BiliVista

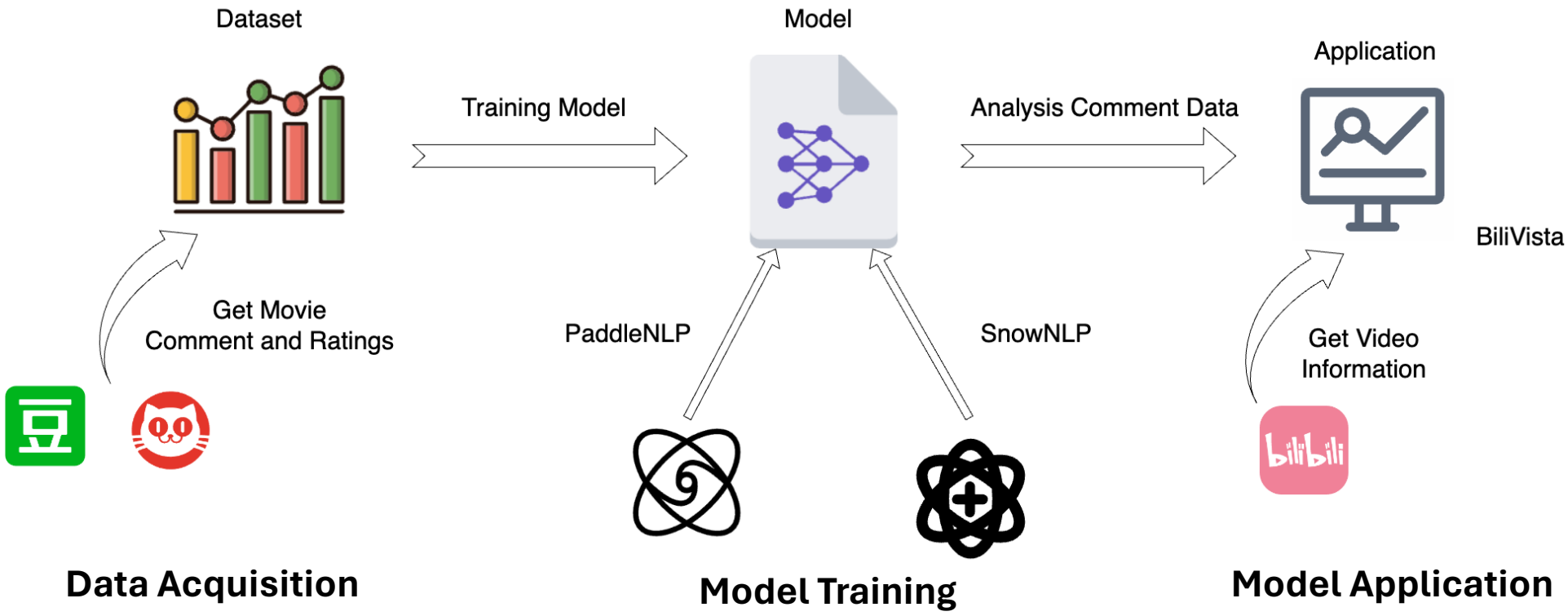
Bilibili data analysis platform

Emotion analysis based on deep learning

Project Overview

Data Acquisition

- *Crawling Movie Comments*
  - *Douban: Collect comments from the `Douban` movie site.*
  - *Maoyan: Collect comments from the `Maoyan` movie site.*
- *Data Preparation*
  - *Perform data cleaning to ensure the quality and consistency of the collected comments.*

