد المونديال	none

Figure 4: showing a very small sample of the data.

3.2 Data Preprocessing

We started preprocessing the data by tokenizing the data set using the NLTK library and wordpunct tokenizer. We then downloaded the arabic stop words from NLTK corpus and filtered the dataset. The third step was to represent similar representations of the same letter only in one way as in Fig [5] to improve the model's accuracy. We then removed the diacritics(اْلتشكيل) and all the elongated characters, for example : (اْلقااااهرةْ) using the regular expression (re) library. We also removed any punctuation such as "?:!.,;". At the end, we lemmatized the data using qalsadi lemmatizer.



```
"أ" : "ا",
"إ" : "ا",
"آ" : "ا",
"ة" : "ه",
```

Figure 5: Example of the replaced letters

4 System Architecture

The main system architecture components to deal with the aforementioned challenges, we first start by preprocessing the data which is done in a number of steps: First of all, we tokenize and segment the data using the NLTK library and wordpunct tokenizer, then we remove stop words, unify letter representations, and remove diacritics. Afterwards, we remove elongated characters and punctuation. The final data preprocessing stage is when we perform lemmatization to return all words to their original root. After the data is preprocessed, we start with the feature extraction phase using for example bag of words or tf-idf. Following the feature extraction is the data splitting into training and testing portions to prevent data leakage. Finally, we select a classification model, train it and test it against the testing data. We then evaluate it by calculating the confusion matrix among other evaluation metrics. The whole system architecture is shown below in figure 4.

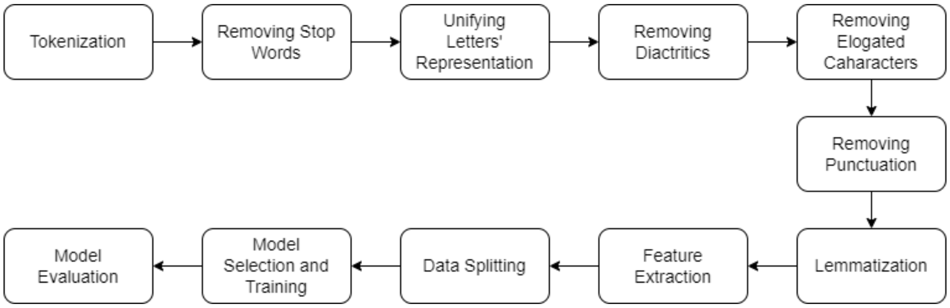


Figure 6: Flow Chart of System Architecture

These system components are subject to change according to their effect on the accuracy of the classification model. We will investigate the effect of each preprocessing step on the model performance and choose the best combination of components that result in the best results. We also might consider removing emojis or replacing them with a standard emoji for each class/emotion and testing the effect of this step on the model evaluation results.

References

[1] Amr Al-Khatib and Samhaa El-Beltagy. Emotional tone detection in arabic tweets. 04 2017.

[2] Talaat Khalil, Amal Halaby, Muhammad Hammad, and Samhaa El-Beltagy. Which configuration works best? an experimental study on supervised arabic twitter sentiment analysis. 04 2015.

[3] Samhaa El-Beltagy, Talaat Khalil, Amal Halaby, and Muhammad Hammad. *Combining Lexical Features and a Supervised Learning Approach for Arabic Sentiment Analysis*, pages 307–319. 01 2018.

[4] Samhaa El-Beltagy. Nileulex: A phrase and word level sentiment lexicon for egyptian and modern standard arabic. 05 2016.