Arabic Tweets Emotion Recognition

Supervised by: Dr.Mervat Aboelkheir
ENG.Mayar Osama
Authors:
Mariam Hesham
Mohamed Hesham
Mostafa Mohamed

May 8, 2023

Abstract

Using a publicly accessible dataset of labelled tweets, this research focuses on sentiment analysis of Arabic tweets. Understanding the sentiment of Arabic tweets has become a crucial job as the use of social media platforms as a forum for individuals to express their thoughts and feelings grows. Researchers and organisations can learn more about the thoughts and attitudes of people towards various events and issues by analysing the sentiment of tweets in real-time. The complexity and diversity of the Arabic language present difficulties for sentiment analysis of Arabic tweets. The goal of this project is to create efficient ways for analysing the sentiment of Arabic tweets using NLP and machine learning techniques.The performance of various algorithms will be evaluated and compared, with the goal of achieving accurate and reliable sentiment analysis of Arabic tweets.

1 Introduction

Sentiment analysis is a field within natural language processing that uses computational techniques to extract subjective information from text and determine the overall emotional tone, which can be positive, negative, or neutral. This technique has various applications in areas such as marketing, politics, finance, and customer service, among others. Social media platforms, particularly Twitter, have become a popular source of data for sentiment analysis due to their real-time nature and the large volume of data they generate. Arabic is a complex language with nuanced features, making sentiment analysis of Arabic text a challenging task. For this project, a dataset of labeled Arabic tweets was used to train and evaluate machine learning models for Arabic sentiment analysis.

2 Motivation

The motivation behind this project is to develop an accurate and effective model for sentiment analysis of Arabic tweets. With the growing popularity of social media platforms in the Arab world, understanding the sentiments and opinions of users on these platforms has become increasingly important for businesses, organizations, and governments. By accurately identifying the emotions conveyed by Arabic tweets, we can gain valuable insights into the attitudes and preferences of the Arabic-speaking population, which can be used to inform decision-making processes and improve communication strategies.

3 Dataset

The dataset used in this project is a publicly available collection of Arabic tweets that have been manually labeled for sentiment. There are a total of over 2600 tweets in the dataset, with 2060 tweets used for training and 688 tweets used for testing. Tweets are collected from Twitter, covering a diverse range of topics and events. The tweets are labeled based on the overall sentiment conveyed by the text, as perceived by human annotators.

3.1 Data Preprocessing

Preprocessing is a crucial step in any natural language processing project, and it involves cleaning and transforming the raw text data into a format that can be easily processed by machine learning models. In this project, we performed several preprocessing steps on the Arabic tweets dataset to prepare it for sentiment analysis.

3.1.1 Prepossessing steps

1. Tokenization: Tokenization is the process of splitting a sentence or text into individual words or tokens. We used the Arabic word tokenizer provided by the NLTK library to tokenize the tweets into words.

2. Stopwords removal: Stopwords are commonly used words in a language that do not carry any significant meaning. We used the Arabic stopword list provided by the NLTK library to remove stopwords from the tokenized tweets. Removing stopwords can help in reducing the size of the dataset and remove noise from the text data.

3. Removal of special characters and symbols: Special characters and symbols like ”@”, ”*”, and ”$” are commonly used in social media and can add noise to the text data. We used regular expressions to remove these characters from the tweets.

4. Removing mentions and URLs: mentions and URLs are widely used in social media platforms specially on twitter and are in English not in Arabic so also they were removed [Al-Khatib and El-Beltagy, 2017]

5. Removing emojis: Emojis are commonly used in social media to convey emotions and sentiments. However, they can add noise to the text data and make it harder to analyze. We used regular expressions to remove emojis from the tweets.

6. Removing punctuation: Punctuation marks like commas, periods can add noise to the text data. We used regular expressions to remove punctuation marks from the tweets.