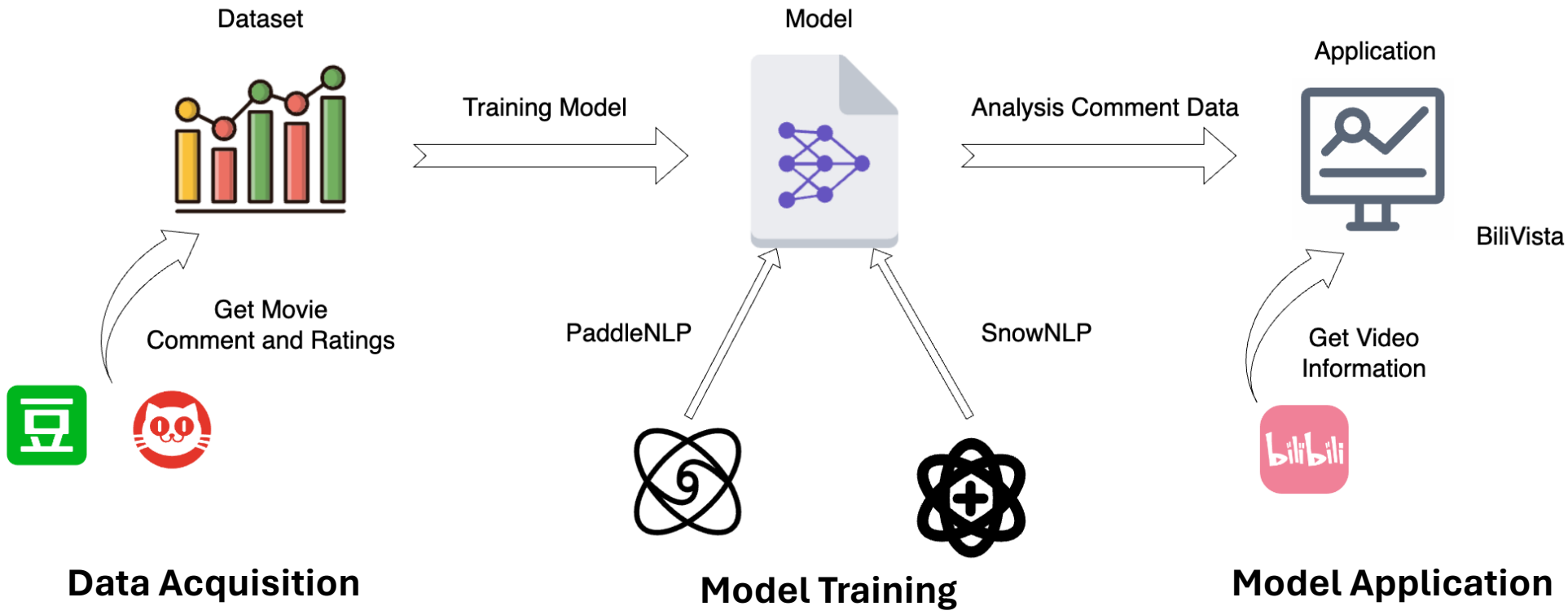


杨钧涯

Project Overview

Model Training

- *Machine Learning Model*
  - *Utilize SnowNLP for sentiment analysis*
- *Deep Learning Model*
  - *Employ PaddleNLP to build and train deep learning models for advanced text analysis.*

