Computer Engineering Department
Faculty of Engineering
Cairo University

CMP Elective 4
Fall 2022

Natural Language Processing
Project Document

Project Description

We all witnessed the COVID-19 pandemic and realized how such a global pandemic has changed many perspectives in our daily lives. One of the most important ways to determine how individuals are affected by such a pandemic is by collecting what they are saying about it. Social media platforms such as Twitter represent a very rich source of people’s opinions and feelings.

During the release of COVID-19 vaccines, people had different opinions about these different vaccines. Some people stand with these vaccines seeing that they are the way to get rid of the virus. Others believed that all these vaccines are useless or may be harmful. Also, these tweets contain different contents (e.g. rumors, advice, information authentication, etc.)

In this project, you will be provided with labeled Arabic tweets that are related to COVID-19 vaccinations. Your task is to build NLP models to detect the stance toward vaccination and the content type of the tweet. Stance detection is the task of detecting whether the author of a piece of text is in favor of the given target or against it.

Dataset Description

The dataset contains labeled Arabic tweets related to COVID-19 vaccines. You will be given the dataset in three different portions (train, dev, and test). Both the train and dev sets will be annotated for stance detection and tweet category classification. The test set contains only the tweet text and your task is to predict the tweet stance and category. The dataset portion sizes are as follows:

1. The training set contains 6988 samples.
2. The dev set contains 1000 samples.
3. The test set contains 2000 samples.

Stance detection labels meaning is as follows:

1. **Positive** (1): means that the tweet author encourages and supports vaccination.
2. **Negative** (-1): means that the tweet author refuses vaccination.
3. **Neutral** (0): means that the tweet neither supports nor refuses vaccination.

Category labels meaning is as follows:

1. **Info_News**: Information about vaccination.
2. **Celebrities**: mentioning celebrities taking vaccinations.
3. **Plan**: Governmental plan or progress of vaccination.
4. **Request**: Requests from governments regarding the vaccination process.
5. **Rumor**: the tweet is a rumor.
6. **Advice**: Advice related to the virus or the vaccination
7. **Restriction**: Restrictions due to the virus e.g. traveling.
8. **Personal**: Personal opinion or story about vaccination.
9. **Unrelated**: Unrelated to vaccination.
10. **Others**: Vaccination related but not one of the above.

Project Pipeline

Your task is to build a system that takes only the tweet text and produces both the stance and the category of the tweet. Here is the pipeline that you will follow:

1. **Data Preprocessing:**
    a. Data cleaning: This includes removing all unnecessary text from the tweets (e.g. URLs).
    b. Tokenization: Use the tokenizer that you think is more suitable for the data.
    c. Lemmatization: You can try if using lemmatizers will be useful for your case or not.
2. **Feature Extraction:** In this phase, you are required to try **at least three** different features (e.g. Bag of Words, TF-IDF, Word Embeddings, etc.) It will also be great if you tried other features either well-known features (e.g. contextual word embeddings) or handmade features (e.g. capturing specific words or patterns in tweets with regex).
3. **Model Building:** In this phase, you are required to build **at least four** different machine learning models. At least one of the four models must be a classical machine learning classifier (e.g. Naive Bayes, SVM). At least one of the four models must be a sequence model (e.g. RNN, LSTM, CRF). Optimizing the model weights will be done using the training set. You will need to use the dev set to pick the model that performs the best since you will only submit only one version of the test set labels.
4. **Model Testing:** In this phase, you will use your best-performing model to produce the stance and category of the given test set tweets.

Grading Criteria

This is a competitive project. The teams will be ranked based on the scores they get in the two tasks: stance detection and category classification. You will be provided with the test set tweets only **ONE DAY** before the final delivery. We will use the macro F1-score as the metric for the ranking process. We will use Kaggle for the ranking process. The overall grading will depend on the following:

1. The team rank in the two tasks
2. The approach you followed. This includes the following:
    a. The preprocessing techniques
    b. The features you used (at least three different features)
    c. The models you trained (at least four different models with at least one classical and one sequence model).
3. The workload division.

Project Schedule

- Project Document Release: Week 6 (Saturday 5th Nov. 2022)
- Final Delivery: Week 13.

Project Instructions

- You will work in teams of 4.
- Your final submission only will be considered for the ranking process.
- There is a penalty for late submissions.
- Any sign of cheating or plagiarism will not be tolerated and will be graded **ZERO** in the project.

Final Deliverables

1. **Final Project Document** containing the following:
    a. Project Pipeline
    b. A detailed description of each phase in your pipeline
        i. Data preprocessing
        ii. Feature extraction
        iii. Model training
    c. Evaluation: Report the macro F1-score (and all other metrics you tried) for all trials you did.
    d. Specify what model you used for the test set submission on Kaggle and the reason for choosing it.
2. **Codes**: All scripts you used.
3. **The final Model**: the weights of the model you used for submission. Use the framework you used for training the model default format when saving.
4. **Presentation**: you will use it for the final project discussion.