CHAPTER 3

Research Methodology & Design

3.1 Introduction

This section introduces the methods and applications that will be used in the study. The overall framework is formulated according to the ACM/IEEE data science life cycle model. First, the project's goals and key issues are clarified. After that, data is collected. This project will write a crawler to obtain data from three different social media and perform data preprocessing. In the exploratory data analysis stage, Chinese word segmentation and new word discovery models will be used in combination with statistical methods and visualization technology to conduct comprehensive text statistical analysis. Finally, this study will use the transform model to perform sentiment analysis on multi-platform network social media text data.

3.2 The framework

3.2.1 Data Science Life Cycle

This method adopts a structured data science project life cycle mechanism to promote the reasonable and efficient completion of data collection, analysis and model creation. There are many definitions of the life cycle of a data science project.

This study adopts the ACM/IEEE data science life cycle model, in which the seven key stages are: problem identification, data collection, data preparation, data analysis, modeling, and evaluation and deployment.

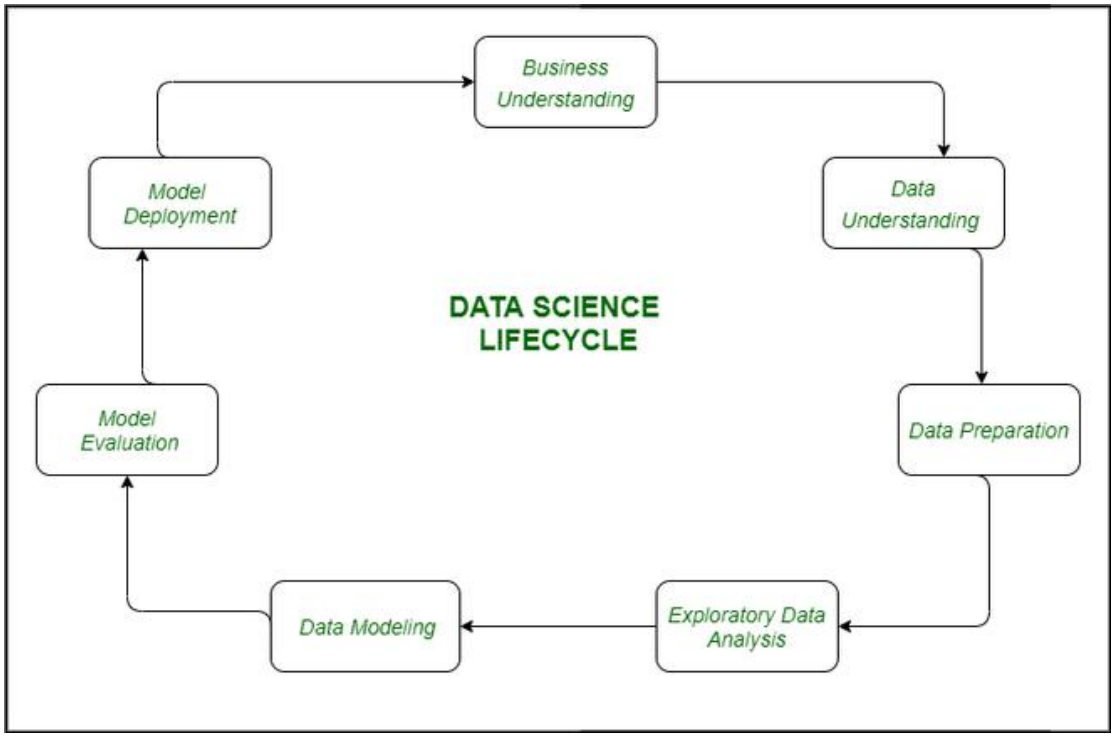


Figure 3.1 Data Science Life Cycle

(a) Problem Identification

The definition actually refers to clarifying the goals and key issues of the project. This phase involves communication with stakeholders to understand needs and expectations. The core goal of this study is to perform sentiment analysis on text data of the same event on multiple social media. This goal can be broken down into two parts: data acquisition of the same event on different social media, and sentiment analysis of short online text data.

