Sentiment Analysis on LGU Kayuan

Yingyao Liu, Ying Xue, Yunfei Ke, and Shijie Shao
School of Data Science
The Chinese University of Hong Kong, Shenzhen
{120090628, 120090452, 120090277, 120090491}@link.cuhk.edu.cn

Abstract

Aiming to improve the mental health of CUHKSZ students, we trained a sentiment analysis model based on LGU Kayuan, a popular anonymous social website targeting at CUHKSZ students. Based on 15000 texts scratched from Kayuan, we trained a BERT-based model with several techniques solving data imbalance, and finally our model achieved an accuracy of 85.33% and an F1 score of 74.37% on the test set. In the future, we will collaborate with student counselling departments to monitor students' emotional levels and provide them with timely mental service support according to the predictions of our model.

1 Motivation

In the recent two years, affected by the pandemic and increasingly intense peer pressure, the mental health state of CUHKSZ students is becoming more and more concerning. Aiming for an efficient approach to monitor students' mental health states, we performed sentiment analysis on LGU Kayuan, an anonymous social website targeted at CUHKSZ students. Due to its anonymity, more and more students prefer to express their real feelings on it, which made it the best public platform to observe CUHKSZ students' emotional levels. Therefore, we managed to train a model based on LGU Kayuan corpus to reveal the underlying emotion under every post on Kayuan.

2 Data Preparation

2.1 Data Collection

Based on Python web scraping and the software Charles, we extract 15,000 posts from LGU Kayuan from March 17th to April 7th and stored the data into JSON files.

Figure 1: LGU Kayuan

Figure 2: Posts obtained from LGU Kayuan in JSON file

2.2 Data Preprocess

To balance the text length, we cut off fractions of texts exceeding 128 words. Then, we manually labeled each post according to their underlying emotions into four categories: neutral, positive, upset, and angry. With examples shown in Figure 4, the detailed logistics of labeling are as follows:

- **Neutral**: Texts without obvious emotions are labeled as "neutral", including suggestion-seeking posts, opinion expressions, etc.
- **Positive**: It refers to posts conveying appreciation and gratitude towards others, optimism, or other positive emotions.
- **Upset**: To be upset is to be disturbed or very unhappy. In Kayuan, these posts are usually correlated with academic pressure and interpersonal relationships.
- **Angry**: Compared with upset, it is an intense emotion against unsatisfactory outer space. In Kayuan, students usually express angry emotions towards some courses or peers.

2.3 Data Analysis

As shown in Figure 3, neutral posts (77%) occupy the largest proportion of Kayuan corpus, which leads to data imbalance. We managed to deal with this unbalance issue in the follow up sections.

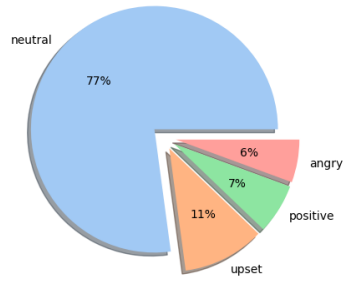

Figure 3: Percentage of emotions

	B	C	D	E	F	G	H	I	J	K	L
1	icknam	reated	ypeNam	content	knonym	dlikeCount	gender	grade	nstitutil	lowPo	Emotion
949		2023-0	Commen	哥大MFE招生负责人原话中国金融机构只有中金在录取的时候被视为bi	TRUE	6	MALE	2021	经管学院	UG9zdD	中性
950		2023-0	Commen	蹲	TRUE	0	MALE	2018	经管学院	UG9zdD	中性
951	中园的	2023-0	Commen	蹲	FALSE	0	MALE	2022	经管学院	UG9zdD	中性
952		2023-0	Commen	作业和quiz基本都是人均满分，就看mid和final	TRUE	0	MALE	2021	经管学院	UG9zdD	中性
953	麦记在	2023-0	Post	日推背春是桑葚色	FALSE	5	MALE	2019	理工学院	UG9zdD	中性
954	DreamD	2023-0	Commen	诱奖FIN2020和FIN没有任何关系，和凸优化有关系[愤慨]	FALSE	0	MALE	2021	经管学院	UG9zdD	中性
955	爱辣螺	2023-0	Commen	音乐吧	FALSE	7	MALE	2019	数据科学	UG9zdD	中性
956		2023-0	Commen	Prof李丹不过好像偏量化…不太记得了	TRUE	1	MALE	2022	经管学院	UG9zdD	中性
957	! ##	2023-0	Commen	远离中宏，人人有责	FALSE	1	MALE	2019	经管学院	UG9zdD	愤怒
958	丝滑德	2023-0	Commen	2111会做题就行了，考试和作业差不多，蛮简单的一门课	FALSE	1	MALE	2020	数据科学	UG9zdD	中性
959	开心到	2023-0	Post	run1	FALSE	2	MALE	2022	经管学院	UG9zdD	中性
960		2023-0	Commen	我	TRUE	0	MALE	2022	经管学院	UG9zdD	中性
961	大橘为	2023-0	Commen	没有吧，这俩课属于那种只要肯下时间是肯定能学明白的，而且期中期	FALSE	1	MALE	2019	理工学院	UG9zdD	中性
962	黑色苦	2023-0	Post	找个代写，明天晚上数学的ddl，帮我把这一周份答案抄一遍，有偿	FALSE	0	MALE	2022	医学院	UG9zdD	中性
963		2023-0	Commen	感觉平均分不会低捏[哭泣]	TRUE	1	MALE	0		UG9zdD	低落
964		2023-0	Commen	Dear applicant,Thank you for your email.Applicants aged 15 c	FALSE	0	MALE	2022	经管学院	UG9zdD	中性
965	紫米奶	2023-0	Commen	蹲一个	FALSE	0	FEMALE	2021	经管学院	UG9zdD	中性
966		2023-0	Commen	bio1001	TRUE	0	FEMALE	2020	医学院	UG9zdD	中性
967		2023-0	Post	刚刚坐公交右手一抓扶手我的妈呀蟲的跟什么似的好恶心啊能不能公共	TRUE	3	FEMALE	2022	经管学院	UG9zdD	愤怒
968		2023-0	Commen	我记得老师最后说了一句明天没有导修	TRUE	0	FEMALE	2022	医学院	UG9zdD	中性
969		2023-0	Post	问问大家申研是只看前三年绩点吗大三升大四暑课和大四上的看不看	TRUE	0	MALE	2020	数据科学	UG9zdD	中性
970		2023-0	Commen	康宁	TRUE	0	FEMALE	2022	数据科学	UG9zdD	中性
971	蓝色白	2023-0	Post	midterm结束啦，两门通识的quiz和pre也都结束啦，暂时没有ddl，我	FALSE	15	FEMALE	2021	经管学院	UG9zdD	积极