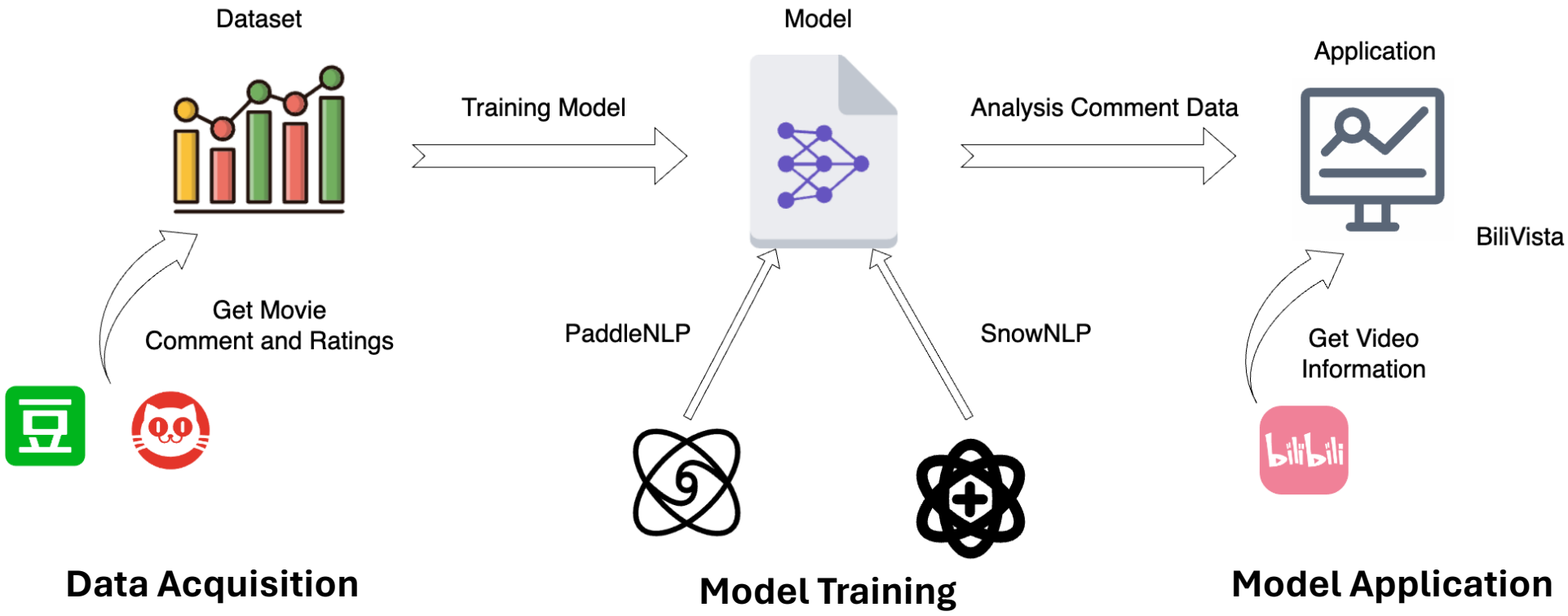


杨钧涯

Project Overview

Model Application

- *Crawling Data from `Bilibili`*
  - *Implement real-time data crawling from `Bilibili`*
- *Backend Development*
  - *Develop the backend using the **FastAPI** Python*
- *Frontend Development*
  - *Implement a user-friendly interface to visualize and interact with the analysis results.*



杨钧涯

Data Acquisition

Crawling Movie Comments

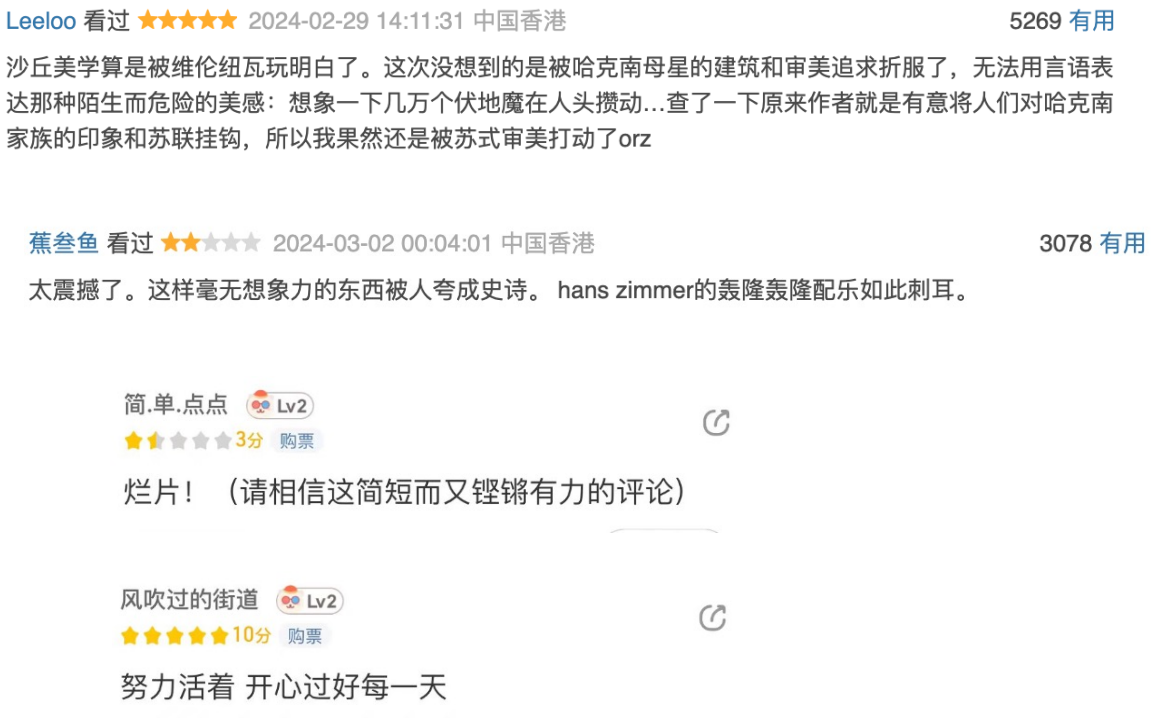
- ***Douban: Collect comments from the Douban movie site.***
- ***Maoyan: Collect comments from the Maoyan movie site.***

Source of data

- *Douban*
- *Maoyan*

Why These Datasets?

- ***Labeled Data:*** This is critical as labeled data provides a foundation for training and evaluating our sentiment analysis models with higher accuracy.



王孟涵

Data Acquisition

Data Preparation

- ***Perform data cleaning to ensure the quality and consistency of the collected comments.***

Data Categorization

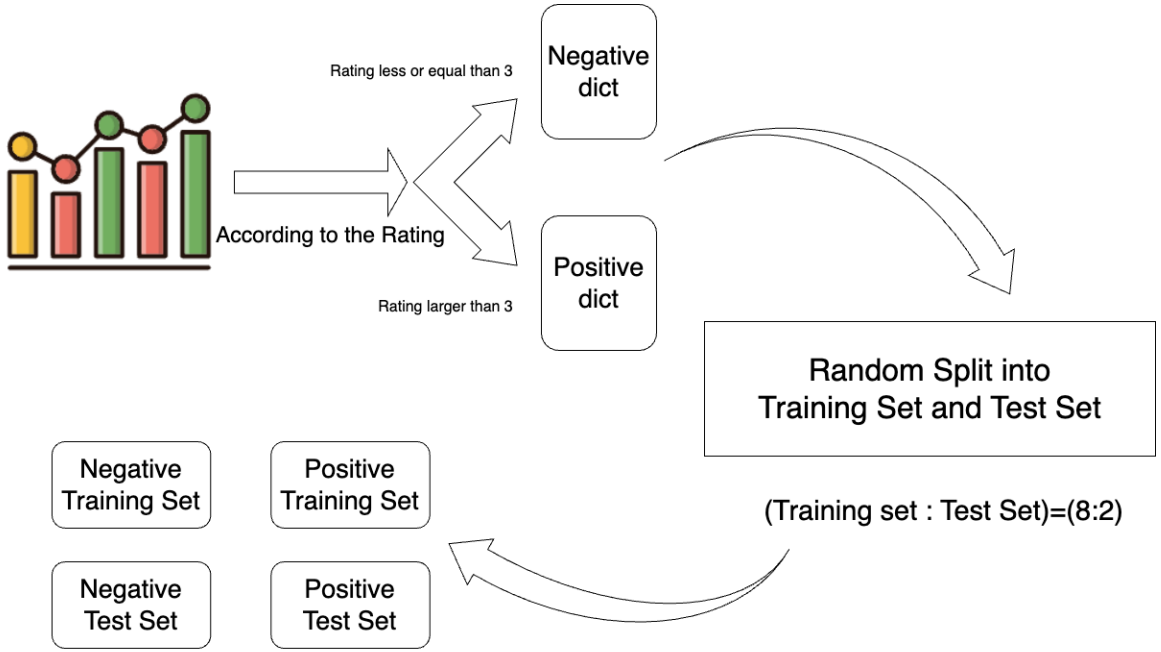
- ***Negative dict*: Ratings less than or equal to 3**
- ***Positive dict*: Ratings greater than 3**

Dataset Splitting

- **Random Split**

Training and Testing set

- **Training set : Test set = 8:2**



王孟涵

Model Training

Machine Learning Model

- ***Utilize SnowNLP for sentiment analysis***

SnowNLP

- SnowNLP is a library focused on natural language processing tasks for Chinese text, such as sentiment analysis and text processing.
- Uses the labeled training data to train the classifier using the Naive Bayes algorithm.

Naive Bayes Algorithm

- Assumes that features are independent of each other
- Estimates probabilities based on the features and labels in the training dataset.