(b) Data Collection

The definition actually means collecting the required data, including internal data (such as company databases) and external data (such as public data sets). At this stage, pay attention to the source of the data and how to obtain it. The data collected in this study are the content posted by users of Weibo, Xiaohongshu, and Bilibili, as well as the comments and replies to the related content. All of the above data are public data from social media, but there is no data source directly used for research, so crawler technology is needed to obtain data. The main implementation of this study is implemented in Python, and the main technical tools are the requests library and the Scrapy crawler framework. And use CSV files and MongoDB database for storage.

(c) Data preparation

The definition actually means cleaning and transforming the collected data, including processing missing values, removing outliers, and converting data formats. This is an important step to ensure data quality. The data obtained in this study is network text data. The original data contains special tags and hyperlinks, which need to be processed. At the same time, since the data comes from multiple platforms, the data format needs to be unified. The specific tool is Pandas in Python.

(d) Exploratory Data Analysis (EDA)

Conduct exploratory data analysis (EDA) and use statistical methods and visualization tools to understand the basic characteristics and structure of the data. This helps identify potential patterns and relationships. Chinese text analysis projects need to use models for new word discovery, Chinese word segmentation, and stop removal at this stage. Then perform data analysis based on specific analysis goals. The visualization tools used during this period are Echarts and Matplotlib.

(e) Data Modelling

Modeling means selecting and applying appropriate algorithms to build models, such as regression, classification, and clustering. At this stage, model evaluation and adjustments are also performed to optimize performance. The model finally established in this study is a sentiment analysis model. For supervised deep learning, it is necessary to select a suitable training data set. There are many specific implementation methods, such as time series models such as LSTM and its variants, Transformers series models, or LLM models. Sentiment analysis belongs to a specific field of text classification. Generally speaking, depending on the application scenario, the Transformers series model and the LLM model have the best performance respectively. This model will be trained based on one of the Transformers series models to obtain an optimal model for sentiment classification in the field of Chinese online short texts.

(f) Model Evaluation

Different evaluation indicators (such as accuracy, precision, recall, F1 score, etc.) are used to evaluate the effect of the model. Cross-validation and other methods are also used to ensure the robustness of the model. The main evaluation criteria for this study are accuracy, precision and recall. Specific evaluation criteria are selected based on the different task types of word segmentation, new word discovery and sentiment analysis, as well as the extreme performance of the model.

(g) Deployment

Apply modeling results to actual business or projects. This includes model launch, monitoring, and integration with other systems. The final model deployed this time needs to be expanded with a long-term training data set after deployment, and the model needs to be continuously adjusted based on the prediction results. The data

collection program needs to continuously update the crawler program according to the changes of social media websites.

Figure 3.2 Research Framework(Complete framework will be added after the project is completed)

3.2.2 Problems to be solved

The primary task of this study is to build a sentiment analysis model for Chinese online short texts. The figure below is a simple flow chart for data acquisition, processing and modeling. To obtain truly effective conclusions, there are still some issues that need to be addressed in each specific link.