3.2 Data Analysis

The goal of data analysis is to understand the underlying patterns and characteristics of the dataset [Rabie and Sturm, 2014], which can help in identifying potential biases, selecting appropriate machine learning algorithms, and improving the accuracy of the model. Visualization is an effective way to explore and analyze data, and it can help in identifying patterns and trends that may not be immediately apparent from the raw data.

3.2.1 Data Analysis steps

The first approach was to check the size of the total words in the dataset before and after the preprocessing steps to compare how cleaning the dataset affected its size. Table 1 shows the results obtained.

Size before	Size after
32597	16937

Table 1: Words count.

The second approach was to examine the distribution of the dataset across the three sentiment categories, which are positive, negative, and neutral. This was achieved to assess any potential bias towards any particular sentiment class. Figure 1 shows the results obtained.

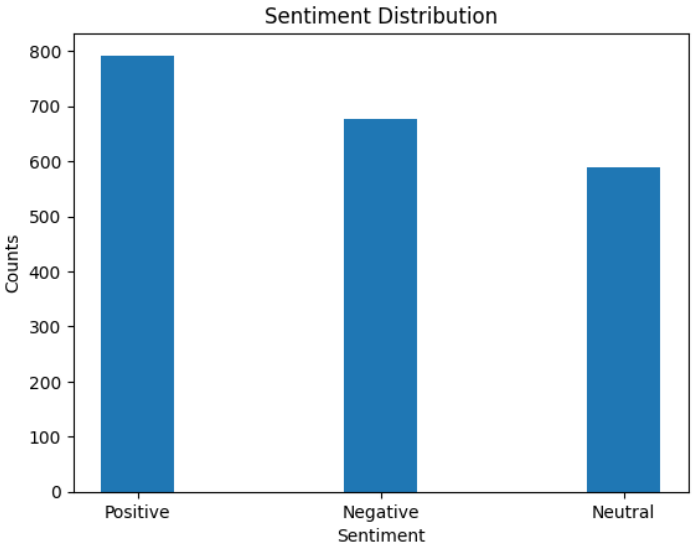

Figure 1: This is the result of the count of words in the three sentiment classes.

The third approach aimed to investigate the tweet length distribution within the dataset to extract insights on the most frequent tweet lengths. This analysis is essential to gain a better understanding of the characteristics of the dataset and to identify potential limitations or biases in the data. To visualize the tweet length distribution, bar charts were utilized to plot the frequency of tweets for different length categories. Figure 2 shows the output of the length distribution.

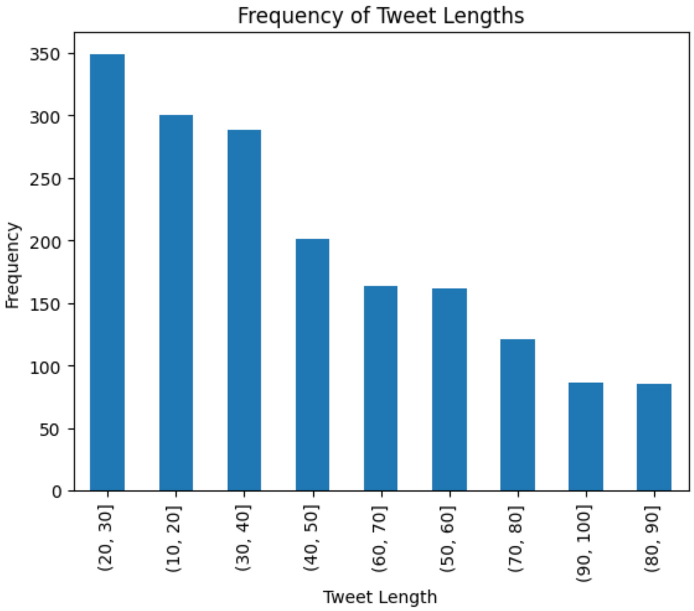

Figure 2: This is the result of the length distribution of tweets.

