Designed report by Abdelrahman Rezk

COVID-19-Arabic-Tweets-Dataset

We have collected more than 3, 000, 000 tweets from twitter API, besides cleaning these tweets for our future work on Analysis.

- Eng: Ayman Mahgoub
- Researcher at electronics research institute
- E-mail: Ayman_mhgb@hotmail.com
- Eng: Abdelrahman Rezk
- Teaching Assistant at Arab Open University
- E-mail: Abdelrahmanrezk12011@gmail.com

Contents

Data Cleaning

- Remove diacritics
- Remove punctuation and Urls
- Normalize text
- Remove Stop words
- Other function based on analysis

Corresponding Code Files

Files included in direction:

- config_files

Files Name:

- Configs.py
- get_data_from_twitter_using_your_developer_account.py
- cleaning.py

----------Configs File----------

This file contains some of the main direction data URLs and the setup of your Twitter code to provide.

Data Collection

Clone The repo

We have used the research paper of Large Arabic Twitter Dataset on COVID-19 by Sarah Alqurashi, Ahmad Alhindi, Eisa Alanazi, which lead us to the repo https://github.com/SarahAlqurashi/COVID-19-Arabic-Tweets-Dataset that have the tweets ids.

----
get_data_from_twitter_using_your_developer_accountFile----------

Gather Data using twitter Api

After that we used the twitter developer account that we have to collect the tweets text from the tweet Id they provide and we have collected more than 3,000,000 tweets overall from about 4,000,000.

One file function

We have design code to enhance the returned tweets by breaking the process into first, get one file tweets, and for this file, we return most of the tweets it contains. This function takes one argument which is a file path we need to process and start to get from this file all Tweet ids that contain to get the text of the tweet associated with these ids, after that we save the text of the tweet we got in a separated file.

One direction function

Each direction has its own files that contain files with the year, month, and days they have been collected and each direction has about 30 files represented month days, so we have a loop over each direction for each of them we get each file and store in another path of the same name.

Data Cleaning

The first thing we started with after gathering the data is to analyze the tweets we got and at this point in what we dealing with is an Arabic tweets we start the cleaning process by:

Remove diacritics

The Arabic language has some of its own features that make it an amazing language, but some of these things need to be handled because it consumes other process and memory and maybe lead to missing actual information because each char is required memory and then each of them will require machine learning algorithms to learn about as it will transform to number and take place in our learning process so chars are called diacritics. We need to delete it from our text, and this will help us to normalize our text in one form because some of the words will have diacritic chars while the other not.

Char that we aimed to delete like [like: " ُ ّ "]

Remove punctuations and urls

One of the most important things we have is that most of the tweets have a lot of other chars that maybe have no meaning like [?,] besides of the tweets that have URL of some reference and for that, we have processed to be removed as it should reflect no knowledge in our analysis.

Normalize text

We should process the text we clean later on machine learning models and words like جميله و جميلة will represent different vector as they differ in last characters and this will affect our model to miss leading because of the same words have different representation.

Remove stop words

The stop words are words like [and, or, not and others] while in Arabic are [هذا و هذه], words that are mentioned in the text much more than other words and actually it does not add information to our model when we start to analysis the tweets but also there are different issues of these stop words because some libraries are deleted all of the stop words from these words [not] while in Arabic [لا] and this is just one word that can be deleted if we use other libraries code and because of that, we have initialized our stop word file that can be used for like this analysis which contains more than 500 words that can be deleted without negative effect like word [لا] that if deleted will change the whole meaning of the tweet from Negative tweet to Positive tweets if we need to classify tweet and for this, we have created our stop words file.

Other functions

All of these previous functions and other like Lemmation can be used but all of that based on the analysis that we will go through, So we have designed our code to work in the general case and also will try different methods with Models that will be trained on our data, besides of another End-To-End approach because some of the things that we removed reflect a lot of meaning in its context.