Weibo Sentiment Analysis During COVID-19

Meichen Feng (A0206451N) [*1] **Shuang Wu (A0206521U)** [*1] **Wanchen Zhao (A0206504R)** [*1]
Shao Tianji (A0206469X) [*1] **Zhao Tiansi (A0206520W)** [*1]

Weibo, Sentiment Analysis, COVID-19

1. Introduction

On 31 December 2019, Wuhan Municipal Health and Health Committee issued an emergency notice on patients with unexplained pneumonia, which was caused by beta coronavirus. Like MERS (Middle East Respiratory Syndrome) and SARS (Severe Acute Respiratory Syndrome), the novel coronavirus pneumonia is also one type of SARI (Severe Acute Respiratory Infection) and has been officially named by World Health Organization (WHO) as COVID-19 [1]. Soon afterwards Chinese national health commission confirmed human-to-human transmission, and the disease has rapidly spread from Wuhan to other areas. The WHO has declared the outbreak to be a public health emergency of international concern in January. In the last four months, the virus has killed more than 179,000 people, seen over 2,600,000 confirmed cases and reached more than 180 countries. In mainland China, there are 84,294 confirmed cases of the coronavirus infection, causing 4,642 deaths from the real-time updates up till 23 April 2020 [2]. And WHO have therefore made the assessment that COVID-19 can be characterized as a pandemic [3]. These below two thermodynamic diagrams indicate the epidemic contribution details worldwide on 12 February and 22 April, which gives a clear comparison of the outbreak scope in the past two months and indicates how severe the coronavirus pandemic is.

2. Problem statement

As the pneumonia epidemic outbreak worsens, fear and panic mood began to spread among people. To prevent the further spread of pneumonia-like illness, a series of measures have led to widespread flight cancellations, shuttered cities, shaken financial markets, and so on. In these affected countries, mainland China is the first place that takes drastic measures to stop the spread of COVID-19. Hubei Province was sealed off from 23 January 2020. Wuhan and other nearby cities have successively announced the suspension of all public transport and temporarily closed exit corridors such as airports, train stations, and expressways to prevent the spread of the disease till 8 April. There is a serious lack of medical resources in Hubei and fever clinics are crowded with patients every day. Masks, disinfectant, medical alcohol and other related products are completely out of stock in many provinces early in the outbreak. On Chinese social media such as Weibo, a number of topics about the expansion of this epidemic appeared on the hot search list. Many netizens expressed concern about the surge in cases and fear of the disease. Negative sentiments and rumours began to appear on social network sites. People's emotions have not been relieved and PTSD (post-traumatic stress disorder) [5] may occur in such a situation. Therefore, we would like to conduct analysis on Chinese netizens' sentiment toward COVID-19 during the worst period of this outbreak. We believe that this analysis could help people face mental health problems, relieve anxiety and panic. While new cases have slowed in China, containment efforts have so far failed to ease the increase of new confirmed cases in many countries, like U.S. and Italy, as the virus continues to spread globally, which has been shown in below figures. This research may also help the affected governments better understand human thoughts and behaviours, and realize what people really worried about during the period of pandemic, such as shortage of supplies and medical overload. Governments can better prepare for the pandemic and improve the credibility.

3. Data Collection and Overview

For the purpose of our research aim, defining the scope of the data is our foremost priority and cannot be performed perfunctorily. As we want to observe and analyze the change in sentiment of Chinese netizens towards COVID-19, the most challenging component comes down to how we collect a sufficient amount of data while ensuring the quality and representativeness of such samples. Following this objective, we decided to scrape our own data from one of China's largest social media platforms, Sina Weibo.

Sina Weibo is a Chinese microblogging website, constituting a major business line of its parent firm, Sina. With less

[*]Equal contribution [1]School of Computing, National University of Singapore, Singapore. Correspondence to: Meichen Feng (A0206451N) < .

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020. Copyright 2020 by the author(s).

