

# NLP Twitter Analysis

Hamed Feizabadi

July 11, 2023

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Repository</b>	<b>2</b>
<b>3</b>	<b>Requirements</b>	<b>2</b>
<b>4</b>	<b>Installation</b>	<b>2</b>
<b>5</b>	<b>Project Structure</b>	<b>3</b>
<b>6</b>	<b>Data Collection</b>	<b>4</b>
<b>7</b>	<b>Data Format</b>	<b>4</b>
<b>8</b>	<b>Data Preprocessing</b>	<b>4</b>
<b>9</b>	<b>Labeling</b>	<b>4</b>
<b>10</b>	<b>Statistics</b>	<b>5</b>
<b>11</b>	<b>Augmenting Data</b>	<b>6</b>
11.1	Generated Tweets . . . . .	7
11.2	Augmented Data Statistics . . . . .	7
<b>12</b>	<b>Word2Vec</b>	<b>8</b>
12.1	Training . . . . .	8
12.2	Evaluation . . . . .	8

<b>13 Language Model</b>	<b>9</b>
13.1 Fine Tuning . . . . .	10
13.1.1 Training and Validation Loss . . . . .	10
13.2 Generating Tweets . . . . .	12
<b>14 Resources</b>	<b>13</b>

# 1 Introduction

This is the final report for the NLP course project. In this project we have collected tweets from Twitter and labeled them using ChatGPT model. Then we have augmented the data using GPT-3.5-turbo model. After that we have trained a classifier using the augmented data and evaluated the results.

Word2vec, tokenizer and language model and other stuffs also have been trained on the collected data, which we will discuss in the following sections.

# 2 Repository

You can access the source code of this project at [https://github.com/hamedhf/nlp\\_twitter\\_analysis](https://github.com/hamedhf/nlp_twitter_analysis)

# 3 Requirements

If you have a GPU(especially Nvidia) then you are good to go and can run project locally, as we did. Otherwise you can use Google Colab to run the project. For that purpose you need to clone the github repo inside colab and run the commands provided with main.py file. These commands are also provided in the README.md file of the repository.

# 4 Installation

You should create a file named "users.csv" inside src folder which contains the Twitter username, University name and Actual name of the users you wish to analyze.

Furthermore installation instructions are provided in the README.md file of the repository.