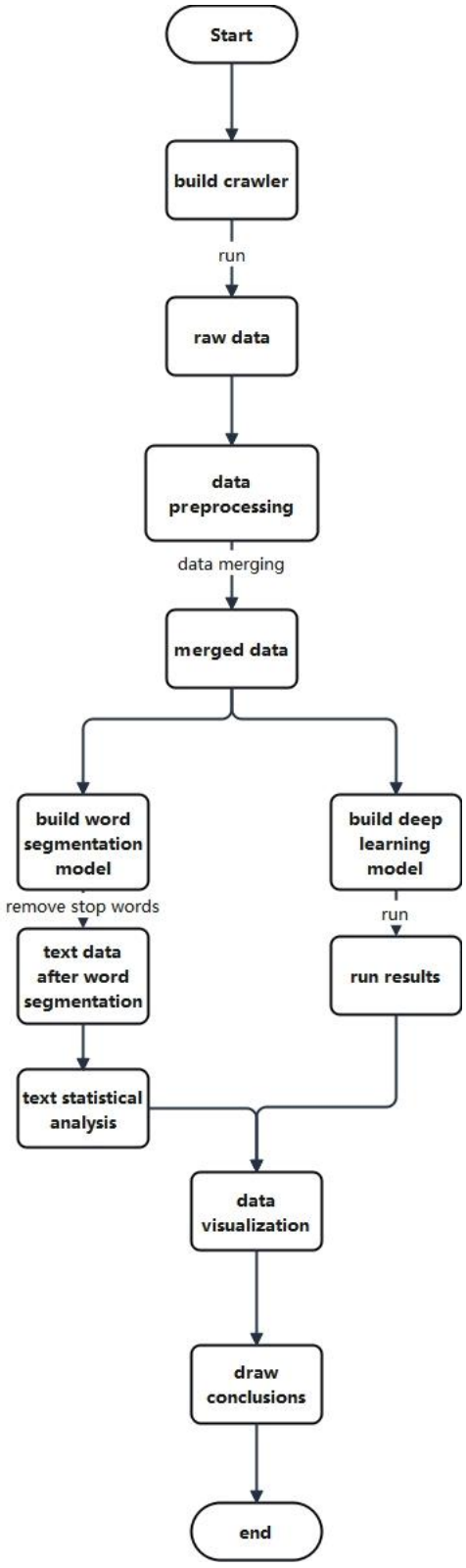


Figure 3.3 Simple data flow chart

(a) Data Collection and Quality

- Web crawler to directly crawl web text data will result in some unnecessary data, such as various icons, emoticons and other special texts, as well as special words such as replies and topics that belong to the main text. These contents are collectively referred to as markers. In addition, there are various network links that do not belong to the main text, so these contents need to be identified and processed.

- The data comes from multiple platforms, so the data format needs to be unified and the final data needs to be merged into the same table.

(b) Chinese word segmentation and stop word removal

- Chinese text has no spaces between characters, so a word segmentation model is needed for word segmentation. However, the word segmentation models in different fields vary greatly, so a reasonable word segmentation model needs to be built. Here we choose the COARSE_ELECTRA_SMALL_ZH model in HanLP as the word segmentation model. The original algorithm of this model is Electra (Clark et al. 2020) small model trained on coarse-grained CWS corpora. Its performance is P: 98.34% R: 98.38% F1: 98.36% which is much higher than that of MTL model. The corpus comes from the Chinese corpus constructed by HanLP, which mainly comes from Chinese published text data, with a total corpus of more than 100 million.The figure below is a flowchart of sentiment analysis, where the input is "这里是测试文本", It means "Here is the test text".

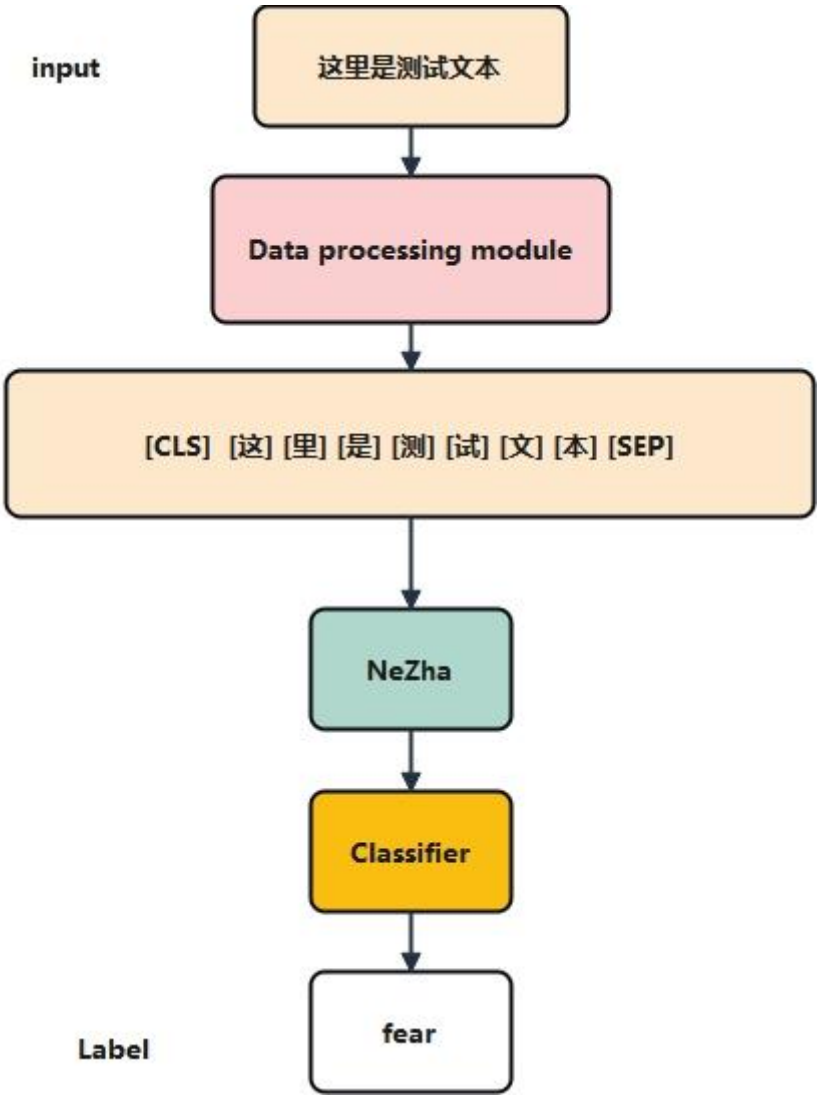


Figure 3.4 Flowchart of sentiment analysis

- Sources of stop words: Different languages have some special words such as textual expressions or modal particles that have no practical meaning. Stop words come from a wide range of sources, and there are certain differences in stop words in different corpora. This stop word list is constructed by me (Tian Fengshou) based on the stop words of Baidu, Google and Sogou, combined with the actual situation of online short texts. It is mainly aimed at online texts, and some commonly used symbols and meaningless words are added.

(c) Building a Sentiment Analysis Model

- Construction and selection of training data set: The actual prediction ability of supervised deep learning models is highly related to their training data, and the data we are going to process this time is short text data on the Internet. This is a sentiment analysis in a special scenario, which is different from the traditional sentiment analysis scenario. The main scenario of traditional sentiment analysis is long text. Therefore, we need to build our own training data set according to the actual situation. Here we choose SMP2020 as our training data set. The annotated data set used in this technical evaluation is provided by the Social Computing and Information Retrieval Research Center of Harbin Institute of Technology. The original data comes from Sina Weibo and is provided by the Micro Hotspot Big Data Research Institute. The data set is divided into two parts. The first part is the general Weibo data set. The Weibo content in this data set is randomly obtained from Weibo content, not targeting specific topics, and covers a wide range. The second part is the epidemic Weibo data set, which is not applicable. Each Weibo is labeled as one of the following six categories: neutral, happy, angry, sad, fear, surprise. The general Weibo training data set includes 27,768 Weibo posts, the validation set contains 2,000 Weibo posts, and the test data set contains 5,000 Weibo posts.

- Selection of pre-trained models: In the field of natural language, if you want to obtain a better model on a small amount of data, the best way is to pre-train on a basic model. This involves how to choose a basic pre-trained model. After multiple rounds of testing, we finally chose the NAZHA model provided by Huawei as the final pre-trained model.

3.2.2 Future Development

More training data. Currently, there is still little training data, and the performance of the algorithm has not been fully explored due to the lack of training

data. Therefore, it is necessary to further increase the training data in the future to improve the performance of the final model.

Correlation tracking of public opinion events: The spread of an event on multiple platforms has a time relationship. This study did not conduct in-depth research on this aspect.

3.3 Data sources and collection methods

Since the data comes from multiple platforms: Weibo, Bilibili, and Xiaohongshu. Therefore, the crawler here is built according to the actual situation. The Weibo data platform is built using the Requests and Scrapy frameworks. Since the platform has two network entrances cn and com, com has login restrictions, so cookies need to be added manually. cn has no login restrictions, and the cookie-free version can be used directly to obtain the required data. The data acquisition of the Xiaohongshu platform also requires cookies to obtain complete data, but the difference between the cookie-free version is not large, so the cookie-free version can also be used directly. The bullet screen and comments of Bilibili do not need to be logged in, so the cookie-free version is used directly.