The fourth approach aimed to identify the most commonly used hashtags in the dataset, which could provide valuable insights into the main topics discussed in the tweets and thus the overall sentiment. This analysis was performed using bar charts to visually represent the frequency of hashtags and help identify the most frequently used ones. Figure 3 shows the output of the comparison.

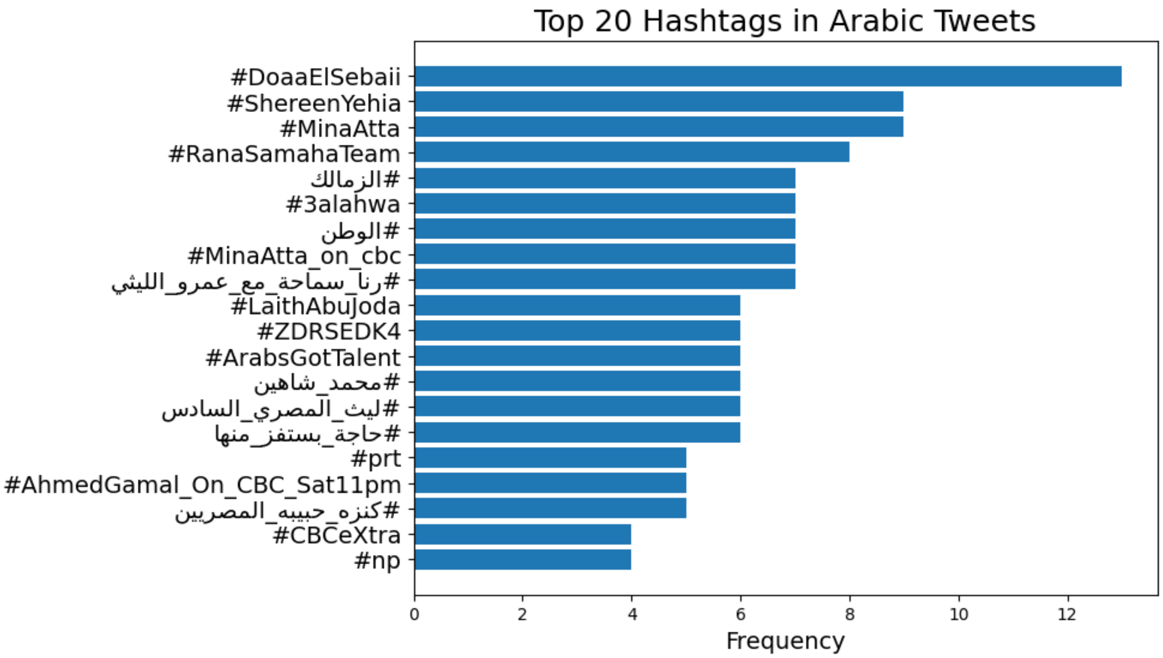

Figure 3: This is the result of the most common hashtags used.

The final approach was to identify the most common 10 words used, this can provide insights into the topics or themes present in the dataset. By analyzing the most frequent words, we can gain a better understanding of the topics that people are talking about and the language that they use to express their sentiments. This information can be valuable for designing effective marketing campaigns, understanding public opinion on specific issues, and improving customer engagement. We visualized this bar chart. Figure 4 shows the output of the most common 10 words used.