## 5 Project Structure

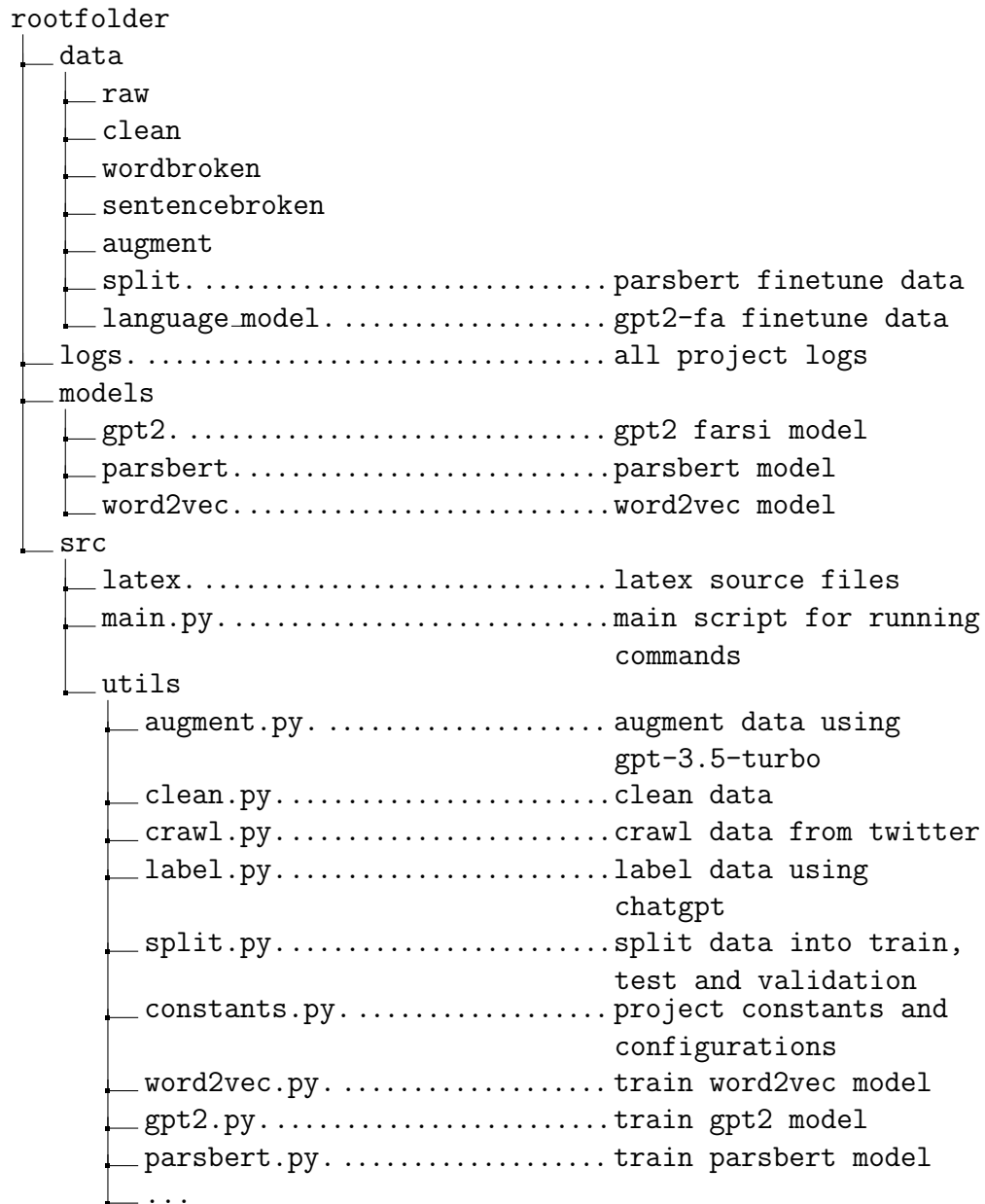


Figure 1: Project Tree

## 6 Data Collection

We used selenium  $\geq 4.6.0$  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 from 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.

## 7 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.

## 8 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.

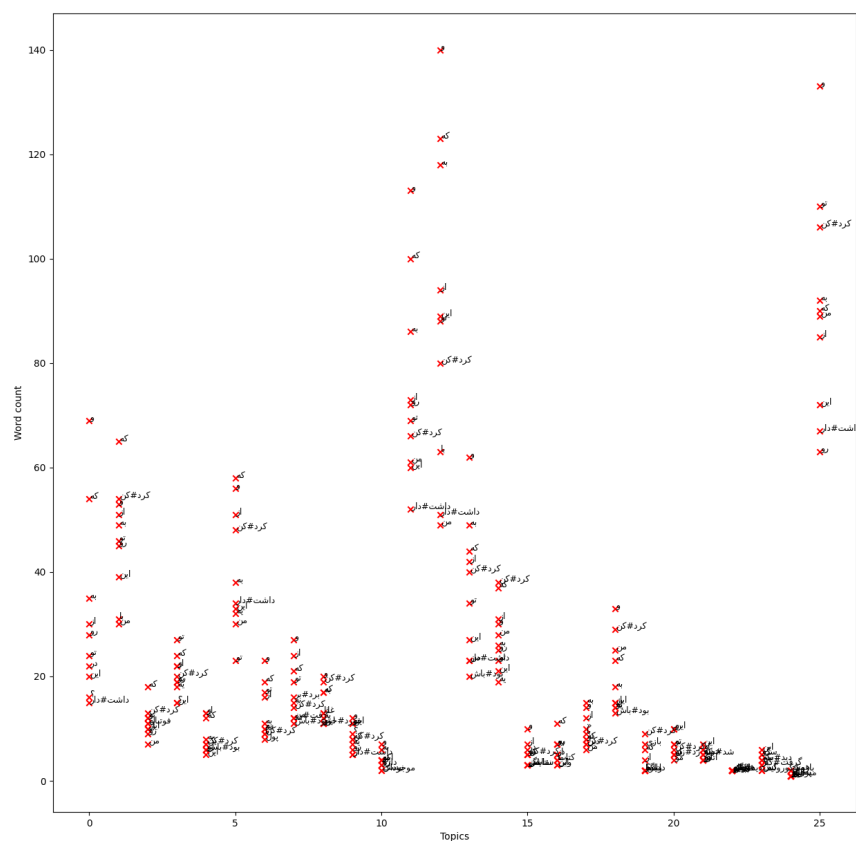
## 9 Labeling

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

constants.py file.

## 10 Statistics

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



## 11 Augmenting Data

For this part to work you need to sign up for an account at <https://platform.openai.com> and provide your openai api key in .env file. Then you can run the augment-data command.

