Chapter4_TianFengshou_EN.docx

by Fengshou TIAN

CHAPTER 4

Initial Findings

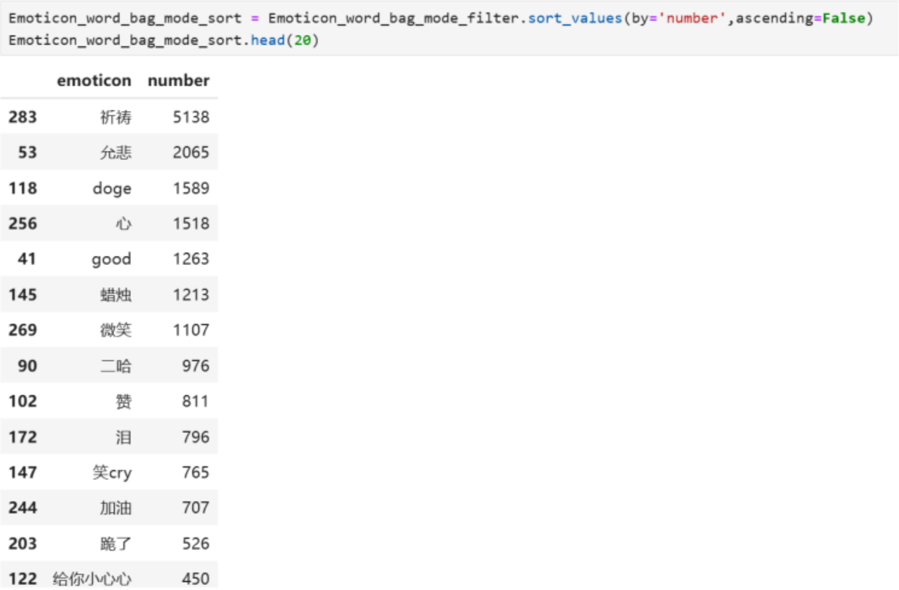
4.1 Data Exploration

After the data is merged, if the text needs to be processed, two types of models are needed. Traditional statistical models need to remove meaningless words, so for this type of model, the emojis in the data need to be removed. For deep learning models, the original data is directly retained, so for deep learning models, the original data also has a practical role.

At the same time, due to the large amount of data obtained, only one independent event, the data related to the "Yibin Earthquake", was selected for subsequent operations. After screening according to the subject and keywords, 55,570 pieces of data related to the event were actually obtained.

4.1.1 Statistical analysis of emoticons

First, we use the emoji discovery system we built earlier to obtain all the emojis used in the dataset, and then we perform statistical analysis on the emojis we found. We find that the most commonly used emoji is "祈祷" (which means "prayer" in English), and the top 20 usage frequencies are as follows:



```
Emoticon_word_bag_mode_sort = Emoticon_word_bag_mode_filter.sort_values(by='number',ascending=False)
Emoticon_word_bag_mode_sort.head(20)
```

	emoticon	number
283	祈祷	5138
53	允悲	2065
118	doge	1589
256	心	1518
41	good	1263
145	蜡烛	1213
269	微笑	1107
90	二哈	976
102	赞	811
172	泪	796
147	笑cry	765
244	加油	707
203	跪了	526
122	给你小心心	450

Figure 4.1 Number of emoticons used

Using a bar chart to visualize the data, the results are as follows:

