

NLP Twitter Analysis

Hamed Feizabadi

July 12, 2023

Contents

1	Introduction	2
2	Repository	2
3	HuggingFace Dataset	2
4	Requirements	2
5	Installation	2
6	Project Structure	3
7	Data Collection	4
7.1	Tools	4
7.2	Process	4
8	Data Format	4
9	Data Preprocessing	5
10	Labeling	5
11	Statistics	5
12	Augmenting Data	6
12.1	Generated Tweets	7
12.2	Augmented Data Statistics	8

13 Word2Vec	8
13.1 Training	9
13.2 Evaluation	9
14 Tokenizer with byte pair encoding	10
15 Language Model	11
15.1 Fine Tuning	11
15.1.1 Training and Validation Loss	11
15.2 Generating Tweets	13
16 Model Selection	14
17 Classification	14
17.1 Fine Tuning	15
17.2 Testing	16
17.3 Inference	16
18 OpenAI API for classification	16
18.1 Accuracy	18
18.2 Inference	18
19 Report Generation	19
20 Resources	19

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

3 HuggingFace Dataset

The dataset is available at https://huggingface.co/datasets/hamedhf/nlp_twitter_analysis/tree/main

4 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.

5 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.

6 Project Structure

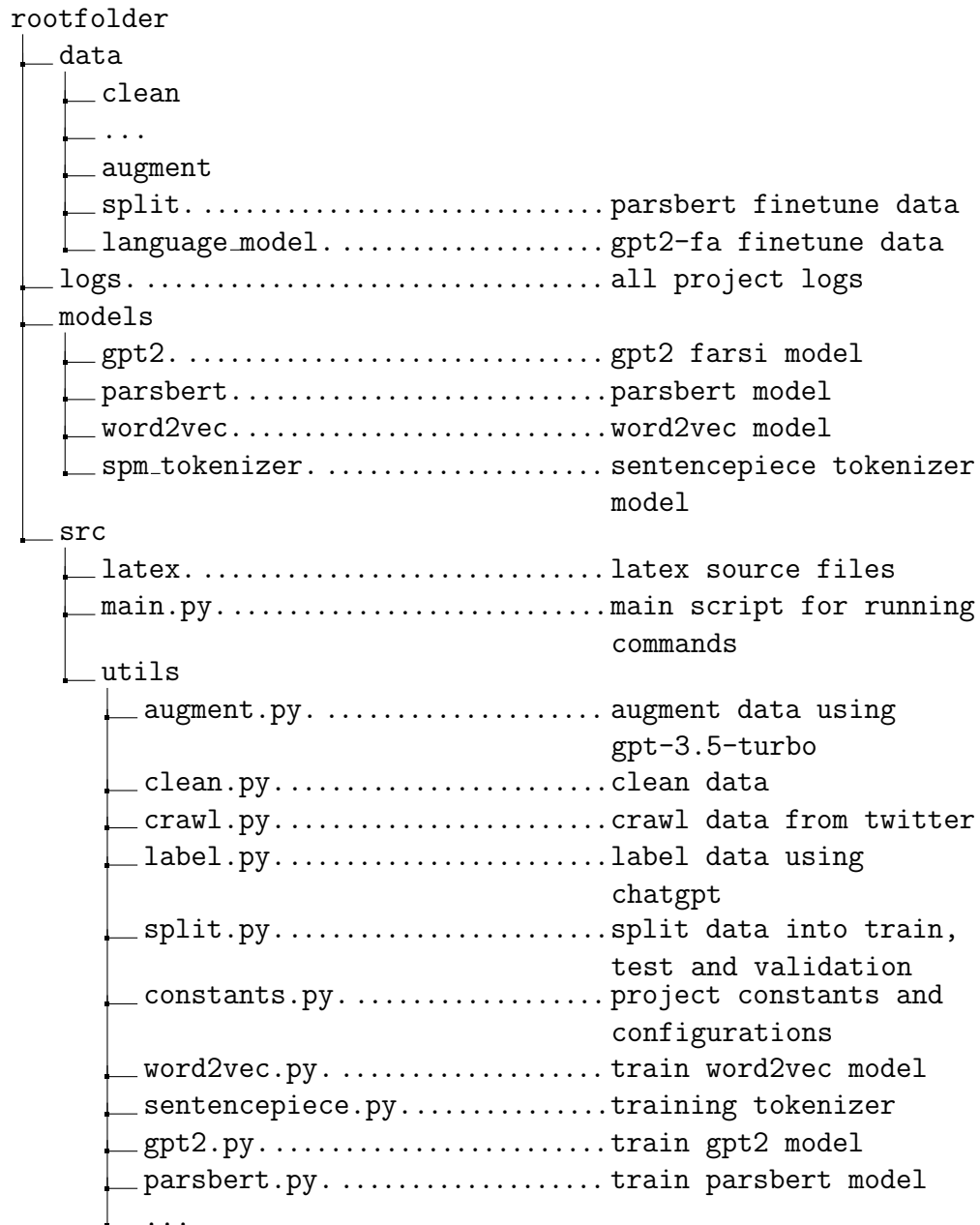


Figure 1: Project Tree

7 Data Collection

7.1 Tools

Selenium is an invaluable tool for our project, as it enables us to efficiently crawl data from Twitter. With its web automation capabilities, Selenium allows us to interact with the Twitter website programmatically, replicating user actions