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

$$x_i = (x^1, \dots, x^n)$$

$$y_i = c_k \quad and \quad k = 1 \dots K$$

$$y = \arg max_{c_k} P(y = c_k) \prod_j P\left(x^j | y = c_k\right)$$

缪润杰

Model Training

Machine Learning Model

- ***Utilize SnowNLP for sentiment analysis***

Naive Bayes Algorithm

- Assumes that features are independent of each other
- Estimates probabilities based on the features and labels in the training dataset.

Split Chinese characters and calculate the probability that the term appears in the set

Comments
很好，揭露了现实打工人的心酸，推荐
还不错 有一些地方挺搞笑的!
非常搞笑，值得推荐
太棒了，太真实了，推荐给领导们看看
搞笑喜剧，挺现实的社会现象

Cut Sentence →

Cut_Comments
很好 揭露 了 现实 打 工人 的 心酸 推荐
还 不错 有 一些 地方 挺 搞笑 的
非常 搞笑，值得 推荐
太棒了 太 真实了 推荐 给 领导 们 看看
搞笑 喜剧 挺 现实 的 社会 现象

Training Result

Method	Test Dataset Accuracy
SnowNLP	78.58%
KNN	78.32%

缪润杰

SnowNLP Model Evaluation

Practical application scenarios

- We actually using this model to annalysis the comment in other platform, it does not work well

Comment of Movie

- Crawl the Comment of the Movie 《*Wandering Earth 2*》 From the Bilibili

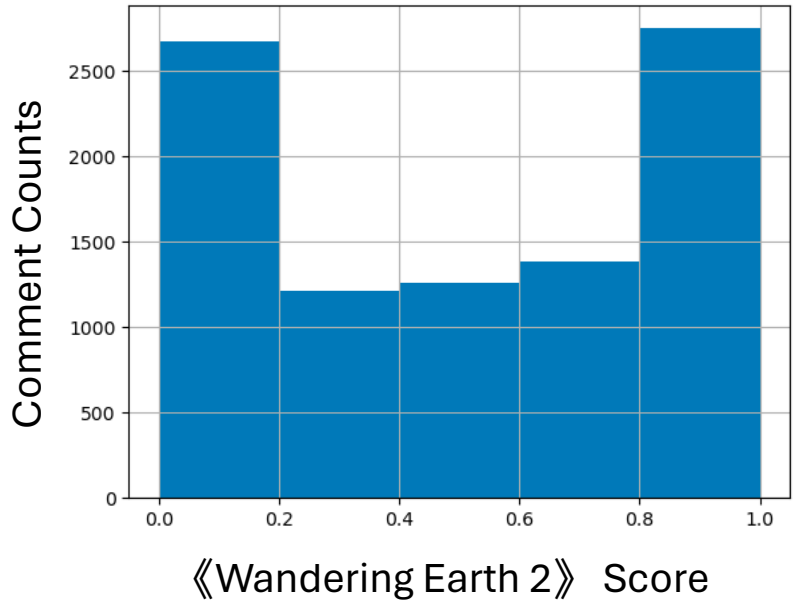
Estimation the comment score

- Get a comment sentiment score using the trained SnowNLP model.
- Average comment sentiment score is only 0.54.

(Assuming that the number of likes is the number of approvals)