Figure 4: Examples of posts with different emotions

3 Model

To achieve a good performance, we chose BERT (Bidirectional Encoder Representations from Transformers) [2], a popular model for NLP tasks nowadays. BERT is a pre-trained model that can be fine-tuned for a wide range of NLP tasks, including text classification, named entity recognition, and question answering. Moreover, Bert uses the Transformer architecture. According to previous studies, it has achieved state-of-the-art results on a wide range of NLP benchmarks.

3.1 Training Process

After data processing, the input text is first tokenized into subwords and then mapped to embedding vectors. A classification layer is added on top of the pre-trained BERT model. Then on top of the BERT model, there is a classification layer. The output of the final encoder layer is fed into the classification layer and makes a prediction on the sentiment label, which is shown in Figure 6.

3.2 Architecture of Bert

The transformer architecture of BERT greatly contributes to its power. Bert consists of a series of encoder layers, each consisting of two sub-layers: a self-attention layer and a feed-forward neural network layer.

As shown in Figure 5 the self-attention layer is the key component of the Transformer model. It allows the model to capture dependencies between different words in a sentence by computing attention weights for each word. These attention weights indicate how important each word is for predicting the output of the model.

The feed-forward neural network layer applies a non-linear transformation to the output of the self-attention layer. This allows the model to learn complex non-linear relationships between words.

BERT has two parameter intensive settings [4]:

- **BERT-base**: 12 layers, 768 hidden dimensions and 12 attention heads (in transformer) with the total number of parameters, 110M;
- **BERT-large**: 24 layers, 1024 hidden dimensions and 16 attention heads (in transformer) with the total number of parameters, 340M

We chose a popular BERT model for Chinese corpus, RoBERTa-wwm-ext-large (Chinese), which is pre-trained with the whole word masking (WWM) task in Chinese [1].

	Masking	Type	Data Source	# Tokens	# Steps
BERT (Google)	WordPiece	BERT-base	wiki	0.4B	?
RoBERTa-wwm-ext-large	WWM	BERT-large	wiki+ext	5.4B	$2M^{MAX512}$

Table 1: Comparison of BERT (Google) and RoBERTa-wwm-ext-large

It is a large bert model which contains 24 encoder layers and is connected with 3 fully connected layers for classification. The output of our model is a 4-d vector representing the possibility of 4 emotions: positive, neutral, angry and upset.

3.3 Techniques for solving unbalanced data

As shown in Figure 3, the labels of our training sample is unbalanced. Therefore, it is vital to observe the F1 score of the predicted labels, which takes precision and recall into account:

$$F1 = \frac{2 * (precision * recall)}{precision + recall}$$

In our chosen base models, we found that some of them have high accuracy, but much lower F1 scores(e.g., word2vec+KNN); In order to solve this issue, we used the following two techniques:

- **Resample**: we repeat sampling from a given sample to augment data with rare labels. After our attempts, we find that by resampling texts with labels "positive" and "angry" to 1800, and texts labeled as "upset" to 2500, the model performance improves.
- **Focal Loss** [3]: we used focal loss ($\gamma = 2, \alpha_t = 1$ for all t) to substitute cross entropy loss as follows:

  $$FL(p_t) = -\alpha_t * (1 - p_t)^{\gamma} * log(p_t)$$

  , where p_t represents the probability that the model predicts the true label t, α_t is the weight of class t, and γ is a tunable parameter used to adjust the ratio of difficult and easy samples. When $\gamma = 0$, Focal Loss is equivalent to cross entropy loss; When γ gets larger, Focal Loss pays more attention to the rare labels.