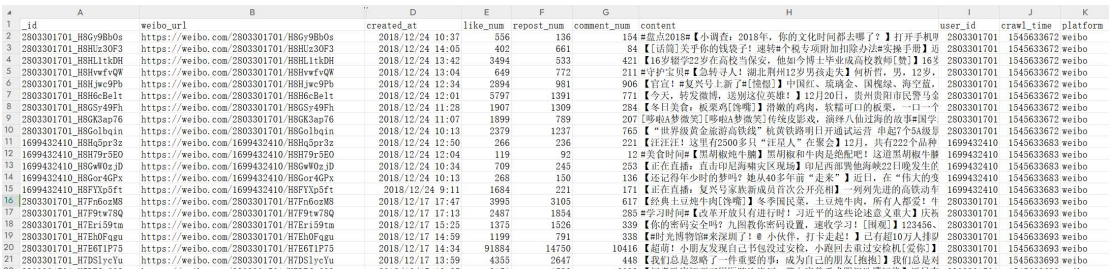
All downloaded data is stored in csv files and saved directly locally. The crawled content includes text, comments, relationships, and users.

The text data is the content officially released by the user. The official content of Bilibili is mainly video, and only links are saved. Weibo videos and Xiaohongshu videos are processed in the same way. Comment data refers to all comments and replies under the official content. User data refers to the personal information of different users. Considering that the data is not directly related to this study, the three platforms use different tables for storage. No data merging is performed. Relationship data refers to the relationship between followers and followed persons. This study does not use this data directly, so it is only stored and not processed.

3.4 Data preprocessing

3.4.1 Data merging

User data and relationship data are only stored, so they are not merged. The text and comment content of the three platforms are not much different, so they are directly stored in a unified format.