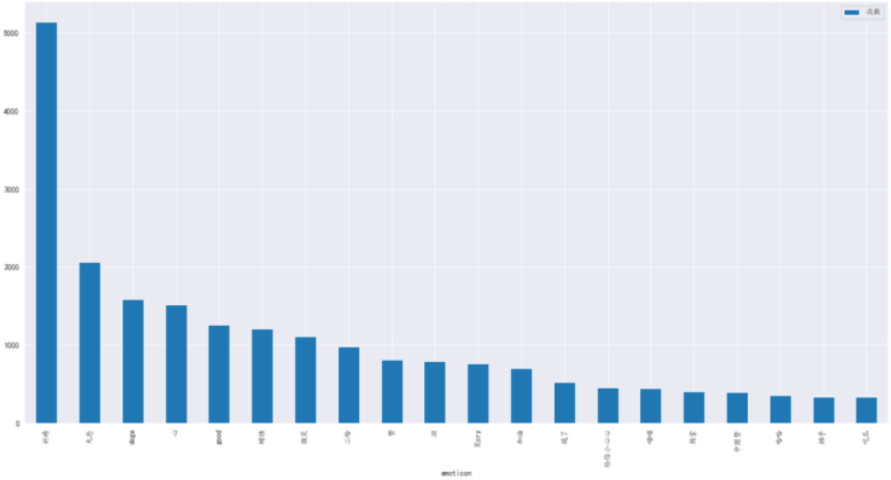


Figure 4.2 Histogram of the number of emoticon packages used

After analyzing the data set, the user ID that used the most emojis was 3853279141, who used the most emojis in this offline analysis, with 52 emojis. A bar chart was used to plot the number of emoji usages, and the results are shown in the figure:

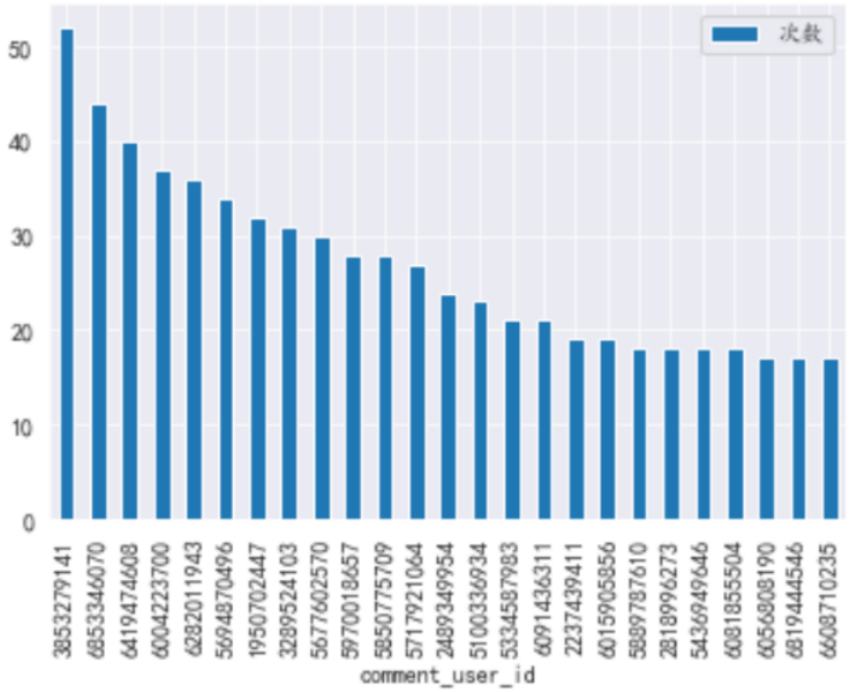


Figure 4.3 Statistics on the total number of emoticons used by users

By counting the number of posts by the id, we can calculate the frequency of each user using the emoji package. We found that the user who used the emoji package the most was user 5694870496.

