# Iranian Militias Classification

Artiom Berengard & Roey Haddad

## About

The purpose of this project is to develop a tool that determines whether a given tweet is anti-Iranian militias or not (categorized as irrelevant or pro-militias). The project consists of two main parts: the first involves collecting and classifying the data, while the second focuses on creating, training, and testing a model.

## Data Gathering

With the assistance of students from the Middle East department, we compiled a list of Twitter accounts categorized as anti-militias, pro-militias, and neutral. The list included a total of 162 accounts: 27 anti-militias, 13 pro-militias, and the remaining were neutral. Using the Twitter API via the Tweepy library, we collected a total of 10,000 tweets from these accounts, excluding retweets.

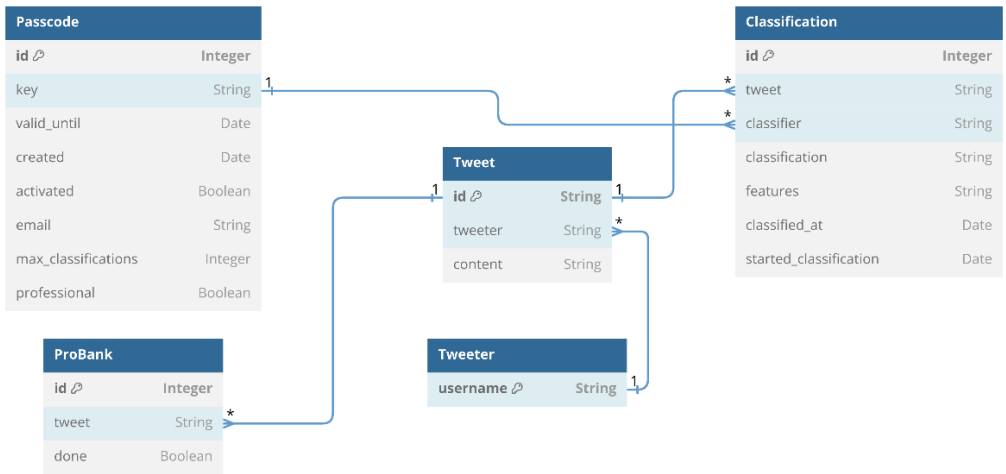
## Database

After collecting the tweets from all Twitter accounts, we preprocessed and inserted them into the database. The preprocessing steps for a tweet included:

a. Removing URLs, mentions, emojis, and extra spaces.
b. Keeping only English and Arabic letters, numbers, and symbols.
c. Deleting tweets with fewer than 6 words.

Following this preprocessing, a total of 3,855 tweets were inserted into the database.

The schema we used is described in the following chart:



## Website for Classifications

We deployed a web application to enable students from the Middle East department to classify the tweets. Both the front end and back end are hosted on Heroku, and the application is accessible at https://webclientside-6db2f2ac8d4b.herokuapp.com.

The front end was built using React and primarily consists of three components:

a. **Login Screen**: We generated a unique password for each of the 18 classifying students and for the 3 professionals.
b. **Classification Screen**: The main screen displays the current tweet, classification options, and features.
c. **User/Admin Panel**: This screen provides relevant statistics about user classifications, such as the distribution of classifications and average classification time. The admin panel offers comprehensive statistics both collectively and per user.

Each tweet is classified twice by two different students. In case of contradictory classifications, the tweet is automatically inserted into the professionals table and is reclassified by one of the instructors.

Please refer to the attached guides for a better understanding of the web application logic and the classification process.

## Classification process

A total of 3,855 tweets were classified with the following distribution: 168 positive (anti-militias), 21 negative (pro-militias), and 3,666 irrelevant. The classification was conducted over four rounds, with the first round serving as a warmup.

In the warmup round, each student classified 200 tweets. The instructors then reviewed and ensured the correctness of these classifications. After the warmup round, the initial classifications were deleted to allow the students to reclassify these tweets during the subsequent three rounds. The students then classified all tweets in three evenly distributed sessions, ensuring an even percentage of tweets pulled from different Twitter users.

## The Model

1. **Algorithm**: We used BERT (Bidirectional Encoder Representations from Transformers) with a pretrained BERT base language model for Arabic, sourced from Hugging Face.

2. **Training and Testing Data**:

   - The model was trained on a dataset consisting of 765 tweets, with 153 tweets labeled as anti-militias and the remaining tweets categorized as irrelevant.
   - The test batch consisted of 75 tweets, with 15 labeled as anti-militias and the remaining tweets categorized as irrelevant.

3. **Classification and Training Process**:

   - The initial training utilized the pretrained BERT tokenizer mentioned above.
   - After the first training, we performed three additional fine-tuning sessions using the previously trained model.
   - Specifications used:
     a. Number of epochs per training: 3.
     b. Training batch size: 16.
     c. Evaluation batch size: 64.
     d. Weight decay: 0.01.
     e. Warmup steps: 10% of the total steps in the training.

4. **Results**:

   - Accuracy: The test set revealed an accuracy of 88%, with the following distribution: 13 out of 15 tweets were correctly classified as anti-militias, and 53 out of 60 tweets were correctly classified as irrelevant.
   - Confusion Matrix: The confusion matrix is best described in the chart below: