Cross-Cultural Differences of Sentiment in Social Media Posts Respond to Major event over Time

Columbia University

Zheng Hui(zh2683)

Zihang Xu(zx2362)

Advisor: Professor John Kender

Abstract

The emotional response of different cultures in major events should be different. People in different countries have their own opinions and emotions based on their culture, thoughts, habits, and customs. Our main goal is to distinguish the cultural differences of posts on social media. These posts have different emotions in major events in COVID-19 in different countries /regions. In the same theme, although some people may be satisfied, others may show resentment or anger. Some cultures will express emotions more straightforwardly, but some cultural expressions are more vague which identify such emotions as a key topic in the field of natural language processing. In this article, we decided to choose Britain and China as our research objects. They have a completely different policy for COVID-19, and major events in great popularity will also have different emotions on social platforms. Therefore, in the same period, the people of these countries may have different attitudes.

In this paper, our goal is to compare the differences and similarities between the two cultures by comparing the emotional fluctuations in major events in Weibo China and the British Twitter for a period of time during Covid pandemic . The reason for the keyword we chose COVID-19 as the theme is that it has maintained the life of the people of the entire world for three years. Only by analyzing the information about COVID-19, we can successfully guess the real situation of each country and compare it.

1. Introduction

To this day, COVID-19's popularity has plagued us for three years. In these three years, many major events such as Wuhan have been blocked, vaccines came out , the state emergency state declaratory and so on. In these major events, people from different countries show different emotions and views on social platforms. In this era of information, social media is a very important part of people's lives that may be the first stop for people to vent or show their emotions. So it is necessary to analyze the large -scale social media sentiment analysis of major events in COVID19. First, this can highlight the emotional response and mental health of different cultural responses to major events. The universal values and policy feedback, the third can remind people that their mental health is popular in COVID-19.

In this project, we have selected two major social media of different countries for analysis, and the countries are Britain and China. For Britain, the social media data we use comes from Twitter. For China, the social media data we use comes from Weibo. The time range of the data of the British Twitter is one month (4 weeks), and the time range of Chinese Weibo data is also one month (4 weeks). If we see the occurrence of major events as a point, the selection of the time frame is two weeks before the major event and two weeks after the major event occurred, which is one month. Then we use different deep learning models to learn data, and then perform the sentiment analysis. For emotions, a popular example is Paul Ekman and his colleagues on cross

-cultural research in 1992. They concluded that the six basic emotions of humans are anger, disgust, fear,happiness, sadness and surprise.

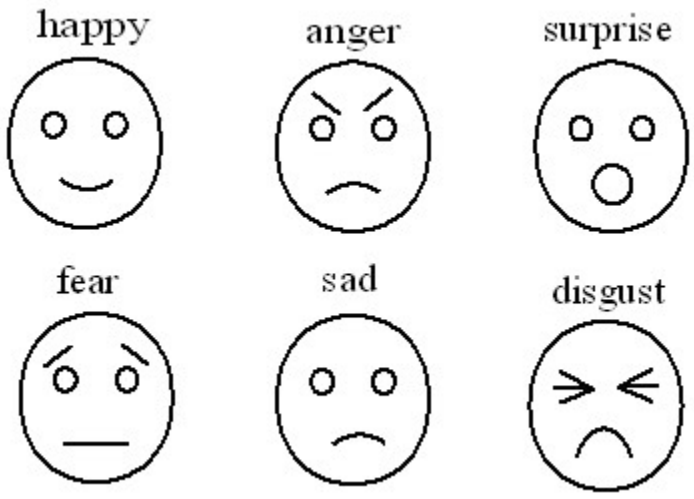


Figure 1: 6 emotions by Paul Ekman

In our project, we simply classify emotions to three in order to better reflect the differences, which are positive, negative, and neutral.

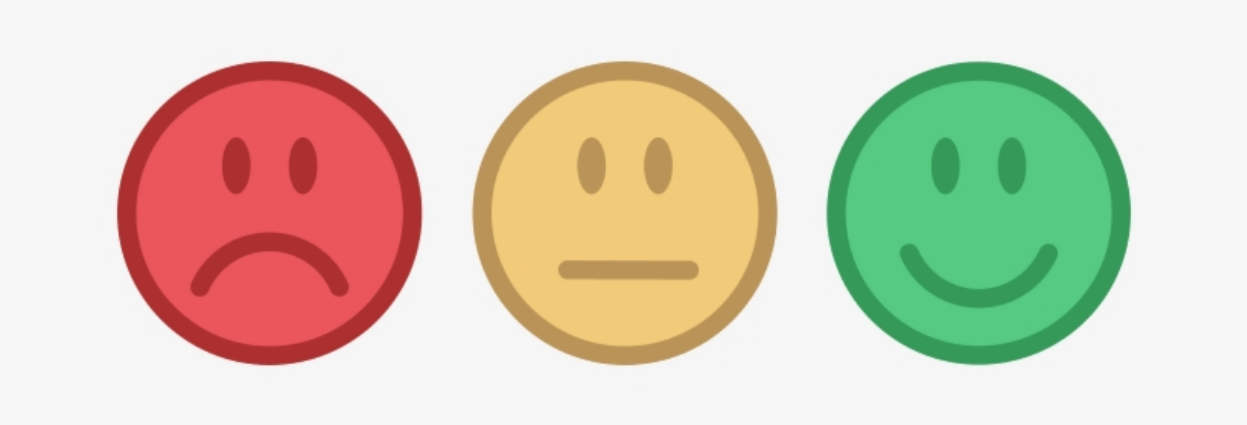


Figure 2: 3 emotions

From dataset collection, we build a web crawler for Weibo since Weibo does not offer APIs and there is no exist Weibo dataset available on the keyword “新冠” which in english means Covid-19. As for the UK Twitter data, we obitian it from an online github

dataset called COVID19_Tweets_Dataset (https://github.com/lopezbec/COVID19_Tweets_Dataset) and we are using the data identity as from the United Kingdom.

After obtaining data, we predict the data through two different deep learning models. On Twitter dataset we use COVID-Twitter-BERT from digitalepidemiologylab and use Chinese-BERT on Weibo dataset from China. Through our improvement and fine -tuning, we successfully improved the accuracy of 9% , compared to the Baseline Model.

2.RELATED WORK

2.1 Social Media Covid Database

There are many ways to collect data on social media. The mainstream includes API comments provided by social media such as Yelp, Facebook developer platform, and such as using third -party software such as octopus, or you can choose to build your own crawler yourself. Come to crawl data. Related work includes the construction of crawlers, the use of APIs and the adjustment of the data set. The main data of Twitter is derived from An Augmented Multilingual Twitter DataSet for Studying The Covid -19 Infodemic by Lopez, C. E., Gallemore, C.

2.2 Mental Health on Social Media during Covid-19

People's emotional health is also a topic that everyone cares about during a big period of popularity. Under the influence of social distance and city lockdown, people's negative emotions will have a great impact on emotional health. Many jobs also pay attention to human emotional fluctuations and emotional health on social platforms. And the association between social media and human emotional health is like Social Media Use and ITS Connection to Mental Health: A Systematic Review by Fazida Karim. And natural language processing seems to be the most consistent and most powerful tool. For example Christopher Marshall‘s Using Natural Language Processing to Explore Mental Health Insights From UK Tweets During the COVID-19 Pandemic: Infodemiology Study, or Natural language processing applied to mental illness detection: a narrative review by Tianlin Zhang and Annika M. Schoene.

2.3 Cross-Cultural Differences of Sentiment in Social Media

Another major topic is how people respond to the major incidents in COVID-19 in social media. There is relatively little related work in this field, but the Cross-Cultural Differences of Social Media Related Relates Related To Covid-19 tells how to use the classification model to identify the emotions of social media in different cultures for COVID-19. The keywords used in this article are Wuhan and Covid. The article shows the classification and quantity statistics of the six different emotions on social media

on social media. Data shows that Twitter data and Weibo data have different emotional fluctuations when specifying keywords in COVID-19. The most emotional emotions displayed on Weibo are positive, however for Twitter people are feeling more fear or neutral.