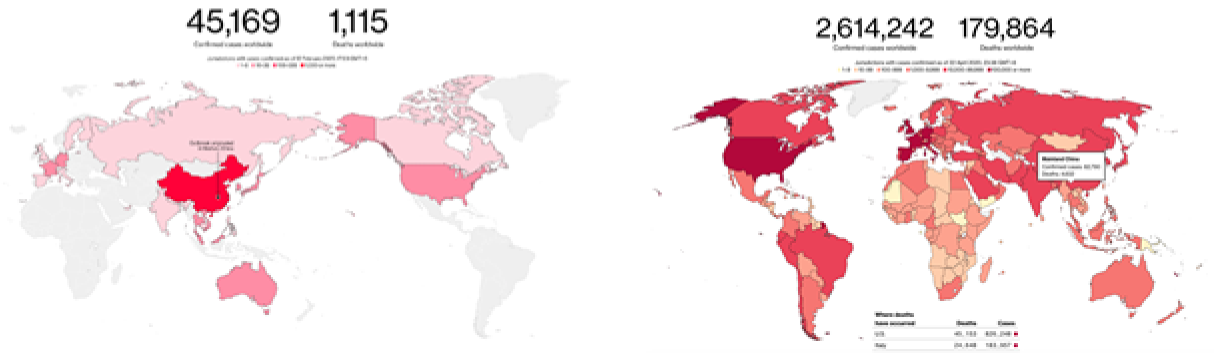

Figure 1. Confirmed cases and deaths worldwide as of 12 February and 22 April 2020, 17:00 GMT+8 [4]

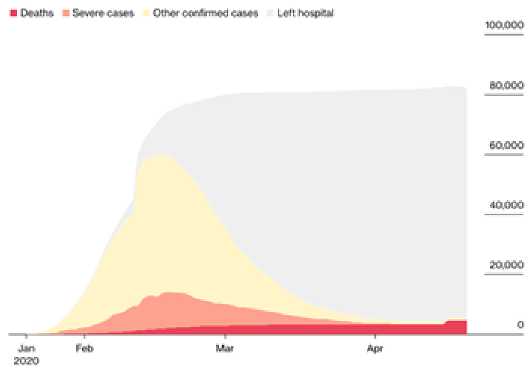

Figure 2. Rise in confirmed cases in mainland China since January 2020

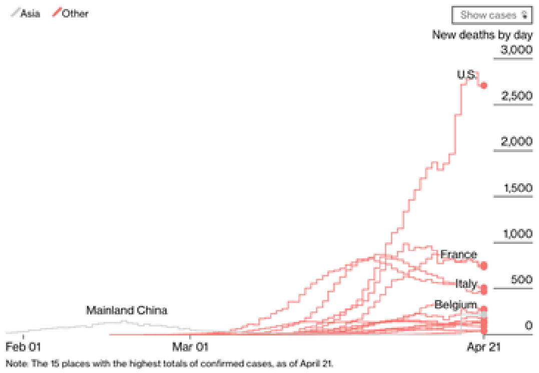

Figure 3. New deaths by day of 15 places with the highest totals of confirmed cases, as of 22 April 2020, 17:00 GMT+8 [4]

than 11 years of history, Sina Weibo was able to retain 445 million monthly active users in 2018[6]. This measure suggests that Sina Weibo is a platform with great exposure to different voices over the world, and it possesses the suitable amount of data that we require.

The Weibo website has a total of three addresses and choosing the appropriate one is essential in obtaining quality data. The three addresses are https://weibo.cn, https://m.weibo.com, and https://weibo.com. The first address is the simplest website to scrape from; however, there are limitations on the scraping amount (where only 1000 tweets can be scraped per day for one keyword). Meanwhile, the second website doesn't support advanced search API so keywords or time frames cannot be filtered. Therefore, we decided to scrape from the last address, https://weibo.com, as it has an advanced search interface, allowing us to customize the search direction and enabling us to scrape a maximum of 24000 posts per day that aligns with our data collection objectives.

During the course of data collection, we employed an open source software, WeiboSpider, and it supports the customization of filters and automation of the whole process. Along with WeiboSpider, we also utilized other web-scraping packages and storage softwares such as scrapy, redis, phantomjs, selenium, and MongoDB to supplement other functions. The keyword that we chose as search filters was 'COVID-19' in Chinese ('新冠肺炎'). This is the official name of COVID-19 in China and is widely used by both official bodies and citizens. The time frame is also critical as seasonable posts are likely to reflect the true sentiments. As a result, we scraped posts that are between Jan 20th, 2020 to Mar 9th, 2020. This is from the time when pandemic just started to breakout massively in China to the time when COVID-19 had already gone viral worldwide. We believed that this time span will capture the sentiment fluctuations that are indicative of the general Chinese netizens.

