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)

April 2023

Outline

Table of Contents

What is Kayuan

LGU Kayuan is an **anonymous** social website aiming for CUHKSZ students. It became popular soon after its publication at early March. Students mainly post contents w.r.t. campus life, academic development, emotions, and peer supports via Kayuan.

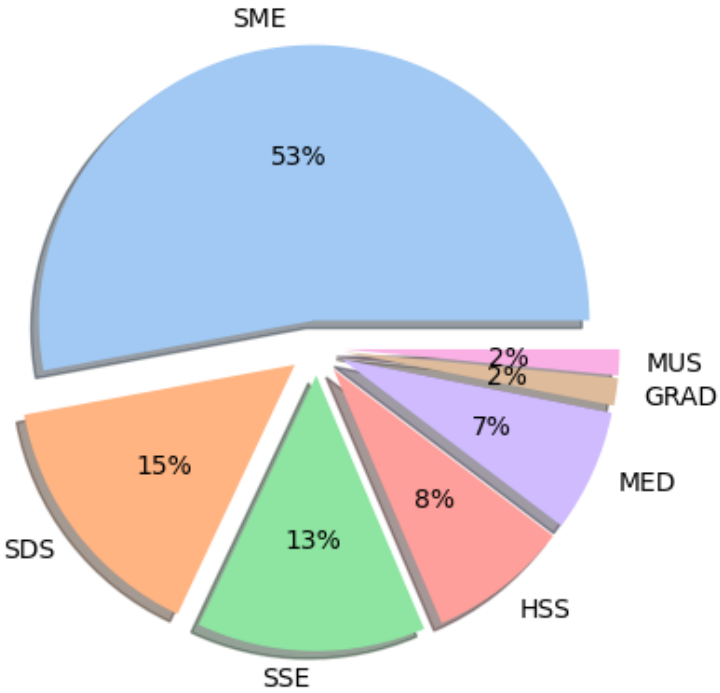


Figure: School Distribution of Kayuan Users

Figure: sample screenshot of Kayuan

Why do we pay attention to LGU Kayuan

- **Background:** Increasingly serious mental health issue among CUHKSZ students.
- **Strength1:** Growing popularity of Kayuan among CUHKSZ students.
- **Strength2:** Due to anonymity, students tend to express their real thoughts on Kayuan.
- **Goal:** Monitor emotions of CUHKSZ students, and further imply their overall mental health status.
- **Application:** Collaborate with responsible departments such as Student Counselling, provide students with timely mental health support.

Table of Contents

Data Processing

Data Scratching: We scratched over 15000 Kayuan posts from March 17th to April 7th.
Text Length Editing: We cut off fractions of texts exeeding 128 words.

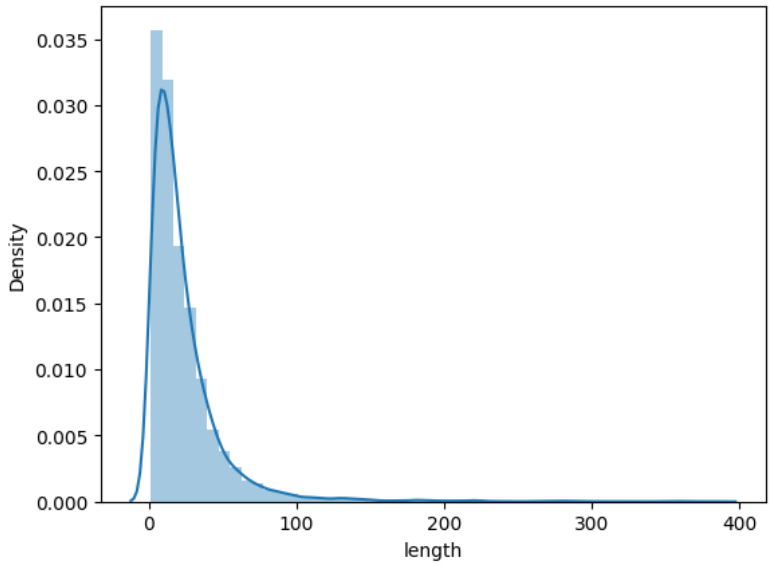


Figure: Length Distribution of Posts

Sentiment Labeling

We classified the contents into **positive**, **negative** and **neutral**. Since we pay more attention to negative emotions, we further divide negative emotions into **upset** and **angry**.

Sentiment Labeling—Neutral

Suggestion Seeking

准备去成都转转，大家推荐点好玩的好吃的地方吧

translation: Planning to journey around Chengdu, please recommend some interesting places.

Making Friends

有无喜欢新世纪纯音乐的uu 看看有无共同兴趣爱好的uu互相分享

translation: Anyone interested in new age music? Wanna find some friends to share!