3.MATERIAL & METHODS

3.1 Data collection

A. Twitter Data

Twitter is an American social networking and microblogging service company. Twitter is a platform that provides real-time global events and discussion of hot topics. On Twitter, real-time commentary conversations showcase4 every side of a story, from breaking events, entertainment, sports, politics, and daily news. Here you can join open live conversations, or watch the event live. Twitter is a very popular social media for the world, and the reason we choose twitter over other social media is people tend to share more about emotion and feeling for society, not just family or friends. Twitter also gets a lot of tweets every day. The figure below shows home the traffic on twitter with covid related keywords and why we believe twitter is a good source to collect our data. The twitter data we are collection from is PanelLab Covid-19 database and COVID19_Tweets_Dataset by Dr. Christian Lopez

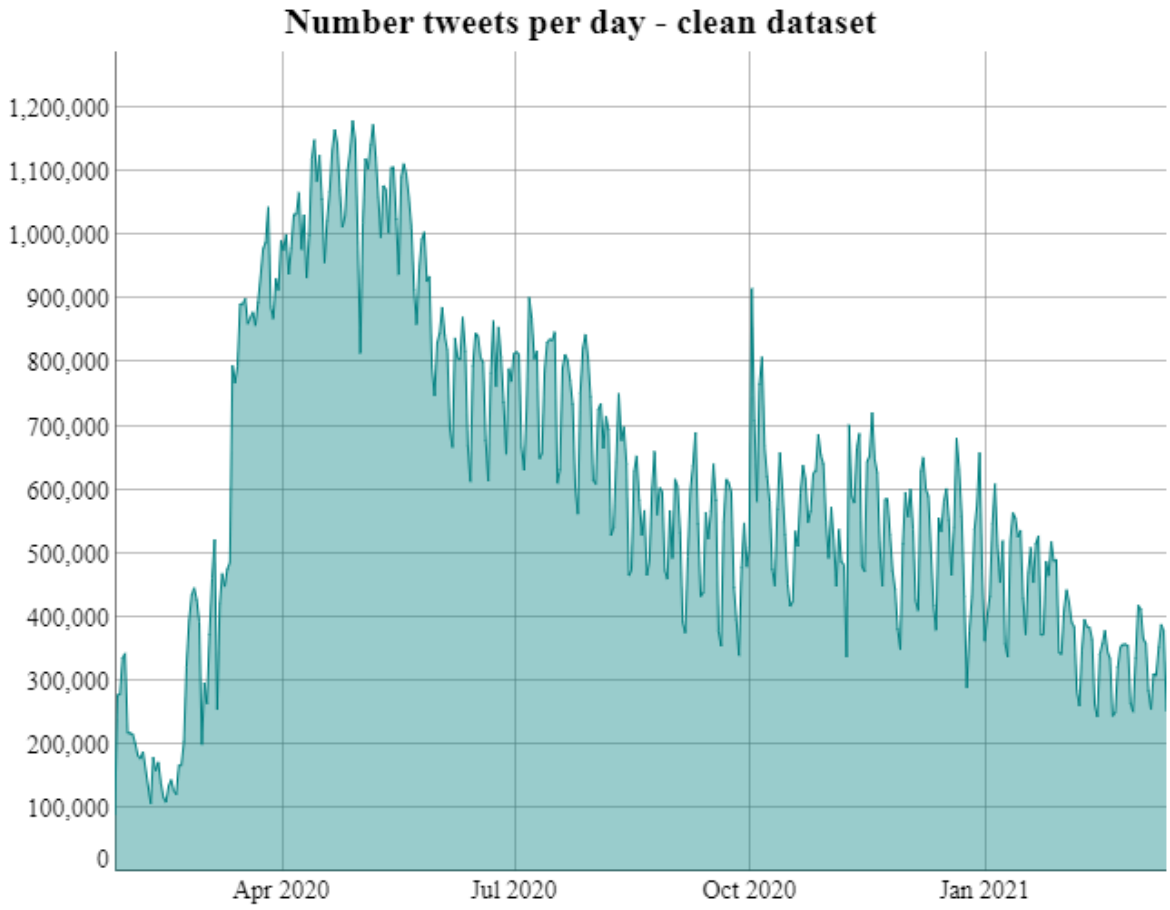


Figure 3: Total Covid related Tweets by date

Based on the code and filter tool provided by the dataset github repository we are about to filter out only UK data based on geolocation provided by the dataset. Then we also add sentiments label add on to the tweet itself by artificial recognition. A code and final after processed data snap is blow:

```python
def is_english_and_usa_tweet(tweet):
    if tweet[1] == 'en':
        if tweet[6] == "UK":
            return True
        else:
            try:
                coordinates = tweet[2][1:-1].split(',')
                coordinates =[float(i) for i in coordinates]
                result_coordinates = rg.search(coordinates)[0]
                if result_coordinates['cc'] == "UK":
                    return True
                else:
                    return False
            except:
                return False
        # Please use this else statement if you remove the reverse latitude functionality
        # else:
        #     return False
    else:
        return False
```

Figure 4: Code Snap for filter database based on geolocation

	Tweet_ID	Sentiment_Label
2	1245140084321632258	0
3	1245140084350910464	0
4	1245140084367732737	1
5	1245140084371927041	1
6	1245140084401332224	1
7	1245140084413935616	0
8	1245140084417941505	1
9	1245140084455813120	0
10	1245140084464029697	1
11	1245140084489359365	1
12	1245140084493627395	0
13	1245140084497637376	0
14	1245140084522803201	0
15	1245140084522967042	1
16	1245140084535496706	1
17	1245140084556521473	0
18	1245140084568936450	1
19	1245140084581715977	0
20	1245140084803829760	1
21	1245140084824739840	0
22	1245140084879265793	1
23	1245140084988485633	1
24	1245140084996874240	1
25	1245140085021917185	0

Figure 5: TweetID with sentiment Label from UK on April 1, 2020

For figure 5, since Twitter provides the universal identifier for each tweet so we are only keeping the identifier as premier key and for sentiment label we are using 0 for negative, 1 for neutral and 2 for positive.

B. WeiBo Data

Since WeiBo does not provide a very comprehensive API to fulfill our need for data and sentiment analysis, For Chinese Covid-19 data, we decided to crawl the related posts and comment on Weibo.

Web Scraping

What is a web crawler? Web crawlers are also called web robots, which can automatically collect and organize data information on the Internet instead of people. In the era of big data, information collection is an important task. If information collection is done solely by manpower, it will not only be inefficient and cumbersome, but also increase the cost of collection. To put it simply, a crawler is a detection machine. Its basic operation is to simulate human behavior to wander around various websites, click buttons, check data, or recite the information it sees. Like a bug crawling around tirelessly in a building.

At this point, we can use web crawlers to automatically collect data information, such as crawling and collecting sites in search engines, collecting data in data analysis and mining, and collecting financial data in financial analysis In addition, web crawlers can

also be applied to various fields such as public opinion monitoring and analysis, and target customer data collection.

What is Weibo?

Everyone knows what blogging means. Blogs, also known as weblogs, are websites that are usually managed by individuals and that post new articles from time to time. Weibo is the abbreviation of micro blog. It is a user relationship information sharing. Spread and obtain a platform, so that users can directly set up a personal community through web, wap, etc., and then update their mood, knowledge, news and other information on it, but the microblog information published cannot exceed 140 characters.

Of course, Weibo only provides a platform for users. If you apply for Weibo, you can update some mood or some useful information on your own Weibo. In this way, other people can post short messages to your information after browsing your information. The biggest advantage of Weibo is that it releases information quickly. As long as you post information, it will immediately update the information to your audience, so that Your message is instantly transmitted to your audience.

Compared with regular personal blogs, Weibo is not a long-winded blog post. Instead, post a few short sentences or a sentence on your Weibo. So his difficulty is much easier than that of a personal blog. This is also the reason why more and more people use Weibo. Secondly, the various APIs opened by Weibo enable a large number of users to update their personal information in real time through mobile phones and the Internet. As far as Weibo is concerned, as long as Sina Weibo and Tencent Weibo

occupy the mainstream market. For some friends who are engaged in network marketing, they use Weibo to carry out social media marketing.