and extracting the data we need. We can navigate through Twitter’s pages, search for specific tweets, scrape user profiles, and extract information such as tweets, hashtags, user handles, and timestamps. The flexibility of Selenium enables us to dynamically interact with Twitter’s interface, making it ideal for our project’s data crawling requirements. By leveraging Selenium’s power, we can gather and analyze Twitter data in real-time, facilitating tasks such as sentiment analysis, social network analysis, and trend monitoring.

7.2 Process

It is worth to mention that this part was one of the most time consuming and hardest parts of the project. Because we had to mock the behavior of a real user and navigate through the pages.

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.

8 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.

9 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.

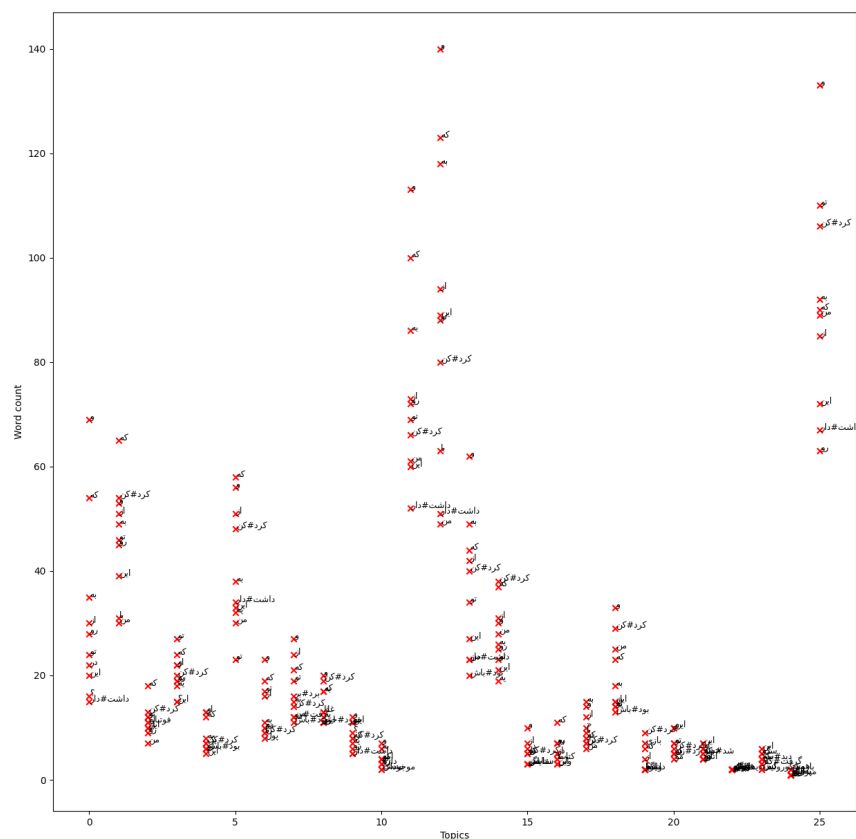
For cleaning the data, we used the following steps: remove emojis, remove urls, remove hashtags, remove mentions, remove numbers, remove punctuations. We used the hazm, cleantext and nltk libraries for this purpose.

10 Labeling

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

11 Statistics

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



12 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/ utils/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.

