Arabic Text Reviews Sentiment Analysis

Youssef Agiza*
youssefagiza@aucegypt.edu
The American University in Cairo

John El Gallab
john.elgallab@aucegypt.edu
The American University in Cairo

1 Dataset: Arabic 100K Reviews

After researching multiple datasets, we have decided to go with the Arabic 100K Reviews comprehnsive data set. Despite being spoken by over 320 millions humans around the world, sentiment analysis of Arabic texts is an under represented area in this domain; which is a main reason that made us go with an Arabic dataset.

This dataset combines a compilation of reviews from hotels, books, movies, online products and airlines. All data points were labelled positive, negative, or mixed according to their sentiment. Generally, all reviews that had more than 3 stars out of 5 were considered positive, all reviews below 3 stars were considered negative, and those having 3 stars exactly as mixed. The data set was already cleaned from non-Arabic characters before we start our own preprocessing on it. The data set consisted of 2 columns exactly: the label (positive, negative, or mixed) and the review text itself. No further features were expressed in the data set.

The dataset is balanced as it contains 33,333 unique positive labels, 33,333 unique mixed labels, and 33,333 unique negative labels. All data is written in string format within a tab-seperated file.

2 Preprocessing pipeline

Similar to English textual data preprocessing, the Arabic textual data needs a pipelined preprocessing mechanism to clean it and generate meaningful features of it that can be later used in training and validating the machine learning model. The order of the pipelining steps is described here and each step of the pipeline is explained in its subsection below. We start by replacing the string labels with numerical labels followed by a language detection phase to make sure that these Arabic characters form Arabic words, if any non-Arabic text was found it was pruned from the dataset. This was followed stemming or lemmatization of all the textual reviews to ensure that every word in them is written in a common form (even if not linguistically accurate) but this is to make that words that had the same meaning were actually associated together even if written in different grammatical format. The text was then dedicritized to remove all figuration (تشكيل و حركات) of the Arabic words. We also normalized certain Arabic characters to a common form (ا ، إ ، ئ) were normalized into (ء) and all (ي ، ى) were normalized into (ى). We removed the word elongation known as (تطويل) followed by removing all the stop words. Then, we concluded with frequency encoding our data.

2.1 Label encoding

We replaced the string label positive with a 1, mixed with 0, and negative with -1 using the string.replace function available.

2.2 Language Detection

The three most popular python libraries available for language detection are textblob, langrid, and langdetect. Both textblob and langrid are good options to detect the Modern Standard Arabic language accurately, but we opted to use the langdetect library as it is a port of Google's language detection library and is expected to have much higher exposure to the colloquial Arabic of different dialect (عامية) which is a form of data that is present within our dataset.

2.3 Stemming vs. Lemmatization

Both stemming and lemmatization are closely related. They both aim to remove inflectional forms of words and reduced derived word forms into a common base form which is basically attempting to reduce a set of words in the dataset into their canonical representative. For Arabic language, very few lemmatizers exist but more stemmers exists (although they are in a scarce amount when compared to English/Latin-based stemmer).

Stemming. Commonly used stemmers are the ISRI Arabic Stemmer, Assem's Arabic Light Stemmer, and the Stanford NLP Stemmer. We settled for using ISRI Arabic stemmer for

*Both authors contributed equally to this research.

this phase of the project due to its relatively higher accuracy and because of delayes in obtaining licensing to more accurate lemmatizers that will be described in the next subsection.

Lemmatization. The current state-of-the-art lemmatizers are all unfortunately not available readily for free usage. Farasa, Madamira, and CAMel all require licensing approvals as per their websites. The Farasa lemmatizer is the most accurate of the previous three but unfortunetly we only managed to get access to its web API which limited our access to 10 operations per minute which was infeasible in terms of 99,999 textuals strings. The Columbia Technology Ventures reponsible for the non-commercial licensing of Madamira provided us with a free license but before the submission deadline by a couple of hours; therefore, we were not able to utilize it in this phase despite of its higher accuracy (about 96 percent)

2.4 Dedecritization

Using PyArabic library, we removed all all figuration. Also, all (ئ ، إ ، ا ، أ) were normalized into (ء) and all (ى ، ي) were normalized into (ى). We removed the word elongation known as (تطـويـل). This was made possible using strip-tashkeel(text), strip-harakat(text), and normalize-hamza(text) functions that are made available in pyarabic.araby.

2.5 Step Words Removal

Since we couldn't find a free Arabic stop words dictionary, we manually defined one from the top 10 percent repeated words in the dataset such as (من ، على ، في ، ان).The full list of selected stop words is available in the notebook attached to this paper. Also, all words that appeared less than 100 times were removed to reduce noise present in the dataset

2.6 Duplicate Removal

After all the preprocessing, we tried removing any duplicate reviews in our dataset but not a single one appeared as a result of the preprocessing so we decided to exclude this step from the preprocessing pipeline as it seems redundant for this dataset.

2.7 Frequency Encoding for Feature Generation

For each sentence, we kept track of meaningful words in this sentence and then counted how many times it appeared within that single sentence in a dictionary. All dictionaries were placed in the frequency column with their respective sentence.

3 Post-Cleaning

Post cleaning, we are left with 99997 unique labelled reviews that are almost fully equally balanced with 33,332 positive entries, 33,332 mixed entries and 33,333 negative entries.