In general: Weibo records short language narratives, which can be a few words, on-the-spot recordings, expressing emotions, and expressing emotions.

How to get the data

In order to analyze the Chinese words, we have to use the “jieba”.

What is jieba

Jieba is an open source library developed by Baidu engineer Sun Junyi. It is very popular and frequently used on GitHub.

GitHub link: https://github.com/fxsjy/jieba

The most popular application of jieba is word segmentation, which is also called "Jieba Chinese word segmentation" on the introduction page, but in addition to word segmentation, jieba can also do keyword extraction, word frequency statistics, etc.

jieba supports four word segmentation modes:

- Accurate mode: try to cut the sentence most precisely, and only output the maximum probability combination;
- Search engine mode: On the basis of the precise mode, segment the long words again to improve the recall rate, which is suitable for word segmentation in search engines;
- Full mode: scan all the words that can be formed into words in the sentence;

- paddle mode, using the PaddlePaddle deep learning framework to train sequence annotation (bidirectional GRU) network model to achieve word segmentation.

Part-of-speech tagging is also supported.

After we built the web crawler we finally were able to get our metadata from Weibo.

Code snap for our web crawler is shown in figure 6.

```
        except Exception as e:
            logger.exception(e)

    def get_result_headers(self):

        result_headers = [
            "id",
            "bid",
            "Body",

        ]
        if not self.filter:
            result_headers2 = ["Time_data", "Userid", "Username"]
            result_headers3 = ["origin" + r for r in result_headers]
            result_headers = result_headers + result_headers2 + result_headers3
        return result_headers

    def write_csv(self, wrote_count):
        write_info = self.get_write_info(wrote_count)
        result_headers = self.get_result_headers()
        result_data = [w.values() for w in write_info]
        file_path = self.get_filepath("csv")
        self.csv_helper(result_headers, result_data, file_path)

    def csv_helper(self, headers, result_data, file_path):
        if not os.path.isfile(file_path):
            is_first_write = 1
        else:
            is_first_write = 0
          with open(file_path, "a", encoding="utf-8-sig", newline="") as f:
              writer = csv.writer(f)
              if is_first_write:
                  writer.writerows([headers])
              writer.writerows(result_data)
        if headers[0] == "id":
```

Figure 6: Web crawler for Weibo

After we have the web crawler we are using it to collect data on Weibo, since we crawler the data down in our own way, Weibo data does not have an universal identifier like Twitter, so we give each of the content its own identifier. The data is shown in Figure 7.

Weibo_ID	Content	Sentiment_Label
1	感谢隔离桌寿星送来的小蛋糕	2
2	高端楼隔离最后一天/阿瑟新剧确实好	1
3	#姑娘的碎碎念# 早上到工位后一直到	2
4	隔离日记Day4 #疫情##日常##老师##	2
5	小?看到一个视频拍的高铁窗外的景色	2
6	实时播报：【贵州贵阳今日发现7例核	0
7	zeenew崽崽一定要快快好起来！！！	2
8	哈哈哈好突然就要开始一个人的隔离	0
9	隔离点的大妈们好八卦呀 好有趣的人	1
10	隔离日记	1
11	隔離Day 3干飯人干飯魂#刘雨昕2060元	0
12	我的快乐源泉你妈，都来看，因为疫	0
13	网友来信：你好，我不在新疆，但我	0
14	当代逃避生活表现："最近有点想去隔	0
15	好像那样想像的隔离那么严	1
16	前两天跟朋友聊到了太阳的后裔 想到	1
17	#2022小事谱# 2022.10.8 星期二 晴早	1
18	我真无语了?我亲爱的妈沫刚被村里通	0
19	隔离结束?绿码恢复 正常恰饭	2
20	铁窗泪 终于浅浅见了一面大哥隔离前	1
21	抑郁症uu们网上问诊有用吗。我快要	0
22	高中时代，想要走遍地理书上的山川	1
23	真有你的kth 不能因为在隔离就不怕被	0
24	#debris share to you# 第一天和最后一	2
25	真的不懂健康码是怎么赋黄码的，在	0
26	广东省 武汉江夏区火车高铁到东莞要	1
27	#南宁疫情防控#我先问下，从南昌到	0
28	隔离期间真的很担心自己的精神状态	0
29	原神代肝 原神交易真是人麻了，隔	2
30	我他妈恨死了 终于等到延安不一刀切	0
31	#川外被曝18元盒饭不断出现异物#川	1
32	妈蛋 互联网女儿居家隔离第一天 想他	1

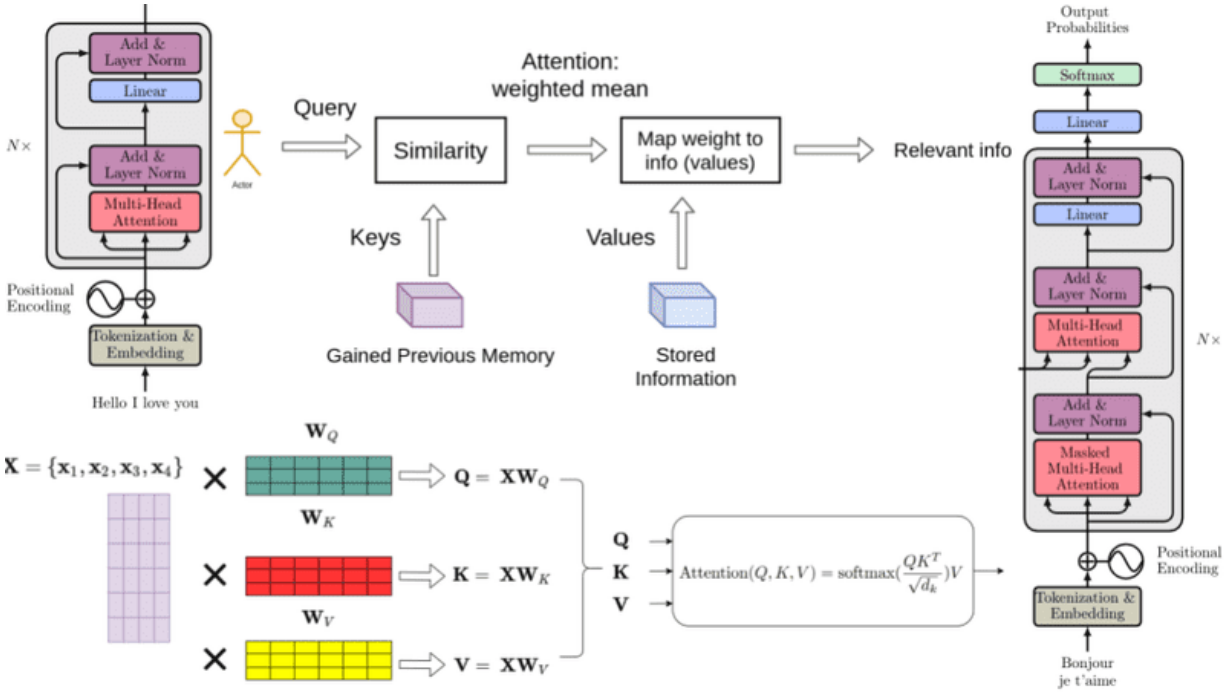
Figure 7: Weibo data from China on November 23, 2022

Note for figure 7, the content showing in the graph is just for visualization and in actual training we only keep a record of Weibo_ID.Sentiments label added to Weibo are done by artificial recognition.Sentiment label we are using 0 for negative, 1 for neutral and 2 for positive.