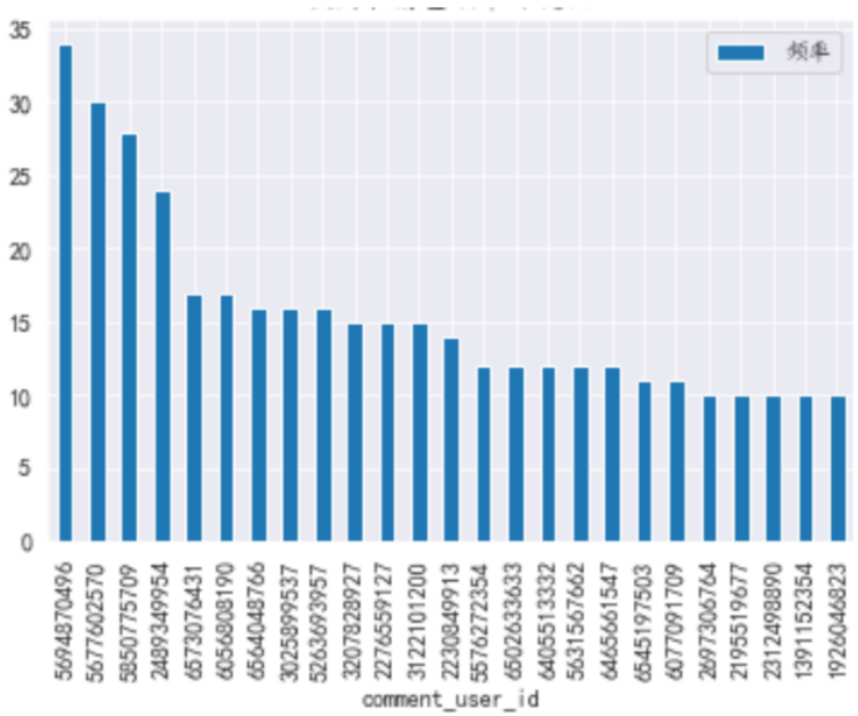


Figure 4.4 Statistics on the frequency of users' emoji usage

4.2 Word segmentation model

The word segmentation model directly uses the structured perceptron model in HanLP, which is trained from a large comprehensive corpus of 99.7 million words, covering multiple fields such as news, social media, finance, and law. It is the largest Chinese word segmentation corpus in the world. The size of the corpus determines the actual effect. The corpus for production environments should be in the tens of millions of words. Natural semantic linguistics experts have been continuously annotating the corpus to keep up with the times and maintain the most advanced word segmentation quality. In terms of word segmentation standards, HanLP provides two granularities: fine-grained and coarse-grained. Fine-grained is suitable for search engine business, and coarse-grained is suitable for text mining business. This study uses coarse-grained word segmentation according to actual needs.

4.3 Word cloud

The data set after word segmentation is clean content words. By statistically analyzing the data, a word cloud can be obtained. By specifying the shape of the word cloud as a heart shape and specifying the color, a word cloud that fits this earthquake event can be obtained.