Opinion Expression

平权主义者，问题出在个体不是性别，每个性别都有很好的人或者人渣

translation: As a egalitarian, I believe we should focus on individuals rather than gender.

Sentiment Labeling—Positive

Gratitude

早上十点二十在道扬后面帮我扶车的男生！感谢你！我是社恐不好意思大声说但还是谢谢你！

translation: To the boy who helped me with the bicycle, thank you so much!

Appreciation

今晚在新图三楼遇到了一个红发姐姐，好米好米[心动][心动][心动]

translation: Met a red hair lady tonight at the library, adorable!!

Optimism

放轻松啦！能来这里很优秀了不要给自己太大的压力！感觉自己能学到东西就很棒了

translation: Take it easy! You are excellent enough to be here. It's good enough if you can learn new things.

Sentiment Labeling—Upset

To be upset is to be disturbed or very unhappy.

Academic Pressure

真的好累啊 可不可以不卷了

translation: Really exhausted, can we stop competing against each other

Interpersonal Relationships

我也，失去喜欢别人的能力了，感觉挺可悲的

translation: I lost the ability to fall in love too, pathetic

Sentiment Labeling—Angry

Angry refers to an intense emotion, including disgust or discontent, against the outer world.

Academic Pressure

为什么这么多课都要做pre啊！ 做你妈的pre!

translation: Why do so many courses require presentations! Fxxk presentations!

Peer Conflicts

逆天舍友天天一点钟吃泡面弄的宿舍一股味[愤怒][愤怒][愤怒]

translation: my annoying roommate eats noodles everyday at midnight and makes the dormitory so smelly[emoji][emoji][emoji]

Table of Contents

Model for Sentiment Analysis

BERT, or Bidirectional Encoder Representations from Transformers, is a deep learning model for natural language processing (NLP) developed by Google in 2018. It 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**.

BERT uses the Transformer architecture, which allows it to capture long-range dependencies and learn contextual representations of language. It has achieved **state-of-the-art** results on a wide range of NLP benchmarks.

Implementations

Masked language modeling is one of the two pre-training objectives used in the BERT model. In this task, a certain percentage of the tokens in a sentence are randomly replaced with a special token "[MASK]".
The objective of the model is to predict the original tokens.

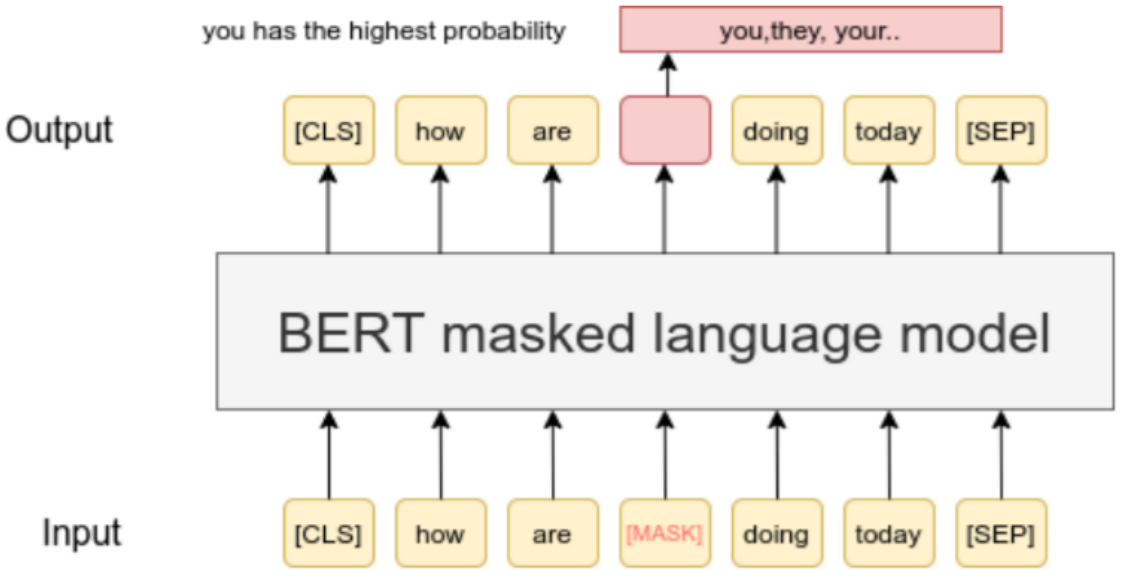


Figure: Bert Model for Mask LM

Question Answering

Bert model can also be implemented for **question answering**.