3.2 Classification Model

A. Transformers

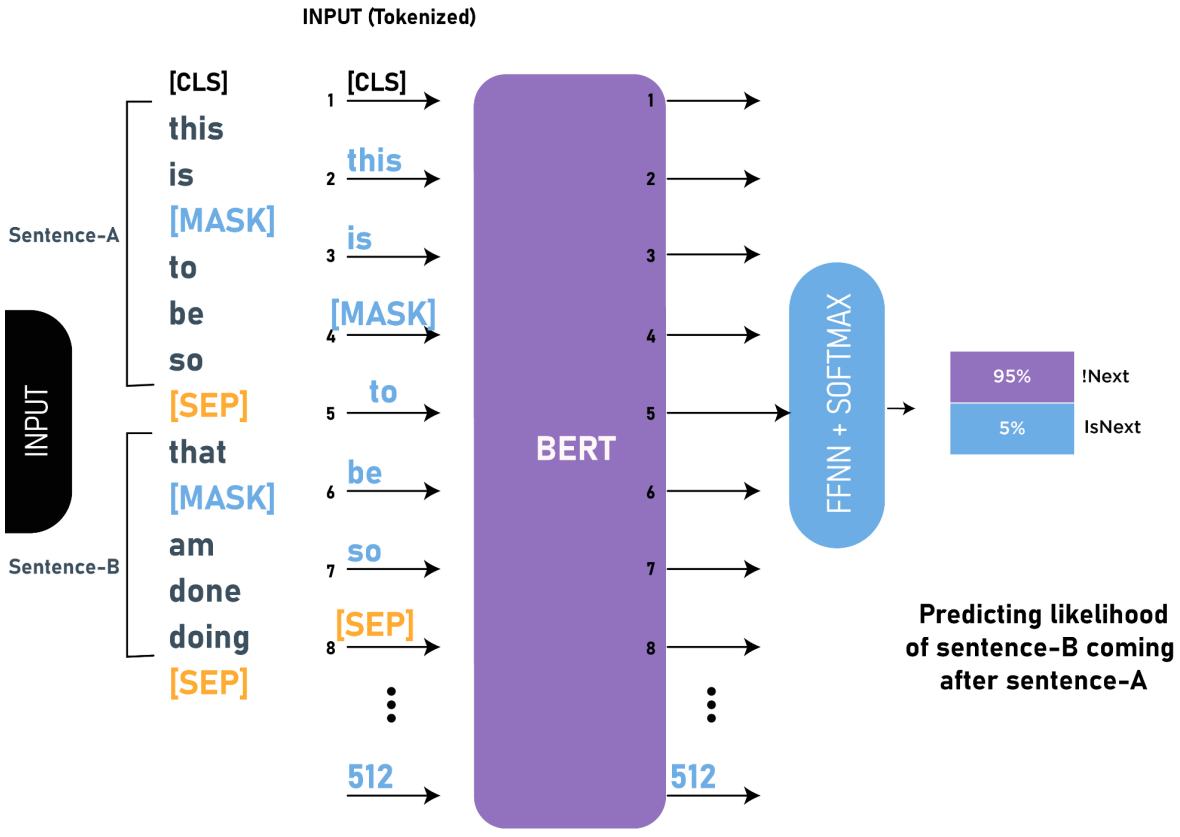
The Transformer in NLP is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. The Transformer was proposed in the paper, Attention Is All You Need. The Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution. “transduction” means the conversion of input sequences into output sequences. The idea behind Transformer is to handle the dependencies between input and output with attention and recurrence completely.



B. Bidirectional Encoder Representations from Transformers (BERT)

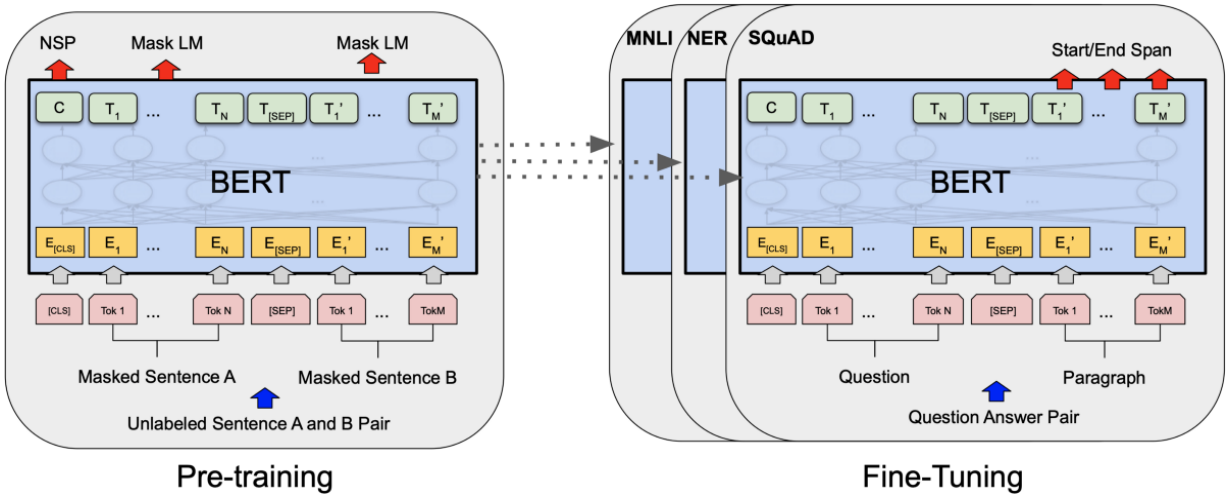
We used Bidirectional Encoder Representations from Transformers (BERT) models for our task. BERT is an open source machine learning framework for natural language processing (NLP). BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context. Historically, language models could only read text input sequentially -- either left-to-right or right-to-left -- but couldn't do both at the same time. BERT is different because it is designed to read in both directions at once. This capability, enabled by the introduction of Transformers, is known as bidirectionality. Although these models are competent, the Transformer is considered a significant improvement because it doesn't require sequences of data to be processed in any fixed order, whereas RNNs and CNNs do. Because Transformers can process data in any order, they enable

training on larger amounts of data than ever was possible before their existence. This, in turn, facilitated the creation of pre-trained models like BERT, which was trained on massive amounts of language data prior to its release. BERT is a semi-supervised learning algorithm, which fits exactly what our project dataset and task. Our project is using Covid Twitter BERT which trained largely on Covid-19 data for Twitter and Chinese-BERT which speciality trained using chinese for Weibo.



C. Pre-trained Model

Both of Covid Twitter BERT and Chinese BERT are pre-trained models which is a saved network that was previously trained on a large dataset, typically on a large-scale image-classification task. And can use the pretrained model as is or use transfer learning to customize this model to a given task. In our case, use the Covid Twitter BERT on Twitter data on sentiment analysis responding to UK announced lockdown and use Chinese BERT on Weibo data on sentiment analysis responding to China Ürümqi fire.



D. Fine-Turing

In deep learning, it is necessary to continuously train and update the parameters (weights) of the model in the deep network to fit a model that can achieve the expected results.

However, when training in a deep neural network, due to the large model size and large amount of parameters, there will be the following problems:

1. The calculation is time-consuming and will take up a lot of computing resources and time costs;

2. For more complex tasks, such as target recognition tasks, if there are more target categories, if you want to improve the performance of the model, you need a large amount of data sets. Also taking the target recognition task as an example, we need a large amount of labeled image data to train the model;

However, there is still a problem. Still take the target recognition task as an example. Suppose there is a trained model A whose task is to recognize (cat, dog, person, chicken, duck, goose) these 6+1 (background) categories When our needs change and we need to add another type of target "pig", if we use the method of retraining a new model B, it will undoubtedly increase the cost and cause a waste of resources - the model requirements of A and B are similar High degree, why can't we use the mature model A?

The solution to the above problems is fine-tune. For example, for the above example, a fine-tuning strategy that can be adopted is to retain the structure of the first few layers of model A and their trained weights, and then change the softmax of the last layer of the model to adjust its mapping to (cat, dog, human, chicken , duck, goose, pig) plus background to form the eight categories, which greatly reduces the training time and computational cost.

In fact, fine-tune can also be understood in this way: our goal is to minimize the prediction loss, and find the optimal point (or a point close to the optimal point) in the

space expanded by each parameter. If you start from the beginning, it is of course slower; But starting from other similar models that have been trained before is equivalent to starting at a point near the optimal point, and the speed and effect of natural convergence will be much better than training from zero.