These two filters were applied to the scraping process and two main datasets were collected, Post Details and User Information. The former dataset contained information about Weibo posts or reposts that are related to the keyword 'COVID-19' ('新冠肺炎') and are within the time frame specified above. It has a total of 40824 observations and 11 features. The latter contained detailed information about various users who posted original posts and it has 1993 observations and 12 features.

The Table 1 and Table 2 provide detailed descriptions of the features of the two acquired datasets.

4. Data Cleaning and Preprocessing

These two datasets will be utilized for different kinds of analysis. The Post Details dataset will be used for sentiment analysis of the impact of COVID-19 on Chinese Netizens. The User Information dataset will be used for geographical analysis of various sentiments among different provinces and cities. In order to better tailor the datasets to the needs of various analyses, some of their features were trimmed to only retain the ones with crucial importance.

The Table 3 shows the features for our final datasets.

For data preprocessing, we removed meaningless posts which are the ones that contained only the word 'repost', '转发微博' (repost in Chinese) and the ones that have no content at all. Additionally, we also removed posts that contained special characters, such as '@' and 'http://'. These posts are unlikely to provide additional insights or useful information to us in the analysis stage. Therefore, they are trimmed to reduce the amount of noise and to ensure that the analysis could be carried out without interference.

5. Sentiment Analysis

The analysis section will be divided into two parts: modeling and visualization. Different methods will be applied to predict the sentiment orientation of the posts and reposts we scraped. Visualization will then be constructed to uncover potential insights.

Since there are no ready-made labels for training and testing, traditional supervised learning cannot be applied here. There are two solutions that have been frequently used. One is sticking to machine learning methods but using existing labeled corpus from other datasets. The other is simple dictionary matching to calculate the sentiment score.

5.1. Basic and Refined SnowNLP

The most popular machine learning package to predict Chinese text sentiment is SnowNLP. SnowNLP is a powerful tool specialized for Chinese text analysis. It can split words using the character based generative model, delete stop words by itself, and predict sentiment scores for new texts using Bayes classifier trained on positive and negative shopping comments.

A baseline model is constructed using the original SnowNLP package. This model will return the positive probability (sentiment score) of each repost. The closer the score to 0, the more negative the repost is, and vice versa. For the sake of accuracy calculation and model comparison, the sentiment score is further binned into three categories: negative (0=¡score¡0.3), neutral (0.3=¡score¡=0.7), and positive (0.7¡score¡=1).

However, since the original SnowNLP is trained by shopping corpus rather than Weibo corpus, the accuracy of predictions is doubtful. Therefore, SnowNLP needs to be refined. Since our dataset is scraped from Weibo, our group first used original SnowNLP to predict the sentiment scores of 5,000,000 Weibo comments and then extracted extreme positive (¿0.8) and negative comments (¡0.3) as new training corpus[7]. After obtaining the new training corpus, the Bayes classifier is re-trained based on the new corpus to predict the sentiment scores again.

Another refinement to the basic SnowNLP model is replacing the word splitting method embedded in SnowNLP by Jieba package, since Jieba package has a more efficient word scanning based on Tier Tree. It can also apply dynamic programming to split words based on highest frequency and deal with new words by the Hidden Markov Model. After refining the model and getting sentiment scores, the sentiment score is also binned into negative, neutral, positive three classes.

5.2. Dictionary Matching

Another popular sentiment analysis method is dictionary matching. The positive, negative, internal negation, and adverb of degree dictionary are first prepared as corpus[8]. Then the Jieba package is employed to split words. Stop words are also deleted later. The final sentiment score will be calculated based on word frequency appearing in different dictionaries. For example, if a word appears in the positive dictionary, the positive score will increase; If the word prior to this word is in the adverb of degree dictionary, the score will be adjusted; If the word prior this word is in the internal negation dictionary, negative score will increase. Judgement for negative words is similar to positive words. Compared to machine learning methods, this approach is less dependent on labelled corpus and has a wider scope of application range when corpus is limited. However, this matching approach is not flexible since text meaning is complex and the accuracy is usually low.