The augmentation script takes a cleaned csv file as input and counts how many tweet we have for each label. Then it will fix the imbalance of the data by generating new tweets for the labels with less tweets. The generated tweets are stored in data/augment folder.

You can see the implementation detail of the augmentation script in src/utlis/augment.py file. We have used gpt-3.5-turbo model for this purpose and the given prompt is like this:

```
1  import openai
2  label = "home_and_garden"
3  temperature = 0.6
4  system_message = "Generate an informal Persian tweet
5  about the given topic without any hashtags, mentions,
6  links, or emojis." # noqa
7  messages = [
8      {
9          "role": "system",
10         "content": system_message
11     },
12     {"role": "user", "content": f"topic: {label}"}
13 ]
14 response = openai.ChatCompletion.create(
15     model="gpt-3.5-turbo",
16     messages=messages,
17     temperature=temperature,
18     timeout=120
19 )
```

Temperature is a parameter that controls the randomness of the generated text. The higher the temperature, the more random the text. The lower the temperature, the more predictable the text. We have used 0.6 for this parameter and it is random enough for our purpose and if we increase it will take much more time to generate the text and this is not practical for our purpose.

Using this approach we have doubled our total data size and each label has at least 200 tweets. It is worth mentioning that because of openai api rate limit, it took us about **2 and a half days** to generate the data.

## 11.1 Generated Tweets

Some of the generated tweets:

هنر و طراحی چهره که زندگی رو جادو و زیبا میکنه تا نگاه به آفرمای مهربانان و طراحان خردمون رو در دنیای خفایان	openai	openai	art_and_design
دین و معنویت به چهره که هر کسی بهش نیاز داره حتی اگه نپذیرد به هیچ دینی تعلق داشته باشه همه ما به دنبال به چهره	openai	openai	religion_and_spirituality
جهان ما تنها محلی است که برای زندگی ما وجود داره بنابراین برای حفظ محیط زیست و پایداری آن تلاش کنیم بیجما ره آفرین	openai	openai	environment_and_sustainability
ورزش و ورزشکاری یکی از بهترین راهها برای سلامتی و خوشحالیه معینه باید به خودمون وقت بایم و ورزش کنیم ترند نبوت	openai	openai	sports_and_athletics
کتاب میرونین گربهها میخوان بهترین دوستای انسان باشن! منم به گربه دارم که همیشه باهامه و خیلی دلم براش تنگ میغه :-	openai	openai	pets_and_animals
دین و معنویت دوسانه میکن که برای روحیه و آرامش خودمون به این دوتیتر نیاز داریم هرکس طبق اعتقادات و معتقدات خود	openai	openai	religion_and_spirituality
اوتشدر از کرمای صورت استفاده میکنم که لوی صبح کارمون رو شروع نمیکشیم تا من معنی مراحل آرایشمو نفهم کنم	openai	openai	beauty_and_cosmetics
ورزش و ورزشکاری یکی از بهترین راهها برای سلامتی و شادابی بدنه همیشه باید فعالیت و ورزش رو تو روزمرهات جا داد و	openai	openai	sports_and_athletics

## 11.2 Augmented Data Statistics

	label	tweet count
1	politics_and_current_affairs	200
2	entertainment_and_pop_culture	200
3	sports_and_athletics	200
4	technology_and_innovation	200
5	science_and_discovery	200
6	health_and_wellness	200
7	business_and_finance	200
8	travel_and_adventure	200
9	food_and_cooking	200
10	fashion_and_style	200
11	environment_and_sustainability	200
12	education_and_learning	216
13	social_issues_and_activism	217
14	inspirational_and_motivational	200
15	funny_and_humorous	200
16	art_and_design	200
17	books_and_literature	200
18	religion_and_spirituality	200
19	family_and_parenting	200
20	gaming	200
21	beauty_and_cosmetics	200
22	home_and_garden	200
23	automotive	200
24	pets_and_animals	200
25	weather_and_seasons	200
26	other	413



## 12 Word2Vec

We used gensim library for training skipgram word2vec model because it is easy to use and fast. The implementation is in src/utlis/word2vec.py file.

All of the available commands are listed in **ReadME.md** file. Here we explain some of them.