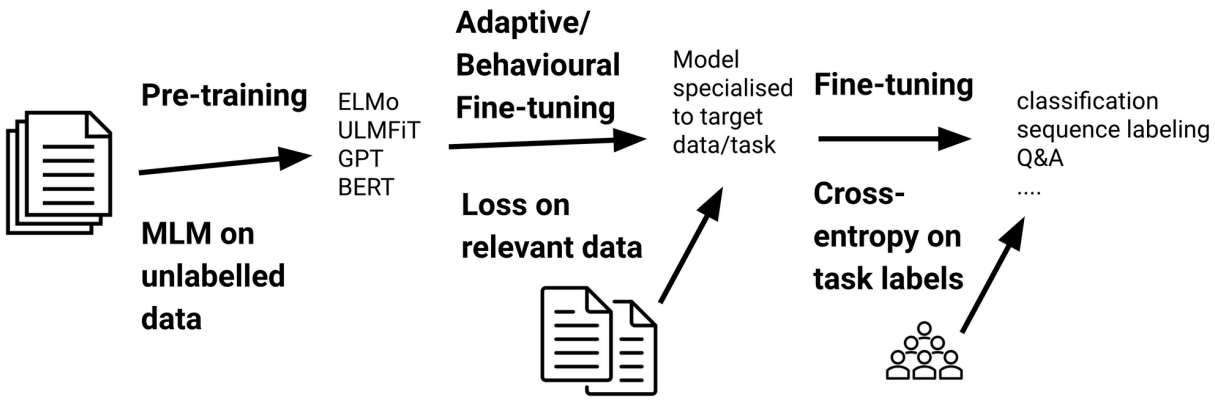
There are different fine-tuning methods for different datasets:

The first type: the amount of data is small, but the similarity is high. In this case, we only need to modify the last few layers or only the output category of the final softmax layer.

The second type: the amount of data is small, and the data similarity is low. In this case, we can freeze the initial layer of the pre-training model (assumed to be K layers), and then train the remaining layers (assuming that there are N layers in total, then the remaining layers are N - K layers). Because the new data has low similarity to the original data, it is more meaningful and efficient to train higher layers on the new dataset.

The third type: the amount of data is large, and the data similarity is low. In this case, the most recommended way is to train your neural network from scratch.

The fourth type: large amounts of data and high similarity. This is the most ideal situation. We only need to keep the architecture of the original model and the initial weights of the model. The model is then retrained on the new data using the weights from the saved pre-trained model.



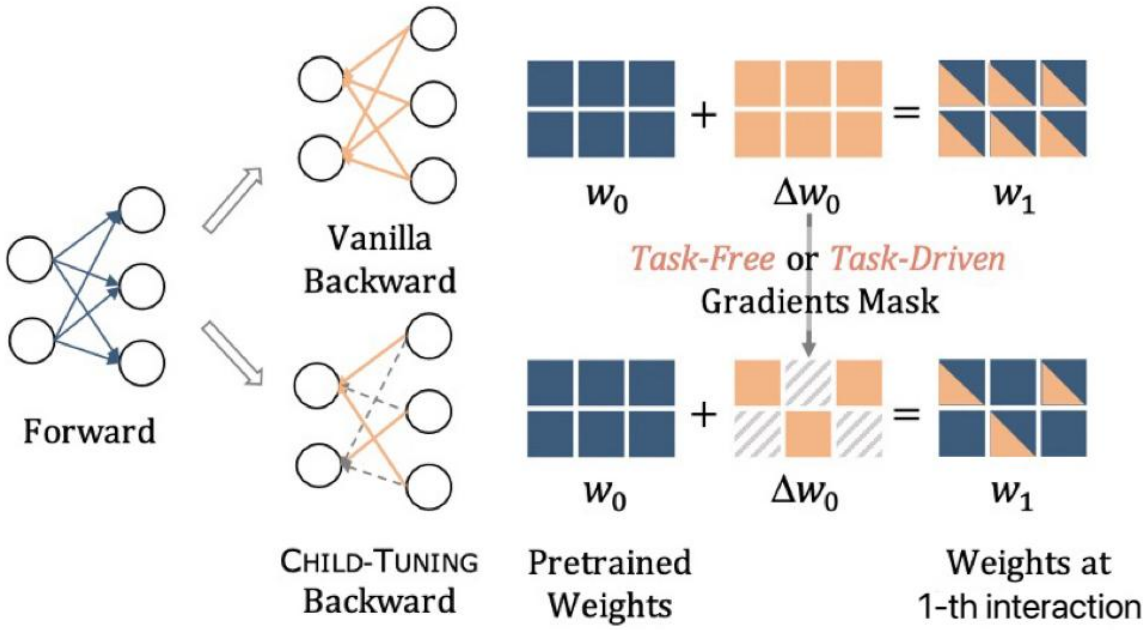
Fortunately, our new dataset has a very high similarity to the original dataset. For COVID-Twitter-BERT, our data also comes from the same platform, twitter, but the data comes from different countries. The pre-training data for COVID-Twitter-BERT is US Twitter data about Covid, while our live data is about the UK. For Chinese-Bert, its pre-training data is a large number of Chinese paragraphs with good emotions. And our current data comes from posts about Covid on Weibo, with emotions marked. But the new data set we provide must not be as comprehensive and complete as the original data set. In the Fine-tuning process, we want to use the powerful knowledge provided by the large-scale pre-training model on the one hand, and on the other hand, we want to solve the mismatch problem between "massive parameters" and "few labeled samples". Can we use this method? to solve the problem? In forward, keep the same as normal Fine-tune, use the parameters of the entire model to encode input samples; when updating parameters in backward, there is no need to adjust a large number of huge parameters, but only a part of them, that is, a Child in the network Network.

Child-Turning:

Step1: Find and confirm the Child Network in the pre-training model, and generate the corresponding Gradients 0-1 Mask of Weights;

Step2: After the backward propagation calculates the gradient, only the parameters in the Child Network are updated, while other parameters remain unchanged.

The whole process is shown in the figure below:



Only update the parameters of Child Network through Gradients Mask

In the two "sub adjustment" steps mentioned above, step 2 is relatively simple to update only the parameters in the "sub network". We can do this by gradient mask, that is, after calculating the gradient of each parameter position, multiply it by the gradient mask of 0-1 matrix. The location of the parameter in the subnet corresponds

to 1, and the location of the parameter that does not belong to is 0. These parameters will be updated later.

The key is how to identify the children's network mentioned in step 1? In this project, we used Child Tuning_ D. It is related to downstream tasks and can adaptively perceive the characteristics of downstream tasks. Its strategy of selecting Child Network can adaptively adjust for different downstream tasks, and select the most important parameters related to downstream tasks to serve as the Child Network. It introduces the Fisher Information Matrix (FIM) to estimate the importance of each parameter to the downstream task, and, consistent with previous work, approximately uses the diagonal matrix of FIM (that is, assumes that the parameters are independent of each other) to calculate the importance score of each parameter relative to the downstream task, and then selects the parameter with the highest score as our Child Network.

$$\mathbf{F}^{(i)}(\mathbf{w}) = \frac{1}{|D|}\sum_{j=1}^{|D|}\left(\frac{\partial \log p\left(y_j|\mathbf{x}_j;\mathbf{w}\right)}{\partial \mathbf{w}^{(i)}}\right)^2$$

Child-Tuning_ D Determine Child Network by calculating Fisher Information of parameters

Algorithm 1 CHILD-TUNING for Adam Optimizer

Require: $\mathbf{w}_0$: initial pretrained weights; $\mathcal{L}(\mathbf{w})$: stochastic objective function with parameters $\mathbf{w}$; η: learning rate; $\beta_1, \beta_2 \in [0, 1)$: exponential decay rates for the moment estimates;

initialize timestep $t \leftarrow 0$, first moment vector $m_0 \leftarrow 0$, second moment vector $v_0 \leftarrow 0$
while not converged **do**
  $t \leftarrow t + 1$
  // Get gradients
  $\mathbf{g}_t \leftarrow \frac{\partial \mathcal{L}(\mathbf{w}_t)}{\partial \mathbf{w}_t}$
  // Get task-free/task-driven child network
  $\mathcal{C}_t \leftarrow \text{GetChildNetwork}()$
  // Generate a corresponding gradient mask
  $\boldsymbol{M}_t \leftarrow \text{GenerateMask}(\mathcal{C}_t)$
  // Employ mask for gradients
  $\mathbf{g}_t \leftarrow \mathbf{g}_t \odot \boldsymbol{M}_t$
  $\mathbf{m}_t \leftarrow \beta_1 \cdot \mathbf{m}_{t-1} + (1 - \beta_1) \cdot \mathbf{g}_t$
  $\mathbf{v}_t \leftarrow \beta_2 \cdot \mathbf{v}_{t-1} + (1 - \beta_2) \cdot \mathbf{g}_t^2$
  // Bias correction
  $\hat{\mathbf{m}}_t \leftarrow \mathbf{m}_t / (1 - \beta_1^t)$
  $\hat{\mathbf{v}}_t \leftarrow \mathbf{v}_t / (1 - \beta_2^t)$
  // Update weights
  $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \eta \cdot \hat{\boldsymbol{m}}_t / (\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon)$
end while
return $\mathbf{w}_t$