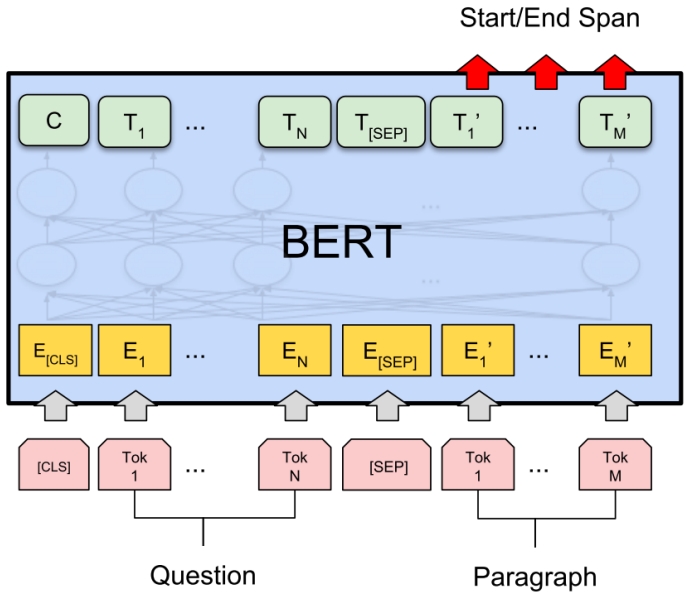


Figure: Bert Model for Question Answering[1]

The goal of the task is understanding the question and paragraph and then answer the question.

[1] https://github.com/vgaraujov/Question-Answering-Tutorial/blob/master/Question$_A$nswering$_B$ERT$_S$panish.ipynb

Text Classification

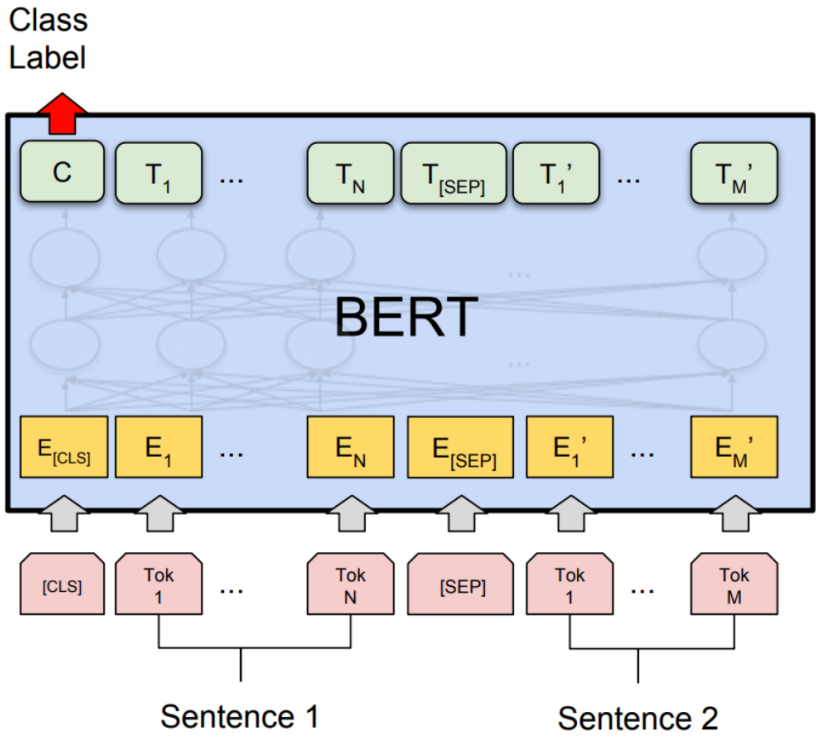
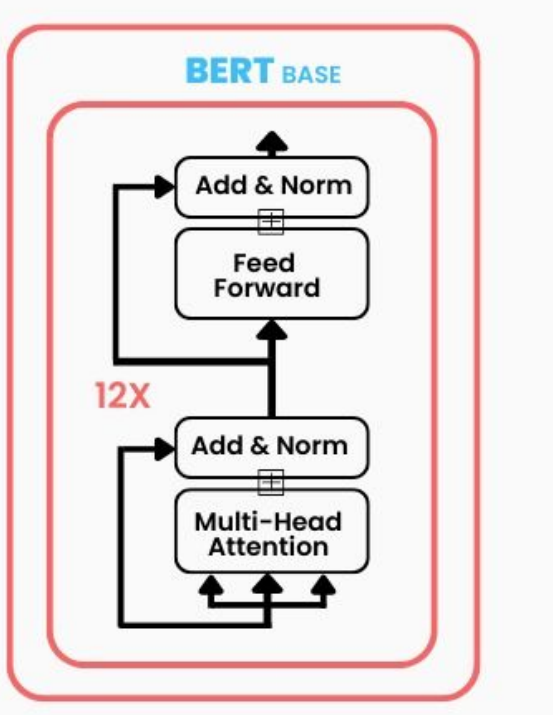


Figure: Bert Model for sentiment analysis

Once the BERT model is pre-trained, it can be fine-tuned for a specific downstream task, such as sentiment analysis. This involves adding a task-specific layer on top of the pre-trained BERT model and fine-tuning the entire network on labeled data for the specific task.