### 12.1 Training

This command trains word2vec for a specific label.

```
1 python src/main.py train-word2vec-label path-to-augmented
2 -csv home_and_garden
```

This command trains word2vec for some preselected labels.

```
1 python src/main.py train-word2vec-preselected path-to-
2 augmented-csv
```

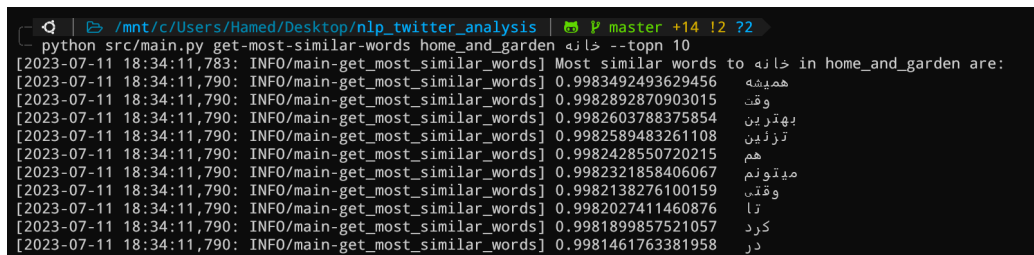
This command trains word2vec for all labels.

```
1 python src/main.py train-word2vec-all path-to-augmented-
2 csv
```

Each model is saved in a models/word2vec/label.npy file.

### 12.2 Evaluation

Let's find that in topic of home\_and\_garden, which words are similar to the Persian word for home (khaneh).



```
python src/main.py get-most-similar-words home_and_garden خانه --topn 10
[2023-07-11 18:34:11,783: INFO/main-get_most_similar_words] Most similar words to خانه in home_and_garden are:
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9983492493629456 همیشه
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9982892870903015 وقت
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9982603788375854 بهترین
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9982589483261108 تزئین
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9982428550720215 هم
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9982321858406067 می‌تولم
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9982138276100159 وقتی
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9982027411460876 تا
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9981899857521057 کرد
[2023-07-11 18:34:11,790: INFO/main-get_most_similar_words] 0.9981461763381958 در
```

Some of the similarity results are shown in the following image. We use cosine similarity for measuring the similarity between two words. The higher the similarity, the more similar the words are.

کلمات مشابه برای کلمه سیاست: [ 'جاری', (0.984733521938324), 'امور', (0.9736297726631165), 'پیچیده', (0.9675320982933044), 'روزها', (0.9606905579566956), 'سیاسی', (0.9490455389022827), 'بازار', (0.9372583627700806), 'اخبار', (0.9359972476959229), 'شده', (0.9271780252456665), 'تجارت', (0.925977885723114), 'گرون', (0.9195088744163513) ]	کلمات مشابه برای کلمه ایران: [ 'تعداد', (0.9875343441963196), 'خوابگاه', (0.9871758818626404), 'حکم', (0.9862680435180664), 'ددلاین', (0.9860983490943909), 'اونور', (0.9860574007034302), 'اولش', (0.9858595728874207), 'انقلاب', (0.9857739210128784), 'تنها\c200uتری', (0.9856471419334412), 'بی\c200uگناه', (0.9855296611785889), 'تومن', (0.9853056073188782) ]
---	--

کلمات مشابه برای کلمه گل: [ 'باغچه\c200uام', (0.9820027351379395), 'گل\c200uha', (0.9808597564697266), 'دکوراسیون', (0.9766149520874023), 'آروم', (0.970939040184021), 'می\c200uگدرونم', (0.9690070748329163), 'گیاهاشم', (0.9681393504142761), 'باغچه\c200uی', (0.9678846001625061), 'باغ', (0.9676481485366821), 'گیاه', (0.9649395942687988), 'تمیز', (0.9636048674583435) ]	کلمات مشابه برای کلمه ماشین: [ 'بخرم', (0.9738640189170837), 'روزی', (0.9613807797431946), 'یکیشونو', (0.9528212547302246), 'قدرتمند', (0.948646605014801), 'بتونم', (0.9442844390869141), 'خفن', (0.9425691366195679), 'خیابونا', (0.9352948069572449), 'میخوام', (0.9346776604652405), 'بشم', (0.9335148334503174), 'خرید', (0.9271440505981445) ]
---	--

## 13 Language Model

We used huggingface transformers library for training the language model. The implementation is in src/utlis/gpt2.py file. We used HooshvareLab/gpt2-fa model for this purpose. The model is trained on huge Persian corpus and it is available at <https://huggingface.co/HooshvareLab/gpt2-fa>. Use of pretrained model helps us gain better results with less data.