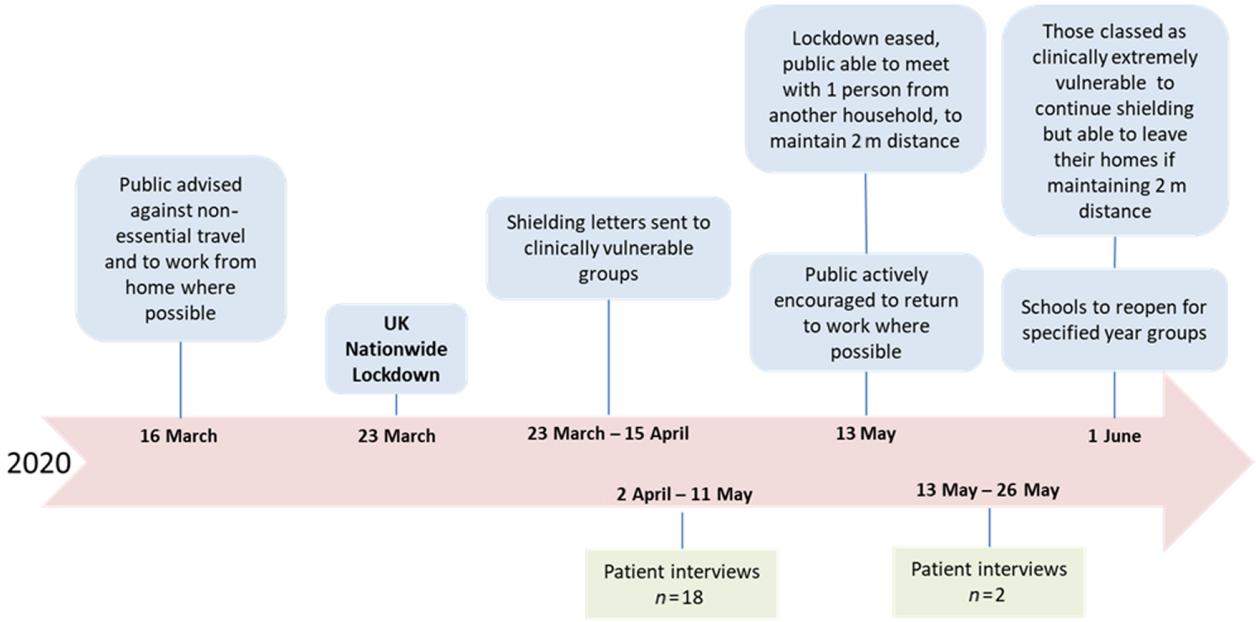
the pseudo-code of CHILD-TUNING when applied to widely used Adam (Kingma and Ba, 2015) optimizer

When we applied child tune to our new data, the accuracy of Covid Twitter Bert increased from 84.4% to 88.9%, and that of Chinese bert increased from 83.7% to 88.3%.

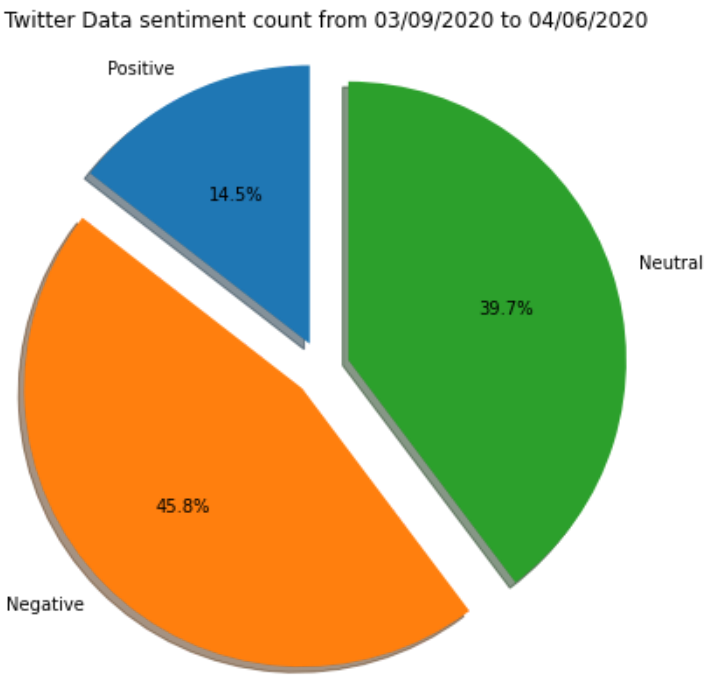
4.RESULTS

4.1 Twitter Data

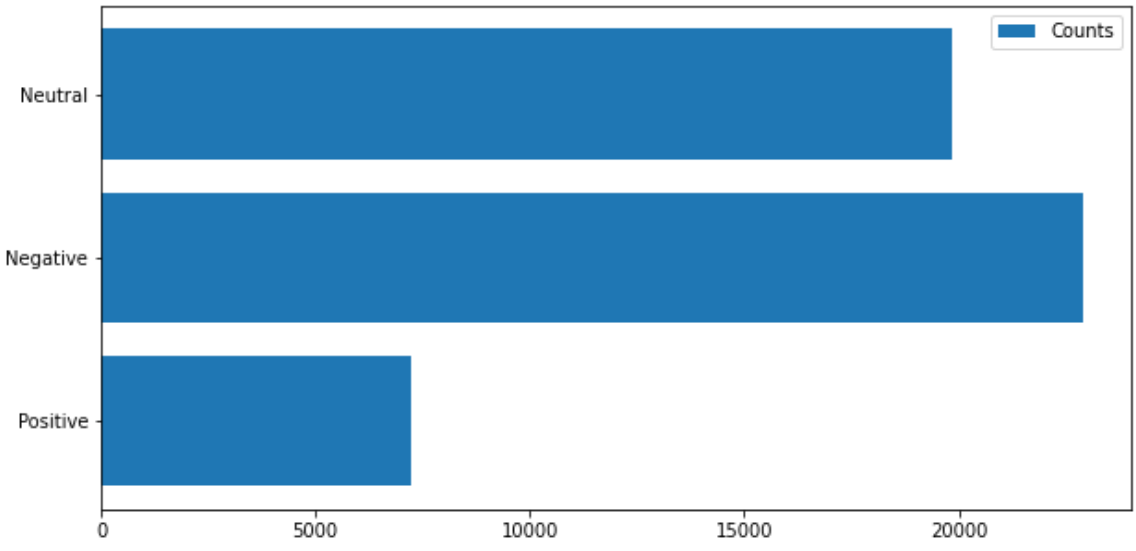
Since we have twitter dataset from github, and the only limitation for choosing data from existing dataset is we want a total tweet number to match the range of Weibo number so we can avoid data unbalanced problems. We have about 48k Weibo data from 11/10/2022 to 12/8/2022. Thus we decided to randomly select 50k tweets from a 80k dataset from UK Twitter for the period of 03/09/2020 to 04/07/2020. The selected method is by using random seed without placements. Below figure shows the timeline of UK government local down.