Comment Score Distribution

- The distribution is polarized



缪润杰

SnowNLP Model Evaluation

Comment of Movie

- Crawl the Comment of the Movie 《*Wandering Earth 2*》 From the Bilibili

Comment	Like	Sentiment	Sentiment x Like
六公主给流浪地球2的颁奖词： 这是中国电影工业的一次全面跨越升级，以硬实力将中国科幻电影提升到前所未有的境界，7万多字原创剧本，2万多名工作人员，90多万平方米的置景总面积，历时1400余天的摄制，奉上一场2小时53的视觉盛宴。 如此庞大精良的制作规模，造就了影片同名话题超11.8亿的网络关注量，收获40.23亿票房。 M大数据显示，影片传播指数达9.8，位列年度影片之首。 片中昂扬的中国精神，如同闪耀的星群，照耀着中国科幻电影前进的方向。	13621	4.14E-11	5.64E-07
其实导演郭帆不是半路出家，而是从小励志拍科幻片。 郭帆15岁时看了卡梅隆的《终结者2》，然后立志以后拍科幻片，他高考本来要考电影学院，但山东省没有招生导演系的，郭帆母亲也劝他考法律后当个政法委总书记就行。 郭帆考上法律专业后想如果自己以后不奔着梦想去，等晚年躺病床、摇椅就特别后悔，所以他就觉得不管选什么专业，只要奔着哪个目标去，然后他大学也拍过电影短片，而且他学过法律学专业很适合工业化方面。 还有郭帆小时候画画很好，也拿过奖，有美术基础的。所以不要总半路出家、非科班、中途转行也能成功，搞得好像人家外行的行我也行，郭帆是自己本来就有这梦想、本来就有相关知识、而且他29岁时还考上北京电影学院管理系研究生。	7597	0.000216	1.639879
中国科幻元年必定是1999年。 郭帆在这一年高考，而且这一年的《科幻世界》压中了高考作文《假如记忆可以移植》，郭导看过并且受到启发拿了高分，同年也是大刘在科幻世界上开始首次投发文章《微观尽头》和《鲸歌》，次年就投发了《流浪地球》。同年高中生谢楠在1月的科幻世界发表了短篇奇想，页码43是吴京的生日4月3号。	6672	8.11E-10	5.41E-06

It turns out that the model does not work well in practice

缪润杰

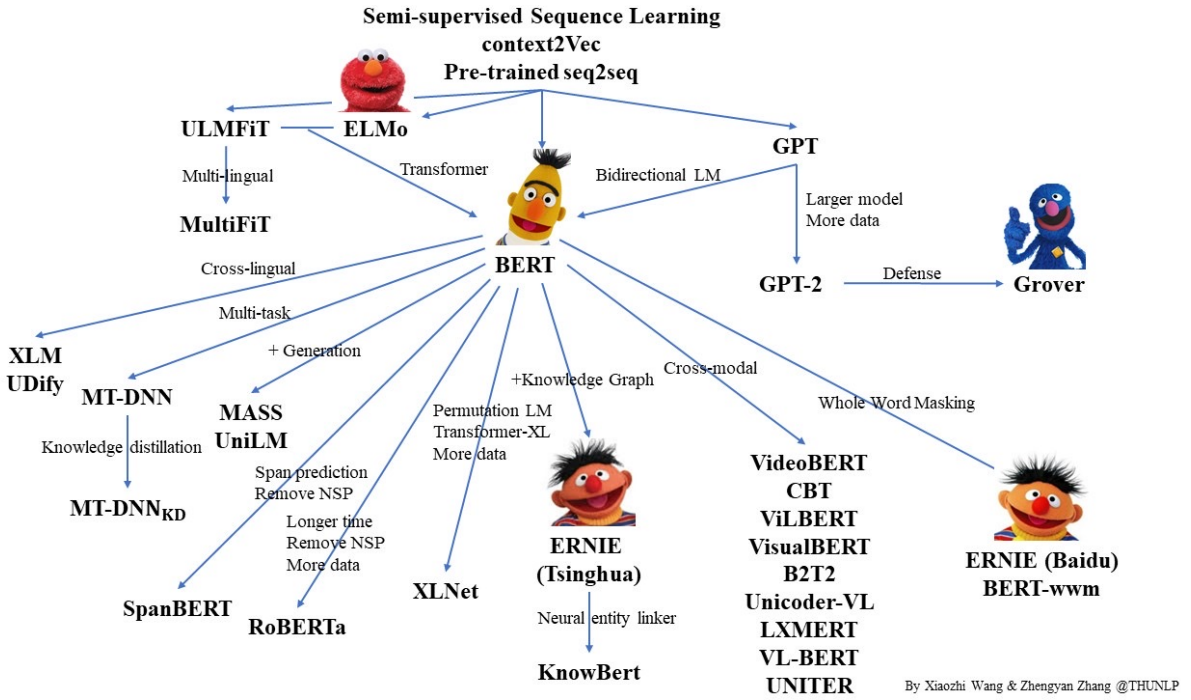
Model Training – Deeping Learning

Pre-training model

- With the development of deep learning, the number of model parameters has increased rapidly, and in order to train these parameters, larger data sets are needed to avoid overfitting.

- Nowadays, studies have shown that ***Pretrained Models (PTM)*** based on large-scale unlabeled corpora can acquire generic language representations and **perform well** when **fine-tuned to downstream tasks**.

- ***In addition***, pre-training models can avoid training models from scratch.