## 13.1 Fine Tuning

The below command shows how to fine tune the model for a specific label. The script will check for dataset in data/language\_model and if it is not available, it will create it using the augmented data. Then it will fine tune the model for the given label and save the model in models/gpt2/label folder.

```
python src/main.py fine-tune-gpt2 ../data/augment/augmented_2023-06-02-10-27-57.csv --desired-label home_and_garden
[2023-07-11 10:49:27.959: INFO/main-fine_tune_gpt2] Fine tuning gpt2...
[2023-07-11 10:49:27.964: INFO/main-fine_tune_gpt2] Dataset already prepared. Skipping preparation step.
[2023-07-11 10:49:28.084: INFO/main-fine_tune_gpt2] Using device: cuda
[2023-07-11 10:49:28.084: INFO/main-fine_tune_gpt2] Fine tuning gpt2 for label: home_and_garden
Model already exists!
Model copied successfully!
tokenizer.json already exists!
tokenizer.json copied successfully!
0%|          | 0/3 [00:00<?, 7it/s]
Beginning epoch 1 of 3
100%|          | 20/20 [00:06<00:00, 3.32it/s]
Average Training Loss: 6.84556020796299. Epoch time: 0:00:06
100%|          | 5/5 [00:00<00:00, 8.69it/s]
Validation loss: 0.7264172911643982. Validation Time: 0:00:01
33%|          | 1/3 [00:06<00:13, 6.59s/it]
Beginning epoch 2 of 3
100%|          | 20/20 [00:05<00:00, 3.94it/s]
Average Training Loss: 0.6327574968338012. Epoch time: 0:00:05
100%|          | 5/5 [00:00<00:00, 8.62it/s]
Validation loss: 0.5444896697998047. Validation Time: 0:00:01
67%|          | 2/3 [00:12<00:06, 6.04s/it]
Beginning epoch 3 of 3
100%|          | 20/20 [00:05<00:00, 3.92it/s]
Average Training Loss: 0.4751459345221519. Epoch time: 0:00:05
100%|          | 5/5 [00:00<00:00, 8.56it/s]
Validation loss: 0.48139827847480776. Validation Time: 0:00:01
100%|          | 3/3 [00:17<00:00, 5.98s/it]
Total training took 0:00:18
```

### 13.1.1 Training and Validation Loss

Following image shows the training and validation loss for some of the pre-selected labels, which you can see the list of them in src/utils/constants.py file. Also you can fine tune the model for them too. If you provide a label then it will be fine tuned for that label, otherwise it will be fine tuned for all of the preselected labels.



(a) education\_and\_learning



(b) environment\_and\_sustainability



(c) home\_and\_garden



(d) politics\_and\_current\_affairs



(e) weather\_and\_seasons

All implementation details are in `src/utlis/gpt2.py` file. Each model is saved in `models/gpt2/label` folder.

## 13.2 Generating Tweets

If we fine tune gpt2-fa on politics label and we give it prompt about politics, we expect it to generate a tweet about politics. The following image shows the result of this experiment.

```

$ /mnt/c/Users/Hamed/Desktop/nlp_twitter_analysis | P master +8 18
python src/main.py complete-prompt-gpt2 "سیاستمدار همه دروغ" politics_and_current_affairs
[2023-07-11 20:20:08,474: INFO/main-complete_prompt_gpt2] Using device: cuda
[2023-07-11 20:20:08,474: INFO/main-complete_prompt_gpt2] Completing prompt: سیاستمدار همه دروغ
This is what the model is given as input: <P>سیاستمدار [startoftext].
[2023-07-11 20:20:13,453: INFO/main-complete_prompt_gpt2] Generated outputs:
سیاستمدار همه دروغاین سیاستا یعتب چه چیزن؟ هر چند سال یه بار باید از این نوبته ها حزب برنیم ولت هیچ نباید از سیاست و
امور خارج قرار کنیم باید همیشه با نوبته قرار کنیم
سیاستمدار همه دروغدولت واقعا باید چیکار کنه که اولاً از رئیس جمهور کنولت یه دیکتاتور بت گناه بت خبر نیامه و دوماً از رهبر د
نوبت همه چیز به نفع خودمونه به نظرم
سیاستمدار همه دروغاوهوم من که هیچ وقت هیچ چیز شکت نیست و هیچ کس سلطان نیست چه خبره از این سیاستا چه خبر است؟
[2023-07-11 20:20:13,453: INFO/main-complete_prompt_gpt2]
$ /mnt/c/Users/Hamed/Desktop/nlp_twitter_analysis | P master +8 18

```

It is important to note that we should set `max_seq` length according to the common length of the tweets in the dataset. If we set it a large number, then model will be overfitted on PAD token and it will generate a lot of PAD tokens.

## 14 Resources

- <https://github.com/hooshvare/parsgpt>
-