12.1 Generated Tweets

Some of the generated tweets:

مدر و طراحی جزیره که زندگی رو جذاب و زیبا میکنه با نگاه به الزامات معرستان و طراحی خردمند رو در دنیای خلاقیت	openai	openai	art_and_design
دین و معنویت به چهره که هر کس بهش نیاز داره حلی اگه نمیتواند به هیچ دینی تعلق داشته باشه معه ما به دنبال به چپ	openai	openai	religion_and_spirituality
جهان ما تنها محلی است که برای زندگی ما وجود دارد بنابراین برای حفظ محیط زیست و پایداری آن تلاش کنیم بهجمله را آموزید	openai	openai	environment_and_sustainability
ورزش و ورزشکاری یکی از بهترین راهها برای سلامتی و خوشحالیه معینه باید به خودمون وقت بدم و ورزش کنیم بزرگ بشویم	openai	openai	sports_and_athletics
کیا می‌تونین گزینه‌ها می‌تونن بهترین دوستای انسان باشن؟ منم به گزینه دارم که همیشه باهامه و خیلی دلم براش تنگ میشه	openai	openai	pets_and_animals
دین و معنویت دونهاییه نیکن که برای روحیه و آرامش خودمون به این دوتایی نیاز داریم هرکسی طبق اعتقادات و معنقدات خود	openai	openai	religion_and_spirituality
اوتیغدر از کرمای صورت استفاده میکنم که شوی صبح کارمون رو شروع میکنم تاس منم مراحل آرایشمو نیموم کنم	openai	openai	beauty_and_cosmetics
ورزش و ورزشکاری یکی از بهترین راهها برای سلامتی و خدایی بدانه همیشه باید فعالیت و ورزش رو تو روزمرهات جا داد و	openai	openai	sports_and_athletics

12.2 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

13 Word2Vec

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

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

13.1 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
```

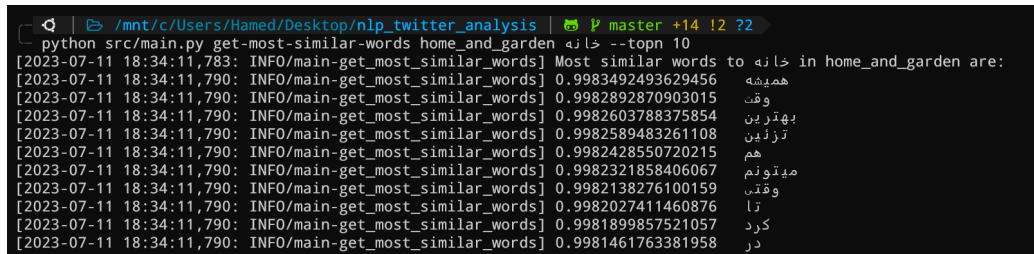
This command trains word2vec for all labels.

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

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

13.2 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 نزله
[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 می‌تونم
[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:	کلمات مشابه برای کلمه ایران:
[('جاری', 0.984733521938324),	[('تعداد', 0.9875343441963196),
('امور', 0.9736297726631165),	('خوابگاه', 0.9871758818626404),
('پیچیده', 0.9675320982933044),	('حکم', 0.9862680435180664),
('روزها', 0.9606905579566956),	('ددلاین', 0.9860983490943909),
('سیاسی', 0.9490455389022827),	('اونور', 0.9860574007034302),
('بازار', 0.9372583627700806),	('اولش', 0.9858595728874207),
('اخبار', 0.9359972476959229),	('انقلاب', 0.9857739210128784),
('شده', 0.9271780252456665),	('تنها\200cتری', 0.9856471419334412),
('تجارت', 0.925977885723114),	('بی\200cگناه', 0.9855296611785889),
('گرون', 0.9195088744163513)]	('تومن', 0.9853056073188782)]

کلمات مشابه برای کلمه گل:	کلمات مشابه برای کلمه ماشین:
[('باغچه\200cام', 0.9820027351379395),	[('بخرم', 0.9738640189170837),
('گل\200cها', 0.9808597564697266),	('روزی', 0.9613807797431946),
('دکوراسیون', 0.9766149520874023),	('یکیشونو', 0.9528212547302246),
('آروم', 0.970939040184021),	('قدرتمند', 0.948646605014801),
('می\200cگدروم', 0.9690070748329163),	('بتونم', 0.9442844390869141),
('گیاهاشم', 0.9681393504142761),	('خفن', 0.9425691366195679),
('باغچه\200cی', 0.9678846001625061),	('خیابونا', 0.9352948069572449),
('باغ', 0.9676481485366821),	('میخوام', 0.9346776604652405),
('گیاه', 0.9649395942687988),	('بشم', 0.9335148334503174),
('تمیز', 0.9636048674583435)]	('خرید', 0.9271440505981445)]

14 Tokenizer with byte pair encoding

With the help of sentencepiece library we trained a tokenizer with byte pair encoding. The implementation is in src/utls/sentencepiece.py file. We used the augmented data for training the tokenizer. The tokenizer is saved in models/spm.tokenizer/tokenizer(vocab_size).model file.

We have trained the tokenizer with different vocab_size and tested them with this metric that how many of the words in the test set will be mapped to UNK token. The results are shown in the following table.

vocab_size	UNK count
100	316
1000	57

As you can see, larger vocab_size leads to less UNK count on the test set.

15 Language Model

We used huggingface transformers library for training the language model. The implementation is in src/utls/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.

15.1 Fine Tuning

The below command shows how to fine tune the model for a specific label. The script will check for dataset in data/ languagemodel 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
```