	A	B	D	E	F	G	H	I	J	K
1	_id	weibo_url	created_at	like_num	repost_num	comment_num	content	user_id	crawl_time	platform
2	2803301701_H8Gy9BbOs	https://weibo.com/2803301701/H8Gy9BbOs	2018/12/24 10:37	556	136	154	#盘点2018#【小调查：2018年，你的文化时间都去哪了？】打开手机里	2803301701	1545633672	weibo
3	2803301701_H8HUz30F3	https://weibo.com/2803301701/H8HUz30F3	2018/12/24 14:05	402	661	84	【[话筒]关乎你的钱袋子！速转#个税专项附加扣除办法#实操手册】近	2803301701	1545633672	weibo
4	2803301701_H8HL1tkDH	https://weibo.com/2803301701/H8HL1tkDH	2018/12/24 13:42	3494	533	421	【16岁辍学22岁在高校当保安，他如今博士毕业成高校教师[赞]】16岁	2803301701	1545633672	weibo
5	2803301701_H8Hvwfv0W	https://weibo.com/2803301701/H8Hvwfv0W	2018/12/24 13:04	649	772	211	#守护"宝贝#【急转寻人！湖北荆州12岁男孩走失】何忻哲，男，12岁，	2803301701	1545633672	weibo
6	2803301701_H8HJwc9Pb	https://weibo.com/2803301701/H8HJwc9Pb	2018/12/24 12:34	2894	981	906	【宜宜！#复兴号上新了#[赞啦]】中国红、祖国金、国槐绿、海空蓝、	2803301701	1545633672	weibo
7	2803301701_H8H6cBe1t	https://weibo.com/2803301701/H8H6cBe1t	2018/12/24 12:01	5797	1391	771	【今天，转发微博，送别这位英雄！】12月20日，贵州贵阳市民警马全	2803301701	1545633672	weibo
8	2803301701_H8GSy49Fh	https://weibo.com/2803301701/H8GSy49Fh	2018/12/24 11:28	1907	1309	284	【冬日美食，板栗鸡[馋嘴]】滑嫩的鸡肉，软糯可口的板栗，一口一个	2803301701	1545633672	weibo
9	2803301701_H8GK3ap76	https://weibo.com/2803301701/H8GK3ap76	2018/12/24 11:07	1899	789	207	[哆啦A梦微笑][哆啦A梦微笑]传统皮影戏，演绎八仙过海的故事#国学	2803301701	1545633672	weibo
10	2803301701_H8Golbqin	https://weibo.com/2803301701/H8Golbqin	2018/12/24 10:13	2379	1237	765	【"世界级黄金旅游高铁线"杭黄铁路明日开通试运营 串起7个5A级景	2803301701	1545633672	weibo
11	1699432410_H8Hq5pr3z	https://weibo.com/1699432410/H8Hq5pr3z	2018/12/24 12:50	266	236	221	【汪汪汪！这里有2500多只"汪星人"在聚会】12月，共有222个品种	1699432410	1545633683	weibo
12	1699432410_H8H79r5E0	https://weibo.com/1699432410/H8H79r5E0	2018/12/24 12:04	119	92	12	#美食时间#【黑胡椒炖牛腩】黑胡椒和牛肉是绝配吧！这道黑胡椒牛腩	1699432410	1545633683	weibo
13	1699432410_H8GwWOzjD	https://weibo.com/1699432410/H8GwWOzjD	2018/12/24 10:34	709	245	253	【正在直播：直击印尼海啸实区现场】印尼西部繁忙海峡221吸发生此	1699432410	1545633683	weibo
14	1699432410_H8Gor4GPx	https://weibo.com/1699432410/H8Gor4GPx	2018/12/24 10:13	268	150	136	【还记得年少时的梦吗？她从40多年前"走来"】近日，在"伟大的变	1699432410	1545633683	weibo
15	1699432410_H8FYXp5ft	https://weibo.com/1699432410/H8FYXp5ft	2018/12/24 9:11	1684	221	171	【正在直播，复兴号家族新成员首次公开亮相】一列列先进的高铁动车	1699432410	1545633683	weibo
16	2803301701_H7Fn6ozM8	https://weibo.com/2803301701/H7Fn6ozM8	2018/12/17 17:47	3995	3105	617	【经典土豆炖牛肉[馋嘴]】冬季国民菜，土豆炖牛肉，所有人都爱！牛	2803301701	1545633693	weibo
17	2803301701_H7F9tw78Q	https://weibo.com/2803301701/H7F9tw78Q	2018/12/17 17:13	2487	1854	285	#学习时间#【改革开放只有进行时！习近平的这些论述意义重大】庆祝	2803301701	1545633693	weibo
18	2803301701_H7Eri59tm	https://weibo.com/2803301701/H7Eri59tm	2018/12/17 15:25	1375	1526	339	【你的密码安全吗？九图教你密码设置，速收学习！[围观]】123456、	2803301701	1545633693	weibo
19	2803301701_H7EbOFqgu	https://weibo.com/2803301701/H7EbOFqgu	2018/12/17 14:59	1199	791	338	[#时光博物馆#来深圳了！@ 小伙伴，打卡走起！]已有超10万人佳贺	2803301701	1545633693	weibo
20	2803301701_H7E6T1P75	https://weibo.com/2803301701/H7E6T1P75	2018/12/17 14:34	91884	14750	10416	【超萌！小朋友发现自己书包没过安检，小跑回去重过安检机[爱你]】	2803301701	1545633693	weibo
21	2803301701_H7DSlycYu	https://weibo.com/2803301701/H7DSlycYu	2018/12/17 13:59	4355	2647	448	【我们总是忽略了一件重要的事：成为自己的朋友[抱抱]】我们总是对	2803301701	1545633693	weibo

Figure 3.5 Screenshot of the text data

The fields in the body are: _id, url, created_at, like_num, repost_num, comment_num, content, user_id, crawl_time, platform. Among them, _id is the unique identifier, url is the body link, created_at is the creation time, like is the number of likes, repost_num is the number of reposts, comment_num is the number of comments, content is the body content, user_id is the publisher id, crawl_time is the crawl time, platform is the publishing platform.

The content of the comment data is similar, but only contains _id, comment_user_id, content, url, created_at, crawl_time, platform. Among them, comment_user_id is the id of the commenter, and url is the id of the content being commented on.