4 Experiments and Discussion

4.1 Model performance

As shown in Table 2, we choose 2 baseline methods: word-to-vector plus KNN, and Chinese Bert finetuned by the Weibo dataset. For the former, it is a classical machine learning method for sentiment analysis. For the latter one, Chinese Bert is an established Chinese sentiment analysis pre-trained model. While Weibo, the most popular social media in China, is usually used for sentiment analysis. Meanwhile, the data of Weibo is similar to data of LGU Kayuan; both consist of posts and comments and are relative to our daily life.

Method	Training Data	Val		Test	
		Acc	F1	Acc	F1
word2vec+KNN	LGU Kayuan	0.7718	0.4021	0.7561	0.3882
Chinese Bert	Weibo[1]	0.1593	0.2275	0.1615	0.2387
Chinese Bert without techniques	LGU Kayuan	0.8123	0.6529	0.8223	0.6911
Chinese Bert(Ours)	LGU Kayuan	**0.8472**	**0.7745**	**0.8533**	**0.7437**

Table 2: Model Performance on LGU Kayuan

However, from Table 2, we see that word-to-vector plus KNN shows low F1 score, and the model finetuned by Weibo data has both low accuracy and F1 score. Therefore, it is necessary to train a new model targeting Kayuan. We tried several techniques to improve the performance, which is shown in Table 3. The result of our final model is much more accurate than the other two.

ID	Weibo Corpus	Resampling	Weighted Cross Entropy Loss	Focal Loss	Acc	F1
i					0.8223	0.6911
ii	✓				0.7729	0.6466
iii		✓			0.8467	0.7115
iv			✓		0.8127	0.6812
v				✓	0.8434	0.7242
vi		✓		✓	**0.8533**	**0.7437**

Table 3: Chinese Bert ablation experiment

4.2 Discussion

Table 3 shows the performance of the Chinese Bert model under different data and loss functions, using accuracy (Acc) and F1 score as evaluation metrics.

(i) The default setting (LGU Kayuan corpus & cross entropy loss) gives 0.8223 accuracy and 0.6911 F1 score

(ii) Compared to not using Weibo corpus, the accuracy drops by 0.0494, and the F1 score drops by 0.0445. This may be because the Weibo corpus has a large difference from the target dataset, leading to the model's inability to adapt to the target domain.

(iii) Using the resampling method will improve the performance of the model. Compared to not using resampling, the accuracy increases by 0.0244, and the F1 score increases by 0.0204. This may be because the resampling method can balance the distribution of different categories in the dataset, reducing the impact of class imbalance.

(iv) Using the weighted cross entropy loss function will reduce the performance of the model. Compared to not using the weighted cross entropy loss function, the accuracy drops by 0.0096, and the F1 score drops by 0.0099. This may be because the weighted cross entropy loss function gives too much penalty to misclassified samples, leading to overfitting of the model.

(v) Using the focal loss function will improve the performance of the model. Compared to not using the focal loss function, the accuracy increases by 0.0211, and the F1 score increases by 0.0331. This is because the focal loss function can transfer model's attention from easy-to-classify samples to hard-to-classify samples, thereby improving the predictive strength across all classes.

(vi) Using resampling and focal loss function together can achieve the best effect, with an accuracy of 0.8533 and an F1 score of 0.7437. This shows that these two methods can complement each other and jointly solve the problems of class imbalance and the difference between easy and hard samples in the dataset.

5 Conclusion

Although limited by insufficient data due to the time-consuming labeling work, we managed to train a well-performing BERT-based model which can accurately detect the emotion under every Kayuan post. We are planning to share our contributions with student counseling departments so that we can monitor students' mental state and provide them with timely support accordingly.

[1]Weibo (a wideused social media) corpus is attained from `https://smp2020ewect.github.io/`

References

[1] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, November 2020. Association for Computational Linguistics.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[4] Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*, 2019.

A Appendix

Here are some figures about the detailed architecture of Bert model:

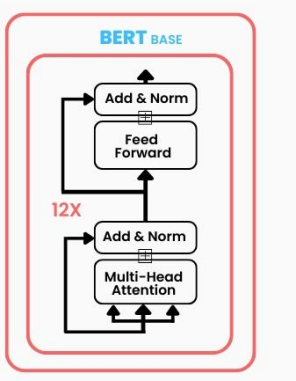

Figure 5: Attention layer

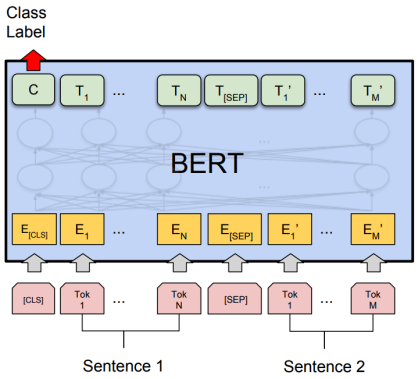

Figure 6: BERT for sentiment analysis