王孟涵

Model Training

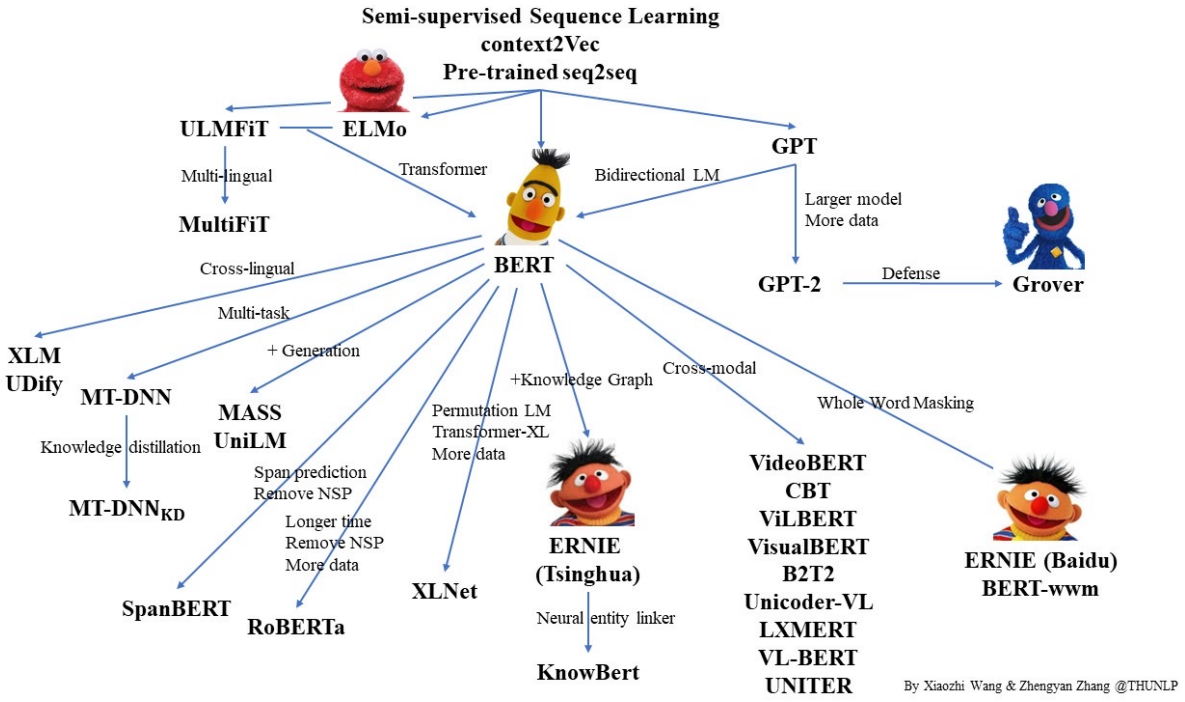
Pre-training model

- ERNIE (Like Bert-wwm)

BERT requires minimal architecture changes for a wide range of natural language processing applications.

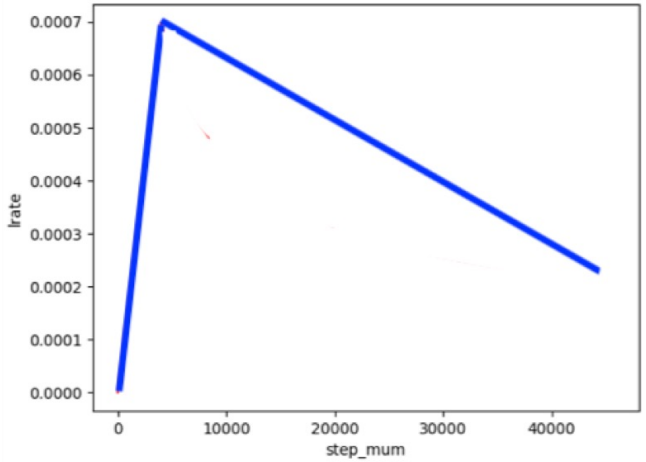
Deep Learning Model

Employ PaddleNLP to build and train deep learning models for advanced text analysis.