In order to compare methods trialed above, our group randomly selected 3000 reposts from the original dataset and

Table 1. Description of Features for the Post Details Dataset.

Features	Description
_id	The unique identification of the user who posted this content given by WeiboSpider
crawl_time	The time used to crawl this post since the beginning of the process
weibo_url	The web address of this particular content
user_id	The unique identification of the user who posted this content given by Sina Weibo
created_at	The time when this content was created
tool	The kind of internet browser or brand of cell phone that this content was posted through
like_num	The number of likes that this content has received
repost_num	The number of reposts being made on this content
comment_num	The number of comment that this content has received
image_url	The image web address if the content contains an image
content	The original text of the content

Table 2. Description of Features for the User Information Dataset.

Features	Description
_id	The unique identification of the user who posted this content given by WeiboSpider
crawl_time	The time used to crawl this post since the beginning of the process
nick_name	The profile name that the user used within Sina Weibo
gender	The gender of this user
province/city	The location information of this user
brief_introduction	A brief introduction provided by the user
vip_level	The level of VIP if the user has purchased such service
authentication	The titles given to the user by Sina Weibo based on, but not limited to the user's popularity and the area of interest
labels	The labels that are self-selected by the user to indicate interests
tweets_num	The total number of posts (both original and reposts) created by the user
follows_num	The total number of users that this user is following
fans_num	The total number of users that are following this user.

Table 3. The List of Features for the Final Datasets.

Post Details	User Information
user_id	_id
created_at	crawl_time
content	nick_name
like_num	gender
post_num	province
comment_num	num_follower
	brief_introduction
	birthday
	vip_level

manually labeled them. After labeling, these labeled reposts were used to calculate evaluation metrics and compare models' performance. Table 4 shows that the performance of SnowNLP and Dictionary Matching is not satisfactory. The highest accuracy is only 0.44, and the highest F1-score is only 0.41.

5.3. Machine Learning Using Labeled Data

Since the performance of employing other corpus to predict our dataset is not satisfactory, our group decided to utilize the subset of data we labeled to train different machine learning models, and choose the one with best performance for final prediction. This approach will return us pseudo labels, which is not as accurate as traditional supervised learning, since we only have a small amount of labeled data but with a large amount of unlabeled data. However, since the subset we labeled is randomly selected, it's reasonable to infer that the labeled and unlabeled data would have the same distribution. Moreover, they are scraped using the same keyword, we have reason to believe that the accuracy

Table 4. Model Comparison (SnowNLP & Dictionary Matching)

Metrics Basic	SnowNLP	Refined SnowNLP	Dictionary Matching
Accuracy	0.40	0.39	0.44
Precision	0.44	0.45	0.51
Recall	0.40	0.39	0.44
F1-score	0.36	0.36	0.41
Comments	Good at predicting positive reposts (positive class has the best F1-score).	Better at predicting negative reposts than model 1, but the overall performance is worse.	Good at predicting extreme sentimental reposts (F1-score is 0.5 for positive and negative classes).
Performance	Moderate	Worst	Best

and credibility of this method would be higher than others, and can provide us some useful insights.

Similar as before, the dataset was first cleaned by deleting memos, url, and punctuations, then words would be split by Jieba package. All split words would be transformed into using the TF IDF method. After vectorization, 3000 labeled reposts were divided into train set and test set (2400:600). Different machine learning methods were trailed based on 2400 training reposts and validated on 600 testing reposts. The comparison results are shown in the table below. Log loss, accuracy, precision, recall and F1-score were chosen as evaluation metrics. From the table we can see that the best performance is obtained from Bagging Classifier (based on logistic regression). After choosing the best model, we re-trained this model based on 3000 reposts and applied this model to predict the remaining 33655 reposts. The results will be used to do visualization analysis in the following part.