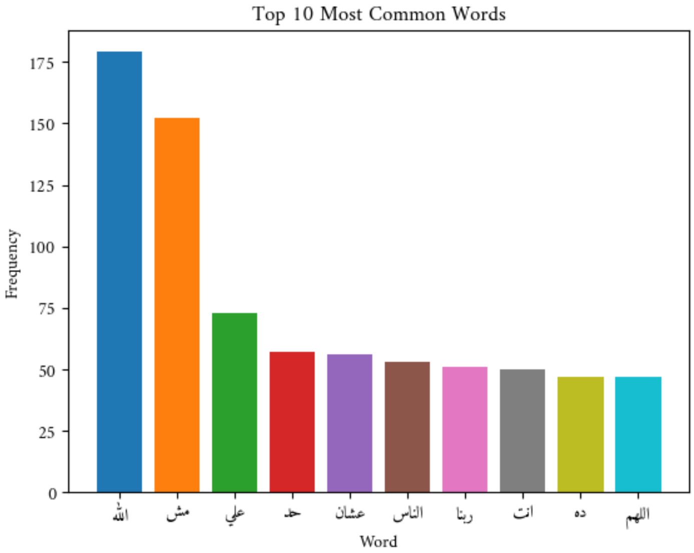

Figure 4: This is the result of the most common 10 words used.

4 Limitations

Limited dataset size: The dataset used in this project consists of over 2600 Arabic tweets, which may not be representative of the diversity of opinions and attitudes in the Arab world. A larger dataset could provide more reliable results and reduce the risk of overfitting.

Limited domain specificity: The dataset used in this project is not domain-specific, meaning it covers a wide range of topics and contexts. This could limit the accuracy of the sentiment analysis results for specific domains, such as politics, entertainment, or sports.

Limited language resources: Arabic language resources, such as sentiment lexicons and labeled datasets, are not as extensive as those available for English, which could limit the accuracy of the sentiment analysis results.

imbalanced datasets in sentiment analysis can be solved by Undersampling: randomly remove samples from the majority class to balance the dataset. Oversampling: replicate samples from the minority class to balance the dataset. Class weighting: assign higher weight to the minority class during model training to compensate for its smaller size.

5 System Architecture

The system architecture for Emotion Recognition in Arabic Tweets encompasses components for data preprocessing, feature identification, and the application of a machine learning model. The preprocessing stage refines the data through processes like tokenization and normalization. Relevant text attributes are then extracted using techniques like word embeddings or TF-IDF vectors during the feature extraction phase. Subsequently, a machine learning model as BERT is trained and fine-tuned on a labeled dataset. The evaluation of the model is based on metrics like accuracy and F1-score, facilitating real-time emotion detection in Arabic tweets.

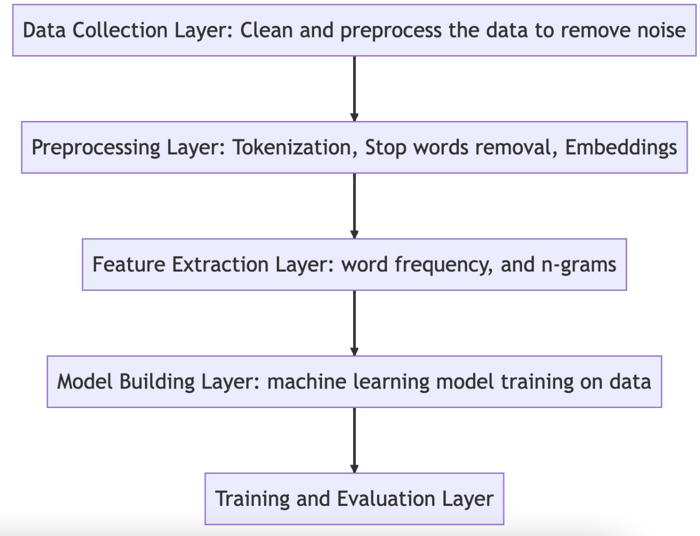
Data Collection Layer: Clean and preprocess the data to remove noise
Preprocessing Layer: Tokenization, Stop words removal, Embeddings
Feature Extraction Layer: word frequency, and n-grams
Model Building Layer: machine learning model training on data
Training and Evaluation Layer

Figure 5: System architecture

References

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

Omneya Rabie and Christian Sturm. Feel the heat: Emotion detection in arabic social media content. In *Industrial Conference on Data Mining*, 2014.