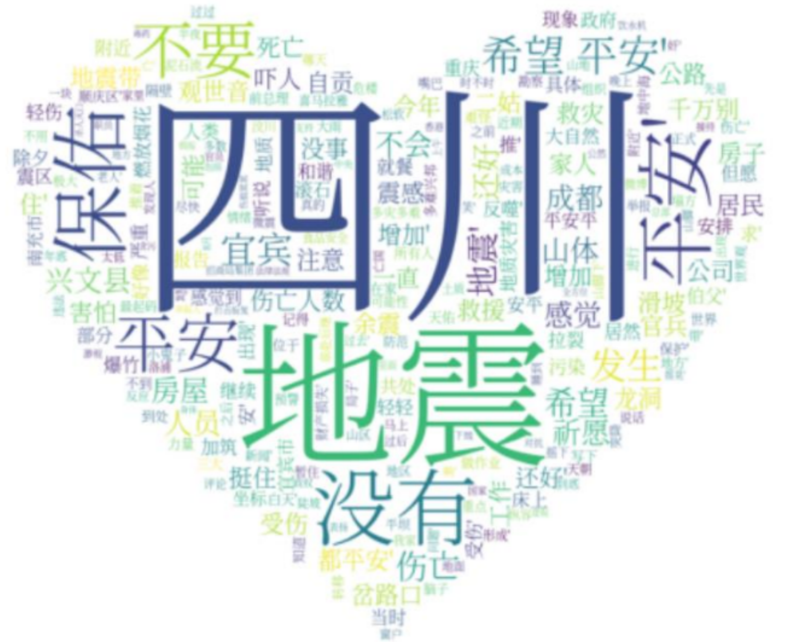


Figure 4.5 Colorful love word cloud

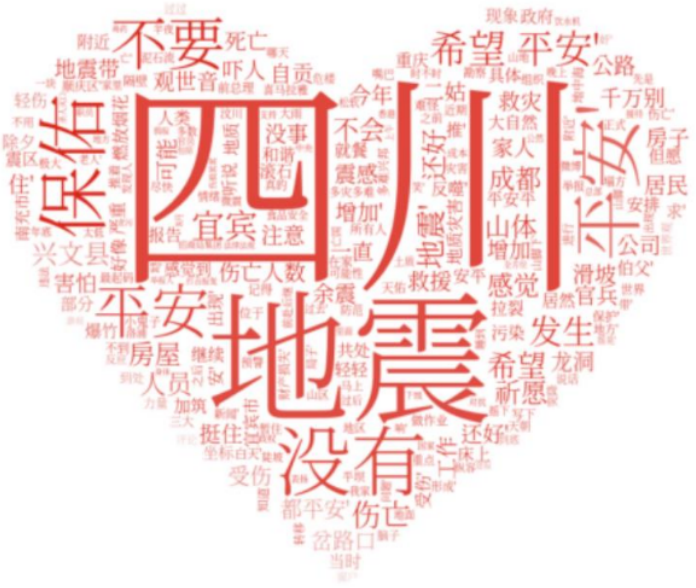
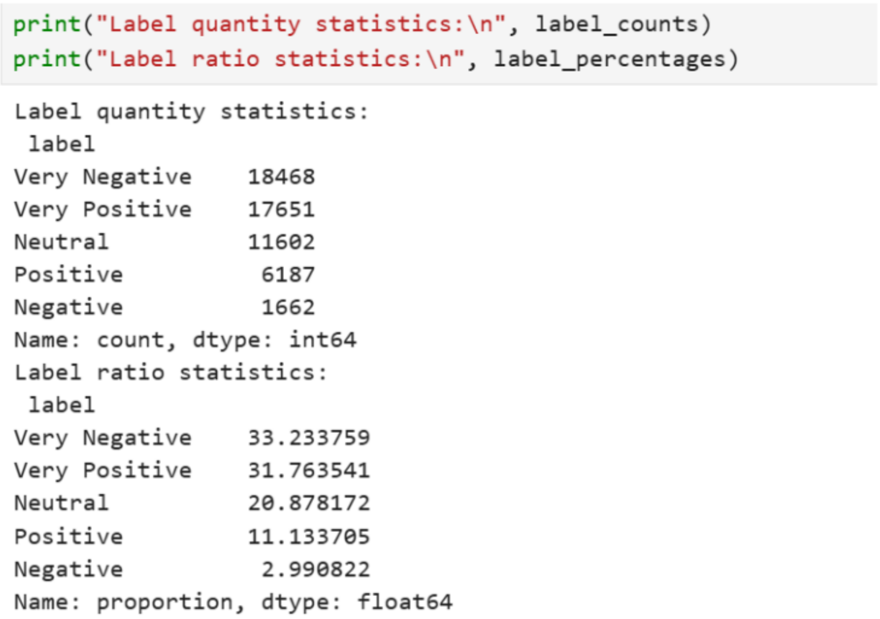


Figure 4.6 Red love word cloud

From the statistical analysis, we can see that the most used word is "Sichuan", which is the place where the earthquake occurred, followed by "earthquake", which means earthquake. The words that are used more frequently are "bless (Baoyou)", "hope (Xiaowei)", "peace (Ping'an)", etc. It can be basically seen that the content and comment areas of various social media platforms are mainly blessings for this incident.

4.4 Sentiment Analysis

Due to the particularity of the selected events, the data inherently contains both negative and positive comments, and different algorithms can be used for sentiment analysis. DistilBERT trained on the SST-2 dataset was first added with synthetic multilingual data generated by advanced LLMs, and then fine-tuned. The final test dataset accuracy was about 93%. The prediction result of the model is label + score. The model was used to perform statistical analysis on the collected dataset, and the final prediction results are as follows:



```
print("Label quantity statistics:\n", label_counts)
print("Label ratio statistics:\n", label_percentages)
```

```
Label quantity statistics:
 label
Very Negative    18468
Very Positive    17651
Neutral          11602
Positive          6187
Negative          1662
Name: count, dtype: int64
Label ratio statistics:
 label
Very Negative    33.233759
Very Positive    31.763541
Neutral          20.878172
Positive         11.133705
Negative          2.990822
Name: proportion, dtype: float64
```

Figure 4.7 Tag statistics results

Data visualization technology can also be used to well display the distribution of sentiment analysis results.

