# NLP Twitter Analysis

Hamed Feizabadi

July 11, 2023

## 1 Introduction

Our source for this project is Twitter. You should create a file named "users.csv" inside src folder which contains the Twitter username, University name an Actual name of the users you wish to analyze.

## 2 Repository

You can access the source code of this project at https://github.com/hamedhf/nlp_twitter_analysis

## 3 Data Collection

We used selenium >= 4.6.0 for collecting data from Twitter. This tool helps us to bring up an actual browser and navigate through the pages. Note that you should have Chrome installed on your system for this to work. Then you can simply install other dependencies form pyproject.toml file using poetry or other package managers. The crawler script reads the users.csv file and for each user, it navigates to the user's profile and collects the tweets. The tweets are stored in a file named unlabeled.db inside data/raw folder. Then labeling script uses this and with the help of ChatGPT model, it generates the labels for each tweet and stores them in data/raw/labeled-run-date.csv file.

# 4 Data Format

The data is stored in a csv file with the following format: tweet_time, tweet_owner, tweet_text, owner_university, owner_name, label. We use tweet_time, tweet_owner as unique identifiers for each tweet. The tweet_owner is Twitter username of the owner. The tweet_text is the actual text of the tweet. owner_university and owner_name are the university and actual name of the tweet owner. The label is the generated label for the tweet.

# 5 Data Preprocessing

We have splitted data with three criteria: split by sentence with hazm sentence tokenizer, split by word with hazm word tokenizer, split by word with hazm lemmatizer.

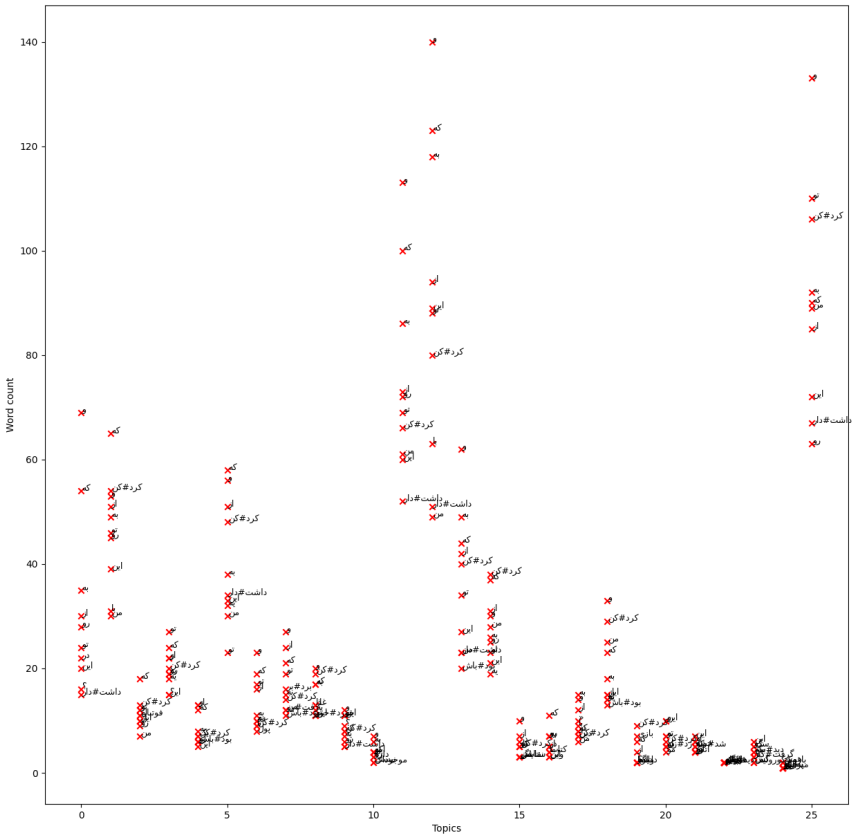
For cleaning the data, we used the following steps: remove emojis, remove urls, remove hashtags, remove mentions, remove numbers, remove punctuations. We used the hazm, cleantext and nltk libraries for this purpose.

# 6 Labeling

We give label to the whole tweet using ChatGPT. For more info about labeling see the src/utils/label.py file. You can also see the labels in src/utils/constants.py file.

# 7 Statistics

| tweet-count | word-count | sentence-count | unique-word-count |
|---|---|---|---|
| 2079 | 31987 | 2944 | 6725 |


Word count
Topics
0
5
10
15
20
25
0
20
40
60
80
100
120
140