Sentiment Analysis For Arabic Tweets

Ahmed Lokma , Ahmed Mohamed , Seif Ramy

May 7, 2023

Abstract

Sentiment Analysis for Arabic Tweets has become a crucial tool for analyzing public opinion towards various issues, thanks to the widespread use of social media platforms like Twitter in the Arab world. However, analyzing Arabic text data poses unique challenges due to the complexity of the language and culture. As a result, sentiment analysis of Arabic tweets requires specialized tools and techniques to address these challenges. This report provides a comprehensive overview of Sentiment Analysis for Arabic Tweets, including its motivation, data analysis, and preprocessing techniques, as well as the challenges that come with this process. The report also proposes an architecture for sentiment analysis of Arabic tweets to improve the accuracy of the analysis

1 Introduction

Sentiment analysis is a technique that uses machine learning and natural language processing to identify whether a text has a positive, negative, or neutral tone, as well as other relevant information such as emotions and attitudes expressed in the text. This technique is useful in fields such as marketing, customer service, political analysis, and social research. It has become even more important with the increasing amount of digital content and social media use.

Social media platforms have become popular sources of data for sentiment analysis, since they allow people to express their opinions and emotions publicly and in real-time. Twitter is one of the most popular social media platforms globally, with millions of users tweeting daily about a range of topics. In the Arab world, Twitter has become a popular platform for discussing politics, social issues, and entertainment, among other topics.

The Arabic language is the fifth most widely spoken language globally, with over 400 million speakers. However, analyzing Arabic text data poses unique challenges due to its complexity, including the use of non-standard characters, diacritics, and variations in grammar and syntax. Therefore, sentiment analysis of Arabic tweets requires specialized tools and techniques to account for these challenges.

2 Dataset

The utilized dataset in this study is the "Arabic Sentiment Twitter Corpus", which comprises a collection of tweets conveying positive and negative sentiments extracted from the Twitter platform. This dataset was procured in April of 2019 and consists of 58,000 Arabic tweets, with 47,000 tweets reserved for training and the remaining 11,000 for testing purposes. Each tweet has been annotated with either a positive or negative label. The dataset is balanced and has been collected using a comprehensive lexicon of positive and negative emojis.

3 Data preprocessing

3.1 Data Cleaning

Data cleaning is considered as an essential stage for data analysis, due to it eliminates unnecessary data with the aim of achieving better results in sentiment analysis. Our data cleaning process involved the following steps:[AR23]

- Remove mentions and URLs.
- Remove any non-Arabic characters.
- Remove Punctuation.
- Remove Stopwords.
- Remove elongation.
- Remove emojis.
- Remove retweets.
- Remove Arabic diacritics marks.
- Remove extra whitespace.
- Tokenization.

Diacritic Marks	Characters
Fatha	◌َ
Tashdeed	◌ّ
Tanwin Fath	◌ً
Damma	◌ُ
Tanwin Damm	◌ٌ
Kasra	◌ِ
Tanwin Kasr	◌ٍ
Sukun	◌ْ

Figure 1: Arabic Diacritics marks[AR23]

3.2 Normalization

Normalization refers to the process of reducing letters to their most basic form. Given the rich morphological complexity of the Arabic language, normalization is essential to simplify the language and make it more manageable for analysis and processing. Normalization was also done in data preprocessing as it's an important step.[AR23]

Letter	Normalized Form
اٱأإآ	ا
ى	ي
ئ	ء
ؤ	ء
ة	ه
ک	ك

Figure 2: Some letters in normalization form [AR23]

3.3 Stemming

Stemming in Arabic NLP (Natural Language Processing) refers to the process of reducing Arabic words to their root form, also known as the stem. This involves removing any affixes or prefixes that modify the word's meaning and converting the word to its basic form. On the other hand, there's some cases where some Arabic words lose their meaning after stemming operation. So we tried to handle this by making a list of words that should not be stemmed.

4 Data Analysis

Data analysis in NLP (Natural Language Processing) involves the application of statistical, machine learning, and computational techniques to extract insights and patterns from natural language data. The goal of NLP data analysis is to uncover meaningful information from large volumes of text data. In our project, we conducted two data analysis for our data:

1. Most common words in tweets
2. Tweet lengths and their frequencies

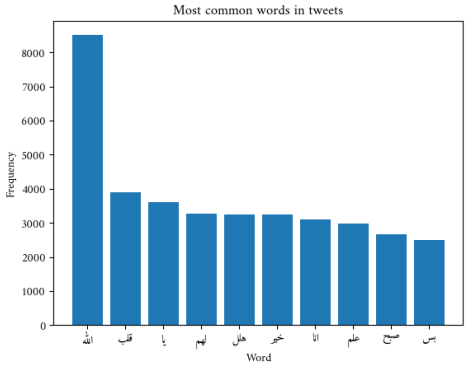

Figure 3: Most common words in tweets

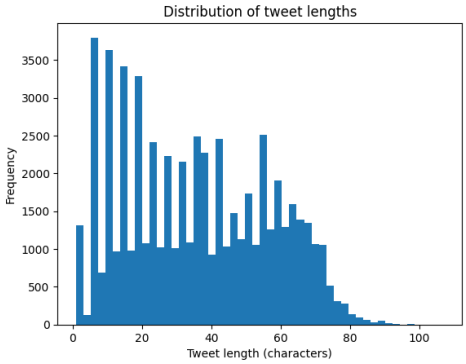

Figure 4: Tweet length and their frequencies

5 Challenges

During the Sentiment Analysis For Arabic Tweets project, several challenges were encountered that need to be addressed to ensure accurate and reliable results. One of these challenges is the limited support for all Arabic words in Matplotlib, which can affect data visualization as Matplotlib default font does not contain all arabic words, so it outputs an empty box ([]) . Another challenge, that some words cannot be stemmed in Arabic language, so we made a list of Arabic words that cannot be stemmed, in order to avoid stemming these words.

6 System Architecture

System architecture is a critical component of any software development project, as it defines the overall structure and organization of the system.Our system architecture is divided into several steps:

1. Data Collection
2. Cleaning dataset and dataset preparation
3. Feature Extraction
4. Sentiment Analysis

Feature extraction is an important step in any NLP pipeline, as it helps to transform raw text data into a set of numerical features that can be easily processed by machine learning algorithms. Feature extraction could be done using several methods such as bag of words, tf-idf, word-embeddings or Arabert. Sentiment analysis could be done using machine learning model such as svm , naive bayes or random forest.

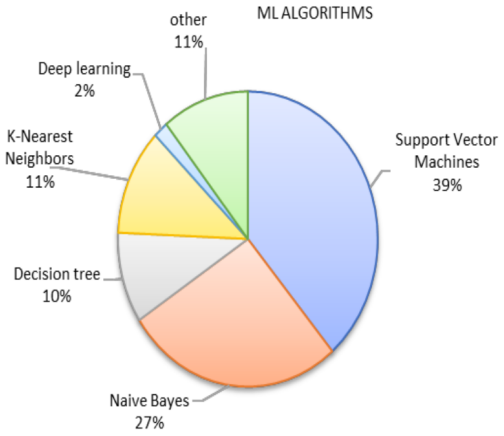

Figure 5: Machine learning models used in sentiment analysis [TAN17]

The system architecture is illustrated in the figure below:

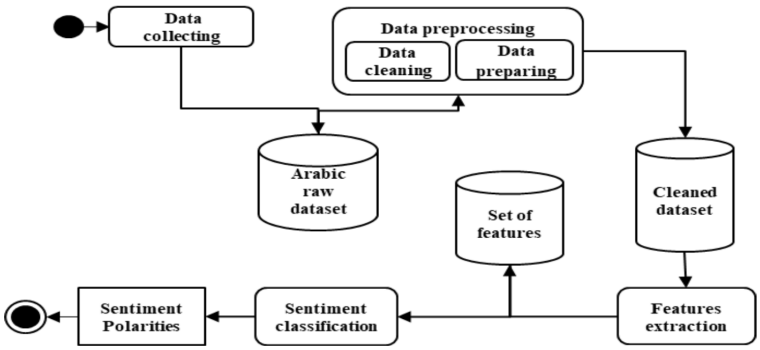

Figure 6: System Architecture [TAN17]

References

[AR23] Arwa Alqarni and Atta Rahman. Arabic tweets-based sentiment analysis to investigate the impact of covid-19 in ksa: A deep learning approach. *Big Data and Cognitive Computing*, 7(1):16, 2023.

[TAN17] Samir Tartir and Ibrahim Abdul-Nabi. Semantic sentiment analysis in arabic social media. *Journal of King Saud University-Computer and Information Sciences*, 29(2):229–233, 2017.