15.1.1 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/utls/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.



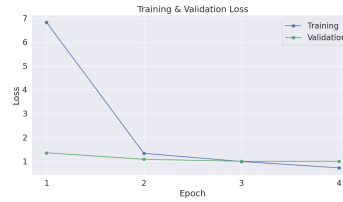
(a) education_and_learning



(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.

15.2 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.

```

[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>دروغ</P>[startoftext].
[2023-07-11 20:20:13,453: INFO/main-complete_prompt_gpt2] Generated outputs:
سیاستمدار همه دروغهای سیاست یهت چه چیزن؟ هر چند سال یه بار باید از این توطئه ها خبر بزنیم وگت هیچ نباید از سیاست و
امور جاری هزار کلمه باید همیشه با توطئه هزار کلمه
سیاستمدار همه دروغدولت واقعا باید چکار کنه که اولاً از رئیس جمهور کنولت یه دیکناتور بت گناه بت خبر نیامه و دوماً از رهبر د
تونی همه چیز به نفع خودتونه به نازم
سیاستمدار همه دروغاوهوم من که هیچ وقت هیچ چیز شکت نیست و هیچ کس مطمئن نیست چه خبره از این سیاستا چه خبر است؟
[2023-07-11 20:20:13,453: INFO/main-complete_prompt_gpt2]

```

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.

16 Model Selection

We have mainly worked on 2 models for classification, one was parsbert v2 and other one is an LSTM based model. Unfortunately we could not use the LSTM based model because of the lack of time. We have finetuned the parsbert v2 model on the augmented data. The implementation of the LSTM based model is in `src/models/lstm.py` file. The implementation of the parsbert v2 model is in `src/models/parsbert.py` file. In next sections we will explain the parsbert v2 model in more details.

17 Classification

We have utilized the `HooshvareLab/bert-fa-base-uncased` model (parsbert v2), a variant of BERT specifically designed for Persian language processing, to extract features from the tweets for our text classification task. The tweets were tokenized using the ParseBERT tokenizer, which splits the text into individual tokens, including special tokens like [CLS] denoting the classification task. The ParseBERT model, consisting of multiple transformer layers, was then employed to process the tokenized input and learn contextual dependencies among the tokens. By leveraging the pre-trained contextualized representations of the `HooshvareLab/bert-fa-base-uncased` model, we could effectively capture the semantic and syntactic information within the Persian tweets. Finally, the contextualized representation of the [CLS] token was passed through a classification head, enabling the model to map the features to the appropriate number of output labels. This approach empowered us to perform accurate and efficient text classification on Persian tweets.

