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 and we have make some analysis to get what is behind these tweets.

- Eng: Ayman Mahgoub
- Researcher at electronics research institute
- E-mail: <a href = "mailto:Ayman_mhgb@hotmail.com"> Ayman Mhgb </a>

- Eng: Abdelrahman Rezk
- Teaching Assistant at Arab Open University & NLP Engineer
- E-mail: <a href = "Abdelrahmanrezk12011@gmail.com"> Abdelrahman Rezk </a>

Corresponding Code Files

Files included in direction:

- config_files

Files Name:

- direction_and_file_handleing.py
- get_tweet_to_classify.py

----------------------------direction_and_file_handleing File-----------------------------

Direction And File Handleing

In this work, we have made some of the helpful functions that help us through the work on different notebooks, because of the direction that contains multiple files and each of them contain a number of tweets, we have made this work to call it in each other notebook and use it as it requires instead of repeat the chunks of the code in each notebook.

We have impletent just one function in this Report and others we have used from previous work:

- one_file_preprocess
- read_direction_preprocess
- handle_direction_analysis
- one_file_analysis
- read_direction_analysis

one_file_preprocess

The function used to read the CSV file and convert the read tweets from this file to the list. After that the function path the list of tweets to the Cleaning Arabic pipeline, which we talked about in the report Gather & Clean in report direction.

The function takes an argument:

- Inputfile: The file path that you need to read and make a process on.

read_direction_preprocess

The function here use the function above which is one_file_preprocess, as we mentioned in other reports we have multiple dierctions[filders], each of them contain some of files, which contain some of tweets.

For each file we read it and make some cleaning on this file using the function above **one_file_preprocess**, after this analysis we aimed to save the newly cleaned data of this file into another dierction[filder] which folder called **preprocessed**, but with the same name of the original file we read from.

Loop over each direction and path each file in this direction to get all tweets then:

- save the new preprocessed data which cleaned tweets we have make another dierction, then
- to save the data in for each file create the same file name and save in another dierction.

The function takes an argument:

- direction_path: The direction you aimed to read files from.

handle_direction_analysis

We have multiple of direction and each of them have its own path so instead of access each of them on its own, we just make a path of one of them and others direction we need to read, we make a simple function that handles the path of the new direction we need to read.

This is work because all directions have the same name but different in the last 1 or 2 char of the name.

For Example:

Use one default direction and replace this with others like below:

- "COVID-19-Arabic-Tweets-Dataset/COVID19-tweetID-2020-01/" - to
- "COVID-19-Arabic-Tweets-Dataset/COVID19-tweetID-2020-02/"

The function takes arguments:

- direction: the folder path you need to change.
- replace_with: replace the folder name.

one_file_analysis

The same as one_file_preprocess but just read and convert to list no cleaning functions are called.

read_direction_analysis

The function used in two different way based on the default argument **convert_list**, and it uses the function above which is **one_file_analysis** to read the file, which works as convert the tweets in this file into a list and return this list, after that the function **read_direction_analysis** either

- to return all tweets of all files which are in one direction into one list that contain all these tweets,
- or return all words in all tweets which are in all files of one direction.

Loop over each direction and path each file in this direction, then get the tweets of one path as a list, then get all of the words of this list, and extend this list to contain all words in all files of one direction.

The function takes arguments:

- direction_path: The dierction you aimed to read files from.
- convert_list: default argument used when it's required.

--------------------------------get_tweet_to_classify File-----------------------------

Classification

After what we have made of the gather, cleaning and analysis these tweets we have separated some random tweets for directions 1 and 2, which will be used for manual labels to generalize this for the problem of classification these tweets into 2 classes [0, 1].

- Zero class for tweets that it not talk about Coronavirus.
- One class tweets that talks about Coronavirus.

The classification problem will dealing with via multiple Machine Learning Algorithms like Logisitc regression, but at the same time we will trying to measure how will our model is in the way of let these tweets clusters using the unsupervised learning methods.

We have impletent just one function in this Report and others we have used from previous work:

- separate_number_of_tweet_each_file
- get_random_number_of_tweets

separate_number_of_tweet_each_file

As we know that each file have multiple of tweets and each file belongs to one of the direction(folder), and we have 4 direction for 4-month start from January to April but what we used as sample is just from the first 2 month or two direction.

As we know we can not label all of the tweets, which in each direction more than 100,000 tweets, so based on the Machine learning Supervised and Unsupervised Approach we have map the problem from label just 3000 sample of each direction to train the model on using the Approach of Supervised learning, and at the same time we will use the Approach of Unsupervised tl cluster these tweets into 2 cluster then we will comapre the two approach to each other.

So the function takes number_of_tweets as a parameter which is related to a number of files in each direction, and we think of a simple math equation to know if the direction contains all of the days then it will have 30 days so will divide 3000/30 will give us that we need 100 tweets from each file to make the sample from all files which define that our sample is represent all tweets from all files, otherwise the 3000 will divide by the number of files in this direction to know which number of tweets we need from each file.

get_tweet_to_classify

After we have been finishing the function separate_number_of_tweet_each_file, and check the result CSV file we that the data do not represent a good sample of the data so we have to make another function and enhance the random way we get the tweet from to arrive after classification to be in balance of the two classes.

Function to read all tweets of all files in one direction into one list, then

- take random tweets from this list, also for each tweet make some regular expression,
  - like replace \n with just space to make the whole tweets in one line,
- then save the result in CSV and XLS file.

The function takes:

- direction_read: Which direction you will read files from.
- direction_save: After you take the random tweets and make a data frame of it which direction you need to save.
- number_of_tweets: which number of tweets at all you need.