

Multilingual Twitter Analysis User Manual

GitHub Repository: <https://github.com/migrationsKB/MRL> (main branch)

Repository Structure:

Only explain the content for updating tweets and analyzing tweets.

(Gray for folders, black for files, text in brackets for comments)

```
| - crawler
    | - config
        | - field_expansions (json files defining fields for crawling tweets1)
        | - keywords (json and csv files for keywords in 11 languages)
        | - credentials.yaml (user-defined yaml file for Twitter API credentials)
    | - main_keywords.py (crawler)
| - models
    | - scripts
        | - inference.py
    | - topicModeling
        | - ETM
            | - data_build_for_inferring_topics.py (script for preparing data for topic modeling)
            | - infer_topic_and_filter.py (infer topics for tweets and filter irrelevant tweets)
| - postprocessor
    | - merging_results.py (merging the results from all semantic analyses)
    | - get_stats_results.py (get the statistics of the results)
| - preprocessor
    | - restructure_data.py
    | - dict2df.py
    | - preprocessing.py
| - utils
| - output (user-create this folder and hierarchy to get the output from the crawling, models
and results)
    | - by_lang (merged the results by language)
    | - crawled (files containing crawled tweets)
    | - merged (merged results from folder results)
    | - models (pre-trained ETM/HSD/SA models, provided as .zip file)
    | - preprocessed (files containing preprocessed tweets for further semantic analysis)
        | - csv (including basic geo information and other meta data of the tweets)
        | - forTP (files feed into ETM and also HSD/SA)
        | - geo
        | - restructured
    | - results
        | - ETM (csv files from Topic Modeling)
        | - HSD (csv files from Hate Speech Detection)
        | - SA (csv files from Sentiment analysis)
```

¹ <https://developer.twitter.com/en/docs/twitter-api/fields>

Crawling Tweets

1. Get Twitter API credentials and put `credentials.yaml` in `crawler/config` folder:

```
migrationsKB:
  bearer_token: XXXX
```

XXXX represents the bearer token for the Twitter API.

2. Specify the country iso2code, batch number (without underline), start year, end year, and the starting index of a list of keywords, change the parameters in `01_run_crawler.sh` or run:

```
python -m crawler.main_keywords "DE" "batch4" 2021 2022 0
```

for example, "DE" is the country iso2code for Germany, "batch4" is the batch number, and 2021 is the start year of the tweets and 2022 is the end year, 0 is the starting index of a list of keywords.

3. The crawled data will be stored in the folder `output/crawled/DE/batch4` with the filename such as:

```
DE_20220722082450_2022-03-31T16:46:09.00Z.gz
```

Preprocessing Tweets

To prepare the data for semantic analyses such as Topic Modeling, Sentiment Analysis and Hate Speech Detection, the tweets need to be preprocessed.

Run the shell script:

```
./02_run_data_preprocessor.sh
```

The preprocessed data stored in `output/preprocessed/forTP` will be used for following steps.

Topic Modeling

Find the pretrained models and responding results of Topic Models in the folder `output/models/ETM`. The file `etm_results.csv` contains the best ETM model in each language and the corresponding topic numbers (K), which are used for define the parameters in the shell script `03_run_topic_modeling.sh`. Change the parameters of language code and number of topics in the script and run:

```
./03_run_topic_modeling.sh
```

The output will be in `output/results/ETM/`.

Sentiment Analysis and Hate Speech Detection

After the tweets are filtered by Topic Modeling, run the script:

```
./04_run_SA_HSD.sh
```

The output will be in `output/results/SA/` and `output/results/HSD/`.

Post Processing

Merging all the semantic analyses results and get the statistics by country and language into jsonl files:

```
./05_run_post_processor.sh
```

The output will be `output/sentiment.jsonl` and `output/hsd.jsonl`.