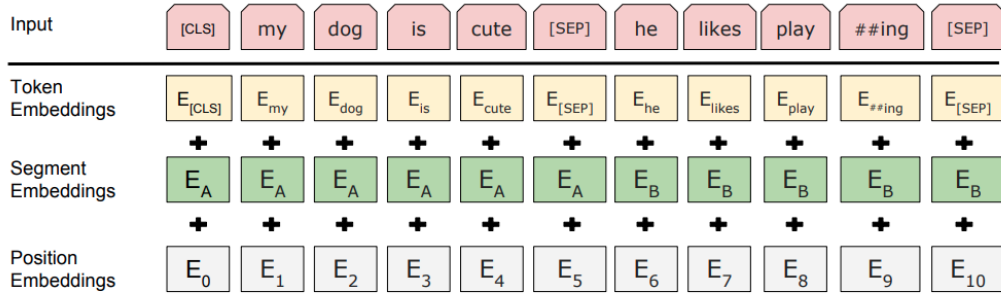


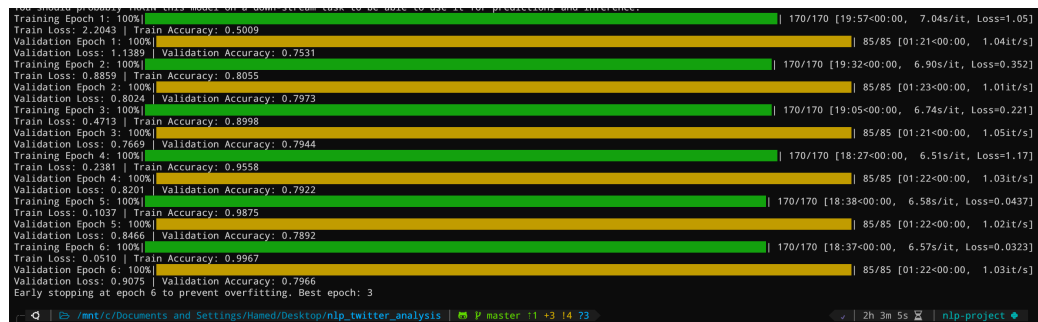
Figure 2: [CLS] token representation is used for classification.

The model is available at <https://huggingface.co/HooshvareLab/bert-fa-base-uncased>. Also you can find the implementation details in src/utlis/parsbert.py file.

17.1 Fine Tuning

```
1 python src/main.py fine-tune-parsbert path-to-augmented-
2 csv
```

The below image shows the process of fine tuning parsbert model. The script will check for dataset in data/split and if it is not available, it will create it using the augmented data.



We have used early stopping technique to prevent overfitting. The accuracy on validation set is roughly 80%.

The model is saved in `models/parsbert/best` and `models/parsbert/final` folder. The best model is the model with the highest accuracy on validation set. The last model is the model at the end of the training process.

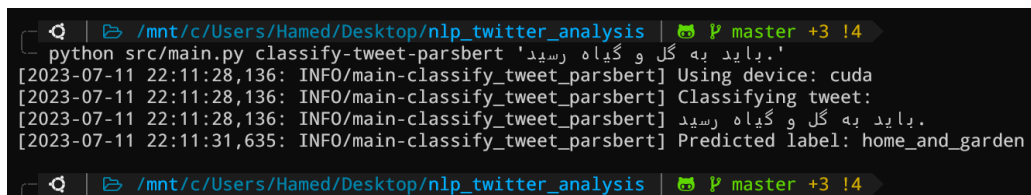
17.2 Testing

```
1 python src/main.py test-parsbert
2
```

test-loss	test-accuracy
0.7573	0.7985

17.3 Inference

Following image shows how to use our main model for inference. The model will be loaded from `models/parsbert/best` folder.



```
/mnt/c/Users/Hamed/Desktop/nlp_twitter_analysis | master +3 !4
python src/main.py classify-tweet-parsbert 'بايد به گل و گياه رسيد.'
[2023-07-11 22:11:28,136: INFO/main-classify_tweet_parsbert] Using device: cuda
[2023-07-11 22:11:28,136: INFO/main-classify_tweet_parsbert] Classifying tweet:
[2023-07-11 22:11:28,136: INFO/main-classify_tweet_parsbert] بايد به گل و گياه رسيد.
[2023-07-11 22:11:31,635: INFO/main-classify_tweet_parsbert] Predicted label: home_and_garden
```

18 OpenAI API for classification

We used OpenAI API to classify the tweets. The following code shows what prompt was given to the API.