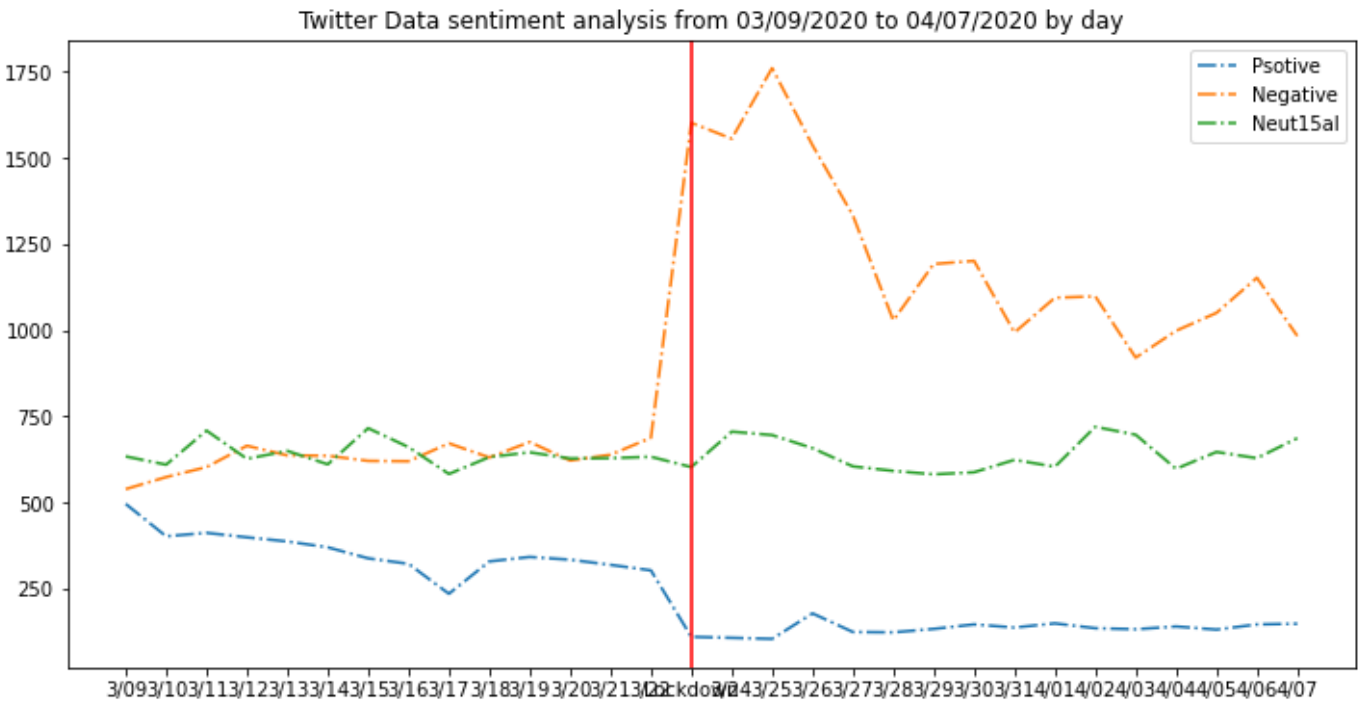


UK data total have 5,0000 Twitter and by our model the detailed principal component analysis are shown in the figure below.



For UK Twitter from 03/09/2020 to 04/07/2020, total negative tweets number are 22884, neutral tweets number are 19865, and positive tweets number are 7251.

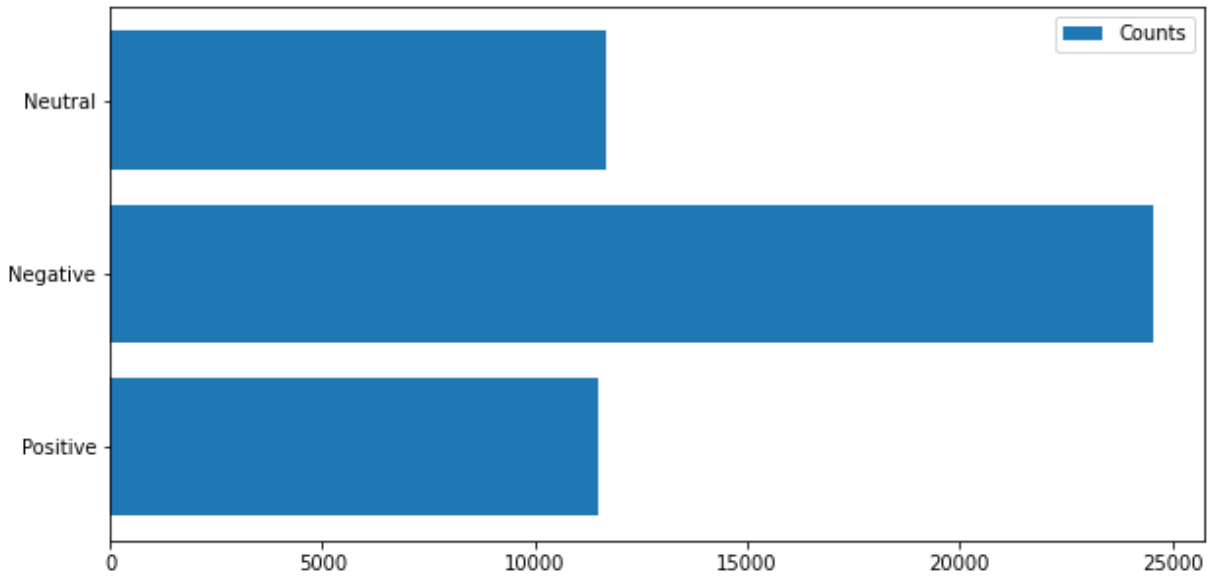




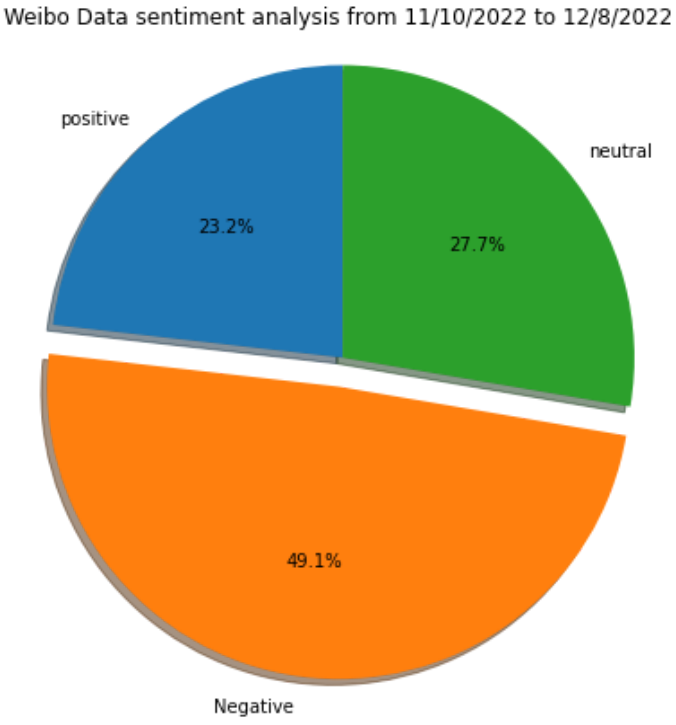
Above figure shows sentiments on social media related to keyword covid-19 over one month period center of event UK announced lockdown.