王孟涵

Model Training



Trainning Process

1. Remove a batch data from the dataloader
2. Feed batch data to the model for forward calculation
3. Pass forward calculation result to loss function to calculate loss. The forward calculation result is passed to the evaluation method, and the evaluation index is calculated.
4. Loss reverse return and update gradient. Repeat the above steps.

Each time an epoch is trained, the program will evaluate the effectiveness of the current model training.

Method	Test Dataset Accuracy
SnowNLP	78.58%
PaddleNLP	85.31%

王孟涵

PaddleNLP Model Evaluation

Practical application scenarios

- We actually using this model to annalysis the comment in other platform, it does not work well

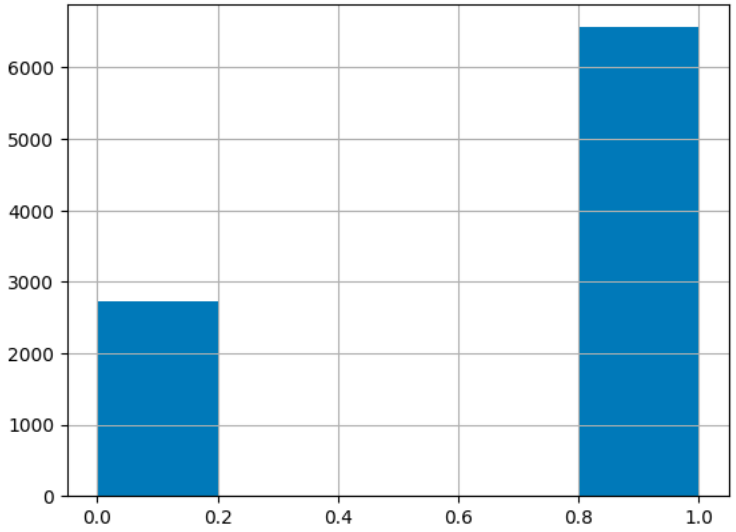
Comment of Movie

- Crawl the Comment of the Movie 《*Wandering Earth 2*》 From the Bilibili

Estimation the comment score

- Get a comment sentiment score using the trained SnowNLP model.
- Average comment sentiment score is 0.89.

(Assuming that the number of likes is the number of approvals)



《Wandering Earth 2》 Score

王孟涵

Model Application

Set the User-Agent and cookie information

- Set the User-Agent and cookie information
- Input the video bvid parameter to obtain the video.

Crawl video basic information

Through bvid access to the basic information of the video including:

- Title
- Author
- Reply count, Favorite count,
- Coin count, Share count
- ...

Crawling Data from Bilibili

- Implement real-time data crawling from Bilibili

```
HEADERS = {
    'User-Agent': '',
    'cookie':"",
}
url = 'https://api.bilibili.com/x/web-interface/view/detail'
params = {
    'bvid': 'BV117411r7R1'
}
```


```
{
  "BV1ct421u7gu": {
    "title": "\u30104\u6708\u3011\u602a\u517d8\u53f7 06",
    "pic": "http://i1.hdslb.com/bfs/archive/95d2e0fcb4bcc9b5dfe334dff7b087595cce2e44.jpg",
    "view": 1572285,
    "danmaku": 10845,
    "reply": 1030,
    "favorite": 1320,
    "coin": 5196,
    "share": 352,
    "like": 18884,
    "author": "\u54d4\u54e9\u54d4\u54e9\u756a\u5267"
  }
}
```


苏泊鑫

Model Application

Crawling Data from Bilibili

- *Implement real-time data crawling from Bilibili*

Crawl video comment content and danmu content

Through bvid access to the comment information of the video including:

- Comment
- Number of comments and likes

Through bvid request screen XML file obtained through basic information:

- Time
- Timestamp
- Danmu text
- ...

苏泊鑫

Model Application

Backend Development

- ***Develop the backend using the FastAPI Python***
- ***FastAPI*** is a modern, fast (high-performance) web framework for building APIs.

Why FastAPI?

Key Advantages:

- High Performance
- Rapid development
- Automatic interactive API documentation

Usage of 3 APIs:

- ***video***: Fetch specific video information.
- ***videos***: Gather statistics on all scraped videos.
- ***newvideo***: Add new video data to the system.

Backend Architecture Overview

杨钧涯

Demo

Frontend Development

- ***Implement a user-friendly interface to visualize and interact with the analysis results.***

肖嘉皓

Thanks