3.4.2 Data cleaning

Data cleaning of online text mainly removes text data that does not belong to the main content. For hyperlinks, we can directly use regular expressions to remove them, or use the hyperlink recognition built into the word segmentation model to remove them. In addition, Weibo and Xiaohong also contain five special markers that need to be processed separately.

1. One is the additional information in the comment, For example, this example comes from actual data obtained, "回复@*:", similar to the following: "回复@鼾甜鼾甜的彼得潘:我知道你们这些都是海归或是海归的父母，搞笑

国内没有好大学？ \和你们这些把国外的[大便]都奉若珍宝人有什么好说的！！！"

2. One type of content is @user. For example, this example comes from actual data obtained, "nihao @dfugo @jb51 haha"

3. One category is emoticons and icons, For example, this example comes from actual data obtained, "铭记历史，勿忘国耻。老爷爷，您一路走好[蜡烛]。" In the original data of Weibo, emoticons are stored in a format similar to `[text]`.

4. Title, Weibo and Xiaohongshu titles will also be directly obtained. For example, this example comes from actual data obtained, "【汪汪汪！这里有 2500 多只“汪星人”在聚会】12 月，共有 222 个品种，超过 2500 只小狗参加第二届波兰卢布林国际犬展。戳视频，一起来看它们的萌样子～[下] [下] [下] 新华视点的秒拍视频"

5. Topics, Weibo or Xiaohongshu content is sometimes related to some internal topics. For example, this example comes from actual data obtained, "【官宣！#复兴号上新了#[憧憬]】中国红、琉璃金、国槐绿、海空蓝，复兴号大家族再添高颜值新成员！时速 350 公里 17 辆长编组、时速 250 公里 8 辆编组、时速 160 公里动力集中等多款复兴号新型动车组首次公开亮相，TA 们既有颜值，更有内涵，最快在明年 1 月 5 日上线！你期待吗？@中国铁路"

Among the five markers, 1, 2, 4, and 5 can be removed using regular expressions, as shown in the code diagram.

```
def weibo_clear(text):
    import re
    result = re.sub(r'【.*】', " ", text)
    result = re.sub(r'#.*#', " ", result)
    result = re.sub(r'回复@.*:', " ", result)
    result = re.sub(r'@([\u4e00-\u9fa5\w\-]+)'," ",result)
    return result
```

Figure 3.6 Code to remove markers

However, the formats of emoticons and icons are special and may be mixed with normal symbols. The two tasks of this study are text statistical analysis and sentiment analysis. In sentiment analysis, the original training data contains emoticons, so they do not need to be removed. In the text statistical analysis stage, emoticons need to be specially processed. Therefore, we have two options. One is to directly remove them with regular expressions, and the other is to build an emoticon discovery system for statistical analysis. The second method needs to be combined with a word segmentation model, so it will be explained in detail in the next chapter.

REFERENCES

Graham, S., Sweeney, P., & Alas, B. (2018). Data science lifecycle: A review and framework. In 2018 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 360-369). IEEE. https://doi.org/10.1109/DSAA.2018.00057

Shin, S., Ryu, W., Park, S., Cho, K., & Lee, D. (2021). Effective Sentence Scoring Method Using BERT for Speech Recognition. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), 5554 – 5566. https://doi.org/10.18653/v1/2021.emnlp-main.451

Clark, K., Luong, M., Le, Q., & Manning, C. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. International Conference on Learning Representations (ICLR). https://openreview.net/pdf?id=r1xMH1BtvB

Li, X., Liu, J., Li, S., & Zhao, X. (2020). SMP2020-EWECT: A dataset for Weibo sentiment classification. Proceedings of the China National Conference on Social Media Processing (SMP 2020). https://smp2020.ewect.com

Wei, J., Wang, X., Tian, H., Bai, Y., Lan, Y., Wu, J., Wang, S., & Li, X. (2019). NEZHA: Neural contextualized representation for Chinese language understanding. arXiv preprint arXiv:1909.00204. https://arxiv.org/abs/1909.00204