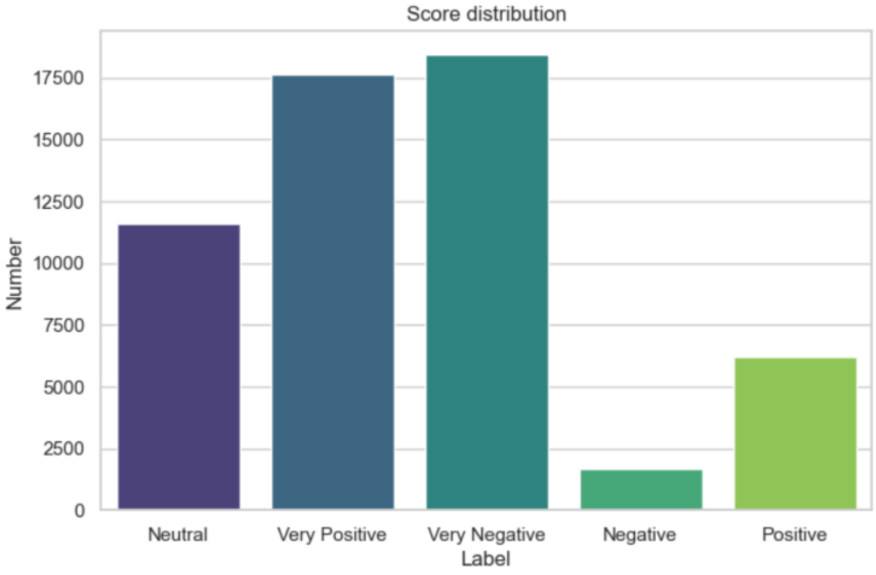


Figure 4.8 Score distribution histogram

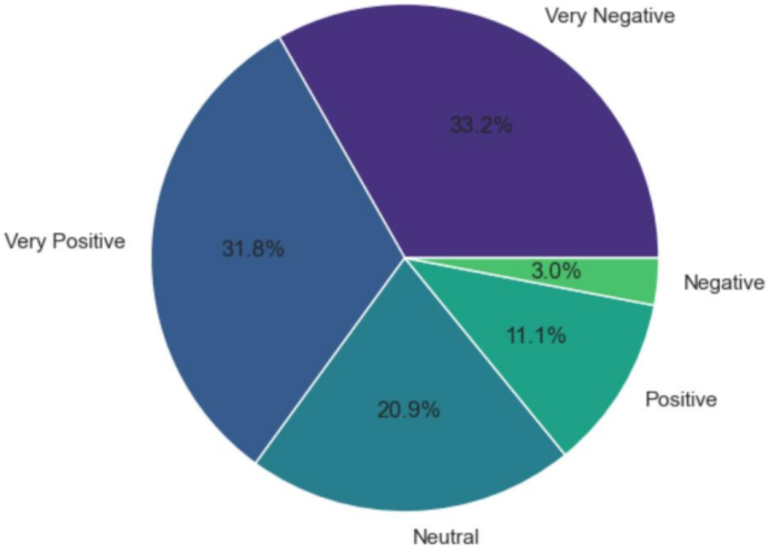


Figure 4.9 Label ratio pie chart

4.5 Summarize

Based on the emoji discovery system, [2] we conducted statistical analysis of emojis, analyzed the frequency of users using emojis, and preliminarily implemented sentiment analysis of cross-platform data of a single event [2] based on the sentiment analysis model. The entire process can basically achieve the design goal.

REFERENCES

1. Bai, X., Chen, Y., & Zhang, Y. (2022). Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 6001–6015). Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.415
2. Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Departmental Papers (CIS)*.
3. He, H., & Choi, J. (2021). The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 5555–5577). Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.451
4. Wang, C., & Xu, B. (2017). Convolutional neural network with word embeddings for Chinese word segmentation. In *Proceedings of the Eighth International Joint Conference on Natural Language

Processing* (Vol. 1, pp. 163–172). Asian Federation of Natural Language Processing. https://www.aclweb.org/anthology/I17-1017

5. Li, B., Yuan, Y., Lu, J., Feng, M., Xu, C., et al. (2022). The first international Ancient Chinese word segmentation and POS tagging bakeoff: Overview of the EvaHan 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages* (pp. 135–140). European Language Resources Association. https://aclanthology.org/2022.lt4hala-1.19/

Chapter4_TianFengshou_EN.docx

ORIGINALITY REPORT

5%	0%	4%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Xuejing Chen, Luyuan Xie, Yonghong He, Tian Guan, Xuesi Zhou, Bei Wang, Guangxia Feng, Haihong Yu, Yanhong Ji. "Fast and accurate decoding of Raman spectra-encoded suspension arrays using deep learning", The Analyst, 2019 Publication	2%
2	"Chinese Lexical Semantics", Springer Science and Business Media LLC, 2024 Publication	2%
3	Submitted to University of Exeter Student Paper	2%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		