6. Findings and Insights

As mentioned in the project proposal, there are several potential explorable aspects in the data we collected. During the COVID-19 outbreak, there have been some critical moments in the timeline of the epidemic's progression, such as rising death tolls, border controls, number of countries with confirmed cases rising, city-wide quarantines, progressions in medical research and vaccine development, just to name a few. Therefore, the first aspect we want to explore is the correlation between these critical events and Weibo users' sentiment fluctuations.

Figure 1 is the sentiment timeline derived from the model result. The higher the sentiment score, the more positive the sentiment. It can be seen that the sentiment fluctuated quite a bit in the period from mid-January to early March. Moreover, there are several peaks and low-points on the line. After cross-referencing with the timeline of critical events, we conclude that correlation exists to a certain extent between critical events and user's sentiment on Weibo.

On Jan 23rd, Hubei lockdown started, resulting in about 11 million people quarantined in Wuhan and over 57 million in fifteen other cities. This time point corresponds to the first low point in the graph. Negative sentiments prevailed at that time expressing concerns and worries. The following day Jan 24th was Chinese new year' eve. There was an increase in the sentiment score. On Jan 25th, the actual day of Chinese new year, the sentiment kept at a higher level. On Jan 28th, WHO report said the risk of the virus was “very high in China, high at the regional level and high at the global level, which may cause the low point on the sentiment line. On Feb 3rd, a small peak may coincide with the event of completion of Wuhan Huoshenshan hospital. On Feb 6th, Chinese doctor Li Wenliang died from coronavirus contracted from an infected patient at age 33. He was considered a whistleblower of the possible outbreak of a disease similar to SARS and was summoned by the police for ‘making false comments on the internet’. Social media users thought highly of him and his death triggered significant anger and grief, which corresponds to the low point on the sentiment timeline on Feb 6th. It is not clear what may have caused the sentiment low point on Feb 15th. On Feb 26th, news from Japan that 891 tested positive for corona-virus and Olympic may delay due to the current situation, which may be the possible reason for another low point on the line at 02-26 point.

The second aspect is the homogeneity of sentiment nationwide. As we can access the province in which the Weibo user is located. We want to examine whether users from different regions exhibit consistent sentiment across the country or there exist levels of differentiations across provinces and regions. Model results are shown in Figure2. Colour towards red represents more positive sentiment and colour towards green represents more negative sentiment. The whole country exhibited a similar level of positive sentiment. However, sentiments of Xizang, Taiwan and Hainan Province are extremely different from the rest of the country, where Xizang and Taiwan displayed rather negative sentiment and Hainan showed highly positive sentiment.

Table 5. Model Comparison (based on labeled dataset)

Model Name	**Log Loss**	**Accuracy**	**Precision**	**Recall**	**F1_score**
Naive Bayes	1.765	0.66	0.68	0.66	0.65
Logistic Regression	1.633	0.67	0.68	0.67	0.67
SVM	1.918	0.63	0.64	0.63	0.63
XGBoost	2.094	0.66	0.66	0.66	0.65
Random Forest	1.156	0.44	0.76	0.44	0.27
Bagging Classifier (based on logistic regression)	**1.599**	**0.67**	**0.68**	**0.67**	**0.66**
LightGBM	2.724	0.64	0.64	0.64	0.64
AdaBoost	1.109	0.60	0.61	0.60	0.60

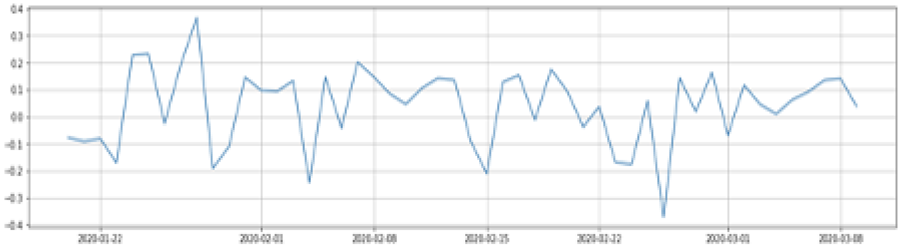

Figure 4. Sentiment timeline

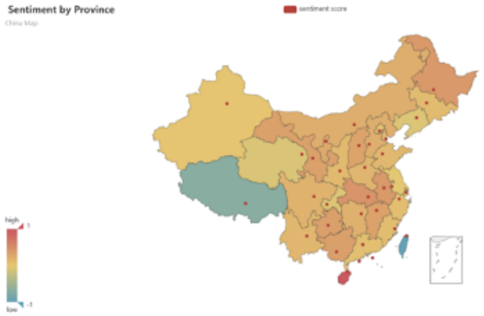

Figure 5. Sentiment across country

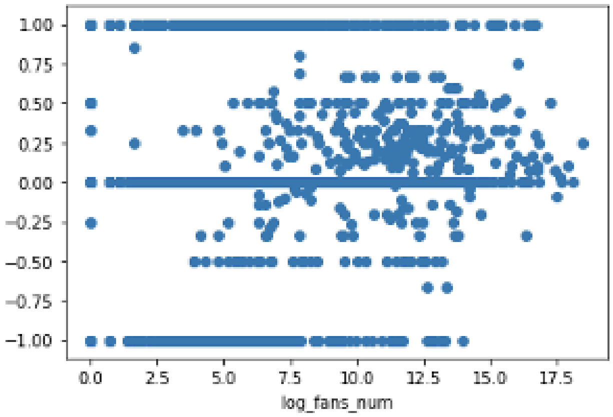

Figure 6. Sentiment vs. user popularity

Another aspect is sentiment's relationship with users' levels of activeness and popularity. In the scraped dataset, there is information about the user's number of posts, which acts as a proxy of level of activeness, as well as the user's number of followers. Therefore, we could investigate whether sentiments differ in terms of user activeness and popularity and whether more popular users tend to spread positivity or even negativity. Figure3 shows the result of sentiment versus user popularity. X-axis is the logged number of fans a user has. There doesn't seem to be a clear trend or relationship between sentiment and user popularity. Users of all levels of popularity show various sentiment, positive, negative and neutral. However, users with a really large amount of followers have less extreme sentiment.

7. Limitations

7.1. Data volume and quality

Weibo imposes a limit on the amount of data that one account can scrape using Weibo Spider. After the account reaches the limit, permission would be denied. Therefore, we managed to get around 3,600 Weibo posts, which is not necessarily representative of the overall sentiment, given the issue's extremely high popularity and the massive Chinese Weibo user population.

What's more, the data quality is not satisfactory as well. Weibo is a casual platform and users express freely, with no limitations on the posts' shapes and forms. This includes the usage of many emojis. In the Chinese web culture, some emojis could represent many different sentiments. One of the most typical is the smiley face 😀 which is known for being able to express anything from joy to sarcasm and mockery. Another typical one is the praying hand emoji 🙏. This is commonly used to express respect towards front-line medical workers, unity to fight against the virus, but also depression and the desperate desire that the epidemic could be over soon. These are two examples of emojis that the machine learning models cannot properly classify.

Data preprocessing could eliminate these emojis, but it may also eliminate key information. For example, the 'thumbs-up' emoji 👍 may not be recognized by Weibo Spider and may be decoded as [赞] ('compliments'). But if we were to consider this in data preprocessing, it could also take out the same character in sentences like "为湖北点赞" ('respects to Hubei'). This could render the sentence meaningless, and change an obviously positive sentiment to neutral.

In addition, our Weibo Spiders also scraped a lot of irrelevant posts because many users attach COVID-19 related hashtags to posts of completely irrelevant contents to gain popularity.

7.2. The study is of machine learning without labels

Unlike classroom examples or binarily classified real-life data, social media data don't come with official sentiments classifications or labels. This imposes challenges to the proper train-test data set split, as well as to criteria of assessing and comparing different machine learning models, because we cannot calculate reliable accuracy scores.

Our solution was to manually label some posts and use them as the train data set. We randomly chose 10% of the posts and labeled them into 3 categories: 1 (positive sentiments; expressions of encouragements and support to those in a hard time, calling for unity to fight against the virus together, and respects towards front-line workers); 0 (neutral sentiments; mostly highly objective news reports on number of newly confirmed cases, public policies around the country and around the world); -1 (negative sentiments; expressions of racism and sarcasm towards other countries and regions, expressions of quarantine depressions, dissatisfaction towards hospitals' service qualities, and sorrow over the loss of medical workers).