4.2 Weibo Data

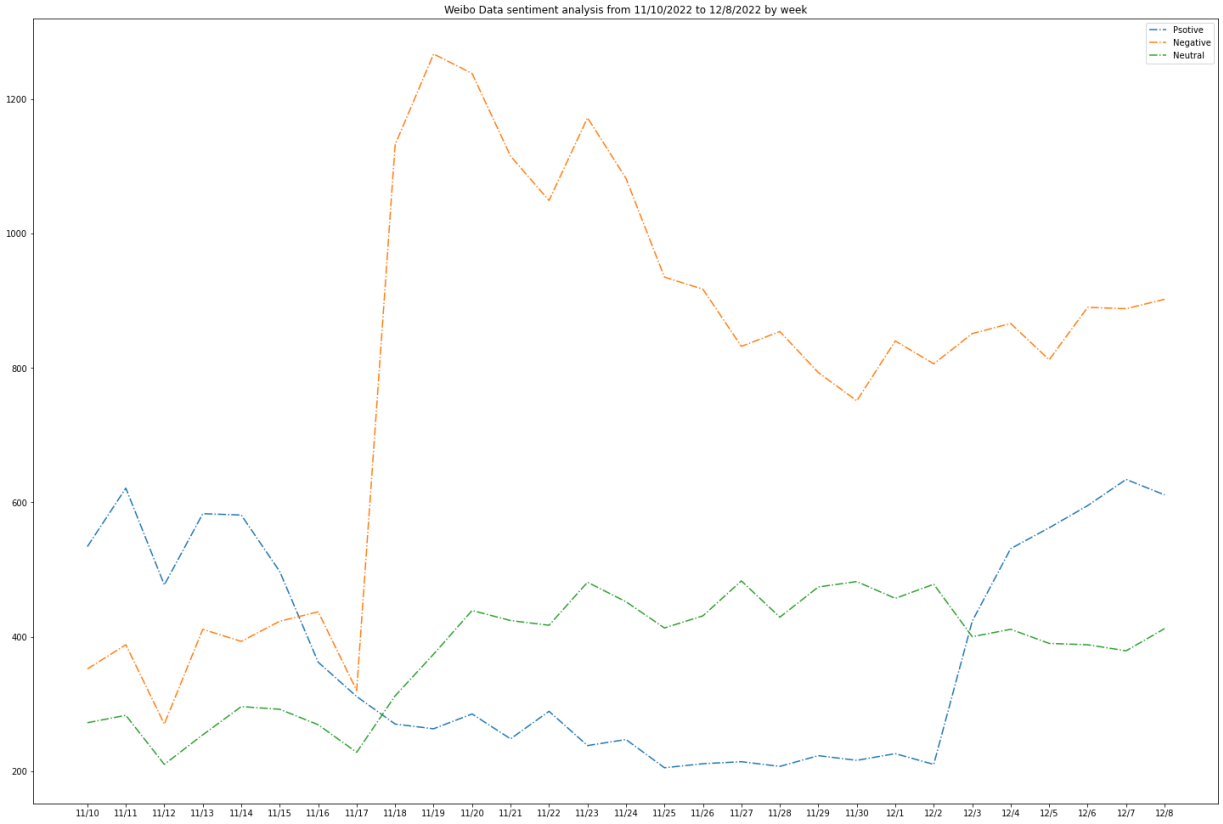
Through crawling from Weibo, we got about 48k Weibo data from 11/10/2022 to 12/8/2022. Total negative tweets number are 24534, neutral tweets number are 11663, and positive tweets number are 11484.



The detailed principal component analysis is shown in the figure below.



The following figure shows sentiments on social media related to keyword covid-19 over a one month period center of event China Ürümqi fire.



4.3 Comparison

From UK data, we can clearly see that the posts of negative emotions occupy a dominant position, while the positive emotions only occupy a small part of the total, and the posts of neutral emotions only account for a little less than those of negative emotions. From the data of China, we can clearly see that the posts of negative emotions are dominant, while the posts of positive emotions and neutral emotions only account for a small part of the total. Posts of positive emotions are slightly less than those of neutral emotions.

Obviously, under the influence of Covid-19, people in both countries are more negative. The proportion of negative emotions in the UK is 45.8%, while that in China is 49.1%. However, the UK's neutral mood is significantly higher than China's. The UK's Neutral mood is 39.7%, while China's is 27.7%. We speculate that this is because the severity of the events in the countries with two disasters has a lot to do with it. The lockdown of the UK only affects the convenience of people's lives, interpersonal communication and the profitability of the company. And the Urumqi fire in China caused the tragic death of real life. In this case, it is easier for UK people to maintain a normal attitude, while it is very reasonable for Chinese people to be more radical.

At the same time, we can find from China's data that negative sentiment showed a downward trend in the days after the fire broke out, which increased tremendously. We suspect this is an attempt by the government to salvage public sentiment. The Chinese government does not want this to have an undue impact. What’s more, we can find that China's positive sentiment has slightly improved after 12/2. After research, it was found that this should be because the Chinese government announced the policy of canceling the health code. But there was also a slight increase in negative sentiment. This should be due to subtle concerns about opening up to Covid-19 while being grateful that life will become more convenient. This has also been confirmed in recent news from China. Because China opened up to Covid-19, people's infection rate started to rise. More and more people are infected, and more and more people die from the infection.

5.FUTURE WORK

A. More Emotion Category

Currently, we only use 3 categories of emotion which are negative, positive and neutral. But we can do more than just 3 different emotions since some human emotions can be shared at the same time. Also in this paper, we did not discuss levels of emotion, we just classified it into 3 big categories. But levels of emotion are also a great indicator. For example, “I am so tired of lockdown” and “If I am still going to lockdown, I will start a fire” both will be classified as negative, however the second one is way more wasted than the first. Which tells it might be something worth looking into.

B. Fusion-based Learning

Another way that can be done in the future is fusion-based learning which means use different models on the text and process of integrating information from multiple sources to produce specific, comprehensive, unified data about an entity. In this way, it can be better suited for the task and hence improve accuracy of the current model.

C. Multimodal Learning

Last but not least, multimodal learning also is very popular and has been considered as one of the further aims of machine learning and deep learning. For our project, we only take the Tweet and Weibo itself as input which is all charters. But when people post on social media a lot of time they will post a picture too. With multimodal learning we can combine nature language processing for the text part and computer vision for

the picture. Which will be better to help to get better sentiment analysis by eliminating ambiguity and determining answers from multiple angles.

REFERENCE

[1] Banda JM;Tekumalla R;Wang G;Yu J;Liu T;Ding Y;Artemova K;Tutubalina E;Chowell G; (n.d.). A large-scale COVID-19 twitter chatter dataset for Open Scientific Research -- an international collaboration.

[2] Lopez, C. E., Gallemore, C., An Augmented Multilingual Twitter dataset for studying the COVID-19 infodemic Soc. Netw. Anal. Min. 11, 102 (2021). DOI: s13278-021-00825-0 https://pubmed.ncbi.nlm.nih.gov/34697560/

[3] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. #COVID-19: The First Public Coronavirus Twitter Dataset. arXiv:cs.SI/2003.07372, 2020

[4] Hongxuan Chen, Columbia University, Cross-Cultural Differences of Sentiment in Social Media Posts Related to COVID-19, 2021

[5] Banda JM;Tekumalla R;Wang G;Yu J;Liu T;Ding Y;Artemova K;Tutubalina E;Chowell G; (n.d.). A large-scale COVID-19 twitter chatter dataset for Open Scientific Research -- an international collaboration.

[6] Karim F, Oyewande AA, Abdalla LF, Chaudhry Ehsanullah R, Khan S. Social Media Use and Its Connection to Mental Health: A Systematic Review. Cureus. 2020 Jun 15;12(6):e8627. doi: 10.7759/cureus.8627. PMID: 32685296; PMCID: PMC7364393.

[7] Marshall C, Lanyi K, Green R, Wilkins G, Pearson F, Craig D Using Natural Language Processing to Explore Mental Health Insights From UK Tweets During the COVID-19 Pandemic: Infodemiology Study JMIR Infodemiology 2022;2(1):e32449

[8] Zhang, T., Schoene, A.M., Ji, S. et al. Natural language processing applied to mental illness detection: a narrative review. npj Digit. Med. 5, 46(2022).https://doi.org/10.1038 /s41746-022-00589-7

[9] Abu Kausar, Mohammad & Dhaka, Vijaypal & Singh, Sanjeev. (2013). Web Crawler: A Review. International Journal of Computer Applications. 63. 31-36. 10.5120/10440-5125.

[10]Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[11]Martin Müller, Marcel Salathé, and Per E. Kummervold. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. arXiv preprint arXiv:2005.07503 (2020).

[12] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[13] Yiming Cui, Wanxiang Che, et al. “Pre-Training with Whole Word Masking for Chinese BERT.” arXiv:1906.08101.

[14] Runxin Xu, Fuli Luo, et al. “Raise a Child in Large Language Model: Towards Effective and Generalizable Fine-tuning.” arXiv:2109.05687.

[15] Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." IEEE transactions on pattern analysis and machine intelligence 41.2 (2018): 423-443.