Probably the most important parameter is the temperature parameter. It controls the randomness of the generated text. The higher the temperature, the more random the generated text is. The lower the temperature, the more predictable the generated text is. The default value is 0.7. We have used 0.2 so that the model be more deterministic.

```
1 def get_tweet_label(
2     api_key: str,
3     api_base: str,
4     tweet: str,
5     sleep_seconds: int = 10
6 ) -> str:
```

```

7      openai.api_key = api_key
8      openai.api_base = api_base
9      topics = list(TOPICS.values())
10     messages = [
11         {
12             "role": "system",
13             "content": f"Classify the topic of the future
14             tweet into only one of the following categories: {' ', ' '.
15             join(topics)}. some of these tweets are in slang persian
16             language. please try to understand them. Just type the
17             topic and nothing else."
18         },
19         {"role": "user", "content": f"Tweet: {tweet}"},
20     ]
21
22     while True:
23         try:
24             print("*" * 100)
25             print(messages)
26             response = openai.ChatCompletion.create(
27                 model="gpt-3.5-turbo",
28                 messages=messages,
29                 temperature=0.2, # we want the model to
30                 be more deterministic
31             )
32             print(response)
33             print("*" * 100)
34             label = str(response['choices'][0]['message']
35             [['content']]).strip()
36             label = get_clean_label(label)
37             break
38         except (ServiceUnavailableError, APIError,
39                 Timeout, RateLimitError):
40             print(f"Some error occurred. Sleeping for {
41                 sleep_seconds} seconds and trying again")
42             time.sleep(sleep_seconds)
43         except KeyError:
44             print("KeyError occurred. Setting label to '
45             unknown'.")
46             label = 'unknown'
47             break
48     return label

```

18.1 Accuracy

Because we are doing classification task, accuracy is the most important metric and we don't need to calculate other metrics like precision, recall and f1 score.

The following table shows the accuracy of the model on small test dataset.

total-tweets	total-correct	accuracy
78	63	0.8077

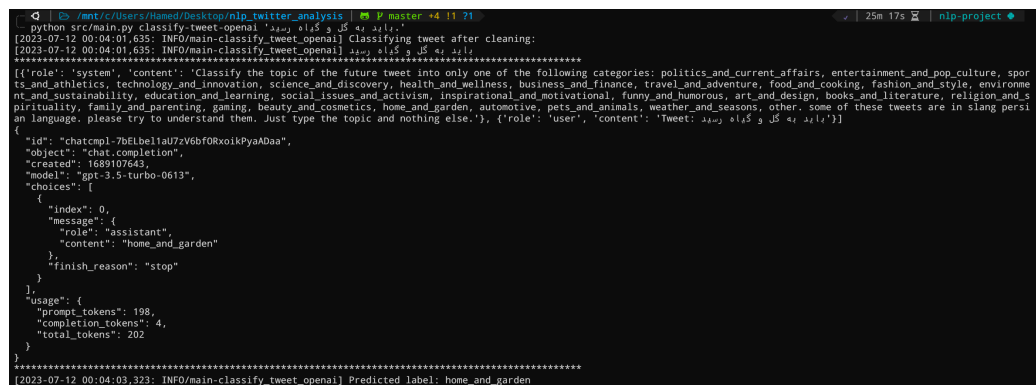
The accuracy is a little bit better than the parsbert model. It was expected because the model is trained on a huge dataset and it is a state of the art model.

There are couple of problems with this approach:

- Very strict API rate limits. To overcome this we had to shrink the test dataset massively.
- The model is not deterministic. We can lower the randomness by lowering the temperature parameter but at end of the day it is still a random process.
- The model is not open source. We have to use the API which is not free.

18.2 Inference

The following image shows how to use the OpenAI API for inference.



```
python src/main.py classify_tweet_openai "دايد به گل و گلاره رسيد"
[2023-07-12 00:04:01.635: INFO/main-classify_tweet_openai] Classifying tweet after cleaning:
دايد به گل و گلاره رسيد
[2023-07-12 00:04:01.635: INFO/main-classify_tweet_openai] Predicted label: home_and_garden
```

19 Report Generation

The report generation process is completely automated. You can find related commands about generating phase 1 and phase 2 reports in ReadME.md file.

20 Resources

- <https://github.com/hooshvare/parsgpt>
- <https://huggingface.co/HooshvareLab/bert-fa-base-uncased>
- <https://platform.openai.com/docs/introduction>
- <https://www.geeksforgeeks.org/python-word-embedding-using-word2vec>
- <https://www.kaggle.com/code/akshat0007/bert-for-sequence-classification>
- <https://www.kaggle.com/code/nulldata/fine-tuning-gpt-2-to-generate-netflix-descriptions>
- https://colab.research.google.com/github/hooshvare/parsgpt/blob/master/notebooks/Persian_Poetry_FineTuning.ipynb
- <https://medium.com/geekculture/easy-sentencepiece-for-subword-tokenization-in-python-and-tensorflow-4361a1ed8e39>