

# ISA5810 / Autumn 2024

## Lab2 Homework

NTHU 112062627 黃星翰

GitHub ID: FrHTate / Kaggle ID: Sakun

### Running Environment

- **System:** Ubuntu 22.04.5 LTS
- **Processor:** Intel(R) Core(TM) i9-10940X CPU @ 3.30GHz
- **Memory:** 64.0 GB
- **GPU:** GeForce 3090 24GiB
- **Python Version:** 3.9.6
- **Python Library:** numpy, pandas, scikit-learn, transformers, torch, tqdm, matplotlib, seaborn, (nlk)

### Private Leaderboard Snapshot

DM2024 ISA5810 Lab2 Homework

Late Submission

Overview Data Code Models Discussion **Leaderboard** Rules Team Submissions

12	▲ 1	YHK115		0.54875	17	5d
13	▼ 1	Yang Hsueh		0.54842	15	3d
14	—	Sakun		0.54433	5	3d

### Preprocessing

1. **Feature Subset Selection:** Since all the data in attributes like '`_index`', '`_source`', '`_hashtags`' are all the same, so there is no helpful information for me to do classification. Hence, I decide to delete these attributes.
2. **Data Cleaning:**
  - I. Check how many duplicated texts, but some of them might be an object of testing dataset, so I will remove them later.
  - II. Check how many empty texts, if there are any empty text, then it should be removed. (Since it won't provide any value for training.)
  - III. Check how many empty hashtags to help me decide how to utilize it.
  - IV. Rename the data to have a better format for me to use.
3. **Split Dataset & Label for Training Dataset**
4. **Remove Duplicated Training Data**
5. **EDA:**

I. **Visualization (to see the relation between emotion and crawldate):** To see whether the data distribution is imbalance, I plot the distribution of emotion. Moreover, I want to know if the time feature can give me any information about the data, so I plot:

- i. The distribution of post date. (for each month)
- ii. The distribution of post time. (for each hour)

Also, I want to see whether 2016 US presidential election affect the distribution of emotion, so I plot the change of emotion over time by month. (only training data)

And I want to see whether people have different emotion at different time, so I plot the change of emotion over time in a day. (only training data)

Finally, I find out that:

- i. In this dataset, the distribution of emotion is very imbalance.
- ii. In this dataset, people don't have significantly change of emotion over time.

**(Quite weird, very unnatural.)** So, the time is independent with emotion.

II. **Visualization (to see the relation between emotion and score):** Except the time feature, I also want to know if the score can give me another information about the data, so I plot:

- i. The distribution of score.
- ii. The distribution of score for each emotion. (only training data)

And I find out that:

- i. In this dataset, the distribution of the score is almost uniform.
- ii. In this dataset, the score is independent with emotion.

6. **Text Segmentation:** I split the word when texts encounter blank space, so I can do preprocess more easily.

7. **Text Cleaning:**

I. **Replace @ symbol:**

- i. **Replace E-mails with <eml>:** I want to help the model recognize the position has an E-mail originally, so it won't 'confuse'.
- ii. **Replace mentions with <usr>:** Reason is similar as the replacement of E-mails.

II. **Replace URLs:** I found out there are about two types URLs in the dataset:

- i. Start with `http://`, `https://`, `www.`, `://`, `//`
- ii. Containing `.com`

And I decide to replace all of them into **<url>** to avoid noises, but let the model know there was an URL at same time.

III. **Replace Special Symbols:**

- i. I observe that `<LH>` seems to be some masked words, so I decide to replace it with **<mask>** (the mask token of RoBERTa)
- ii. **Replace remaining "@" with "at":** I want to make the model know more about

what this sign means. (After the competition, I think this action may be redundant, I should do analyze first so it will be more reasonable.






- iii. **Remove number sign:** I want to drop the number sign, so the model won't be affected by it when seeing a hashtag.

## Model

During the competition, I utilize two model:

1. **Bert (bert-large-uncased):** The reason I choose uncased model is that I think the sentiment classification task on social media has little need of cased things. Since they rarely contain proper nouns in the texts.
2. **RoBERTa (roberta-base):** The reason I choose the model is that I have known it has better performance on sentiment classification task and it utilize more training data to have a better understanding of words compares to BERT.

## Results

	<b>submission5.csv</b> Complete · 4d ago · roberta-base ep2	<b>0.53638</b>	<b>0.54904</b>
	<b>submission4.csv</b> Complete · 4d ago · bert-large ep3	<b>0.53066</b>	<b>0.54608</b>
	<b>submission3.csv</b> Complete · 4d ago · roberta-base ep3	<b>0.54305</b>	<b>0.55704</b>
	<b>submission2.csv</b> Complete · 4d ago · roberta-base ep4	<b>0.54433</b>	<b>0.55648</b>
	<b>submission1.csv</b> Complete · 4d ago · roberta-base ep5	<b>0.54297</b>	<b>0.55484</b>

The first attempts, I upload RoBERTa-base model fine-tuning after 5<sup>th</sup> epoch to be my baseline. Since I don't want to use validation dataset (to make model recognize more pattern) so I upload the 4<sup>th</sup> epoch to see whether my model is overfitting and the result shows that the it's overfitting after 5<sup>th</sup> epoch. With the same concept, I continue upload the 3<sup>rd</sup> epoch and find out that the accuracy is decrease.

After knowing RoBERTa-base model fine-tuning after 4<sup>th</sup> epoch maybe one of the best settings, I start to try if BERT-large-uncased model will be better. And I find out that it will be worse than my 3<sup>rd</sup> attempts, so I decide to upload RoBERTa-base model fine-tuning after 2<sup>nd</sup> epoch betting it will perform well in testing set.

Finally, the result in private stage shows that my third attempt is the best model, though the performance is worse than second attempt in public stage.