Architectures



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.

The feed-forward neural network layer takes the output of the self-attention layer and applies a non-linear transformation to it.

Our Bert Model

In our projects, the model we use is a **large** Bert model which was pre-trained with Chinese text before. This Bert model contains **24 encoder layers** and is connected with a **3 fully-connected layers**. The output of the model is the class label containing 4 categories: positive, neutral, upset, and angry.

Problems in Kayuan Corpus

- The main problem we faced in the Kayuan corpus is the **unbalanced labels** of emotions.
- **F1 score**:

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

, which is the harmonic mean of precision and recall.

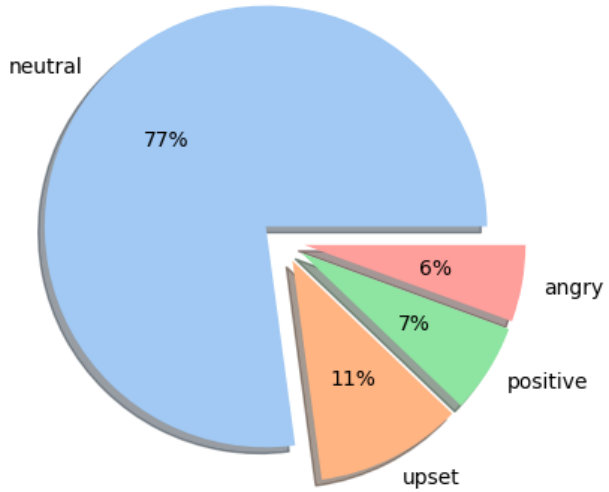


Figure: Emotion Distribution

Techniques for Solving Unbalanced Data

- **Resample**: repeat sampling from a given sample. After our attempts, we find that resample to 700 for each label is a good choice.
- **Focal Loss**[2]:

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

, 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

[2] https://arxiv.org/pdf/1708.02002.pdf

Table of Contents

Baseline Model Performance

Method	Training Data	Val		Test	
		Acc	F1	Acc	F1
word2vec+KNN	LGU Kayuan	0.7718	0.4021	0.7561	0.3882
Chinese Bert	Weibo[3]	0.1593	0.2275	0.1615	0.2387

Table: Model Performance on LGU Kayuan

- The first baseline model shows the importance of chosen model.
- The second baseline model shows the importance of training data.
- It is necessary to train a new model targeting Kayuan.

[3]Similar with LGU Kayuan, both are social media containing posts and comments. https://smp2020ewect.github.io/

More Experiments

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: Chinese Bert experiment[4]

- Add Weibo Corpus (ii): Data augmentation. (Training data is a mix of Weibo and Kayuan.) $\Rightarrow$ Lower Acc and F1.

[4] Default loss: Cross Entropy

More Experiments

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: Chinese Bert experiment[5]

- Resample (iii): Try to solve data imbalance.
  $\Rightarrow$ Categories with smaller percentages have higher accuracy.

[5] Default loss: Cross Entropy

More Experiments

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: Chinese Bert experiment[6]

- Weighted Cross Entropy Loss (iv): Try to solve data imbalance.
  $\Rightarrow$ No improvement.

[6] Default loss: Cross Entropy

More Experiments

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: Chinese Bert experiment[7]

- Focal Loss (v): Try to solve data imbalance.
  $\Rightarrow$ Best solution to data imbalance.

[7] Default loss: Cross Entropy

More Experiments

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: Chinese Bert experiment[8]

- Resample + Focal Loss (vi) $\Rightarrow$ Highest Acc and F1 Score.

[8] Default loss: Cross Entropy

Overall Model Performance

Method	Training Data	Val		Test	
		Acc	F1	Acc	F1
word2vec+KNN	LGU Kayuan	0.7718	0.4021	0.7561	0.3882
Chinese Bert	Weibo[9]	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: Model Performance on LGU Kayuan

[9]Similar with LGU Kayuan, both are social media containing posts and comments. https://smp2020ewect.github.io/

Table of Contents

Contributions, Limitations, and Applications

Contributions:

- Create a sentiment analysis **dataset** for LGU Kayuan.
- Train **a well-performing model** for LGU Kayuan sentiment analysis.

Limitations:

- Insufficient data (Because of the time-consuming data labeling work).

Applications:

- Customize an **online** emotion detection system for LGU students, report the result to the corresponding department on time.