But manual labeling has obvious drawbacks. Different group members have different understandings and labeling criteria over the same posts, each by their own intuitions. And since this process was highly subjective, it was difficult to set a very specified common standard. What's more, it's very common that one post could express several different sentiments, some of which positive and others negative by our standards.

It's also very common that one post could have multiple sentiments. The following is a typical example:

你好，明天武昌医院院长刘智明因患新冠肺炎，不幸殉职。倒在战疫一线，一个最不愿看到的糟糕消息，以我生命，守护生命，燃尽一生，只因我是医生。谁对尘世不曾怀有眷恋？谁不想长久陪伴家人？而他们一着白袍，即为战士，苌弘碧血，令人感佩。且用更显著的抗议成效告慰他们，英雄一路走好！

Hello, tomorrow Liu Zhiming, the chief of Wuchang Hospital lost his battle with COVID-19. Medical workers perishing on the front line is the last kind of news we want to see. They sacrifice their own lives to save others, only because they're doctors. Who doesn't want to live a long life and who doesn't want to have more time with their families? Yet once they put on their white uniforms, they are the most courageous warriors of this battle. It's truly respectable, especially since they've achieved so much. Rest in peace, our hero!

This post expresses both sorrow and regret over the loss of Doctor Liu, which is classified as negative sentiment by our standard; but at the same time it also expresses respect towards medical workers and their generous contributions, which counts as positive by our standards. Users freely express on Weibo, and it's natural that one post could have mixed sentiments.

8. Conclusion

In order to analyze the sentiment orientation on Chinese social platforms during the outbreak, we have applied SnowNLP, dictionary matching, and other machine learning methods to predict the sentiment scores for Weibo netizens. According to the analysis, we are able to conclude that sentiment fluctuations corresponded with the current events that took place in real life. Various incidents imposed a great impact on the Chinese netizens and these influences were in turn reflected on the posts on Weibo. While most provinces showed a nationwide sentiment homogeneity, provinces like Xizang, Hainan, and Taiwan exhibited some extreme cases. This could potentially be attributed to their isolated geographical locations which prevented them from achieving consensus. For the relationship between user popularity and sentiment, no clear correlation was observed among the general public. However, it should be noted that popular users tend to express neutral opinions that prevent them from guiding the public towards extreme perspectives.

As mentioned in Limitations, our project still faces drawbacks from data volume, data quality, and research methods. Further actions should be taken to improve those mechanisms, including, but not limited to, crawling and labelling more data, applying more machine learning methods, and conducting in-depth research on different kinds of emotions (anger, sneer, fear, etc.)

Complete code and data can be found here. https://github.com/BT5153-Group-Seventeen/Weibo-Sentiment-Analysis-During-COVID-19

References

[1] Dario Thuburn, AFP. WHO Has Finally Named The New Coronavirus. ScienceAlert. 12 Feb 2020. [Online] Available at: https://www.sciencealert.com/who-has-finally-named-the-deadly-coronavirus

[2] Novel Coronavirus Pneumonia Epidemic Real-time Updates. Sina News.13 Feb 2020. [Online] Available at: https://news.sina.cn/zt_d/yiqing0121

[3] WHO Director-General's opening remarks at the media briefing on COVID-19. World Health Organization. 11 March 2020. [Online] Available at: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020

[4] Cedric Sam, Chloe Whiteaker and Hannah Recht. Mapping the Coronavirus Outbreak Across the World. Bloomberg. 12 Feb 2020. [Online] Available at: https://www.bloomberg.com/graphics/2020-wuhan-novel-coronavirus-outbreak/

[5] Posttraumatic stress disorder. Wikipedia. [Online] Available at: https://en.wikipedia.org/wiki/Posttraumatic_stress_disorder

[6] Bylund, A. Weibo Added 15 Million Users in Q3. [Online] Available at: https://www.fool.com/investing/2018/11/29/weibo-added-15-million-users-in-q3.aspx

[7] Budao. Utilizing 5,000,000 Weibo Comments to Analyze Sentiment. [Online] Available at: https://zhuanlan.zhihu.com/p/30061051

[8] Chinese Sentiment Analysis Based on Python & Jieba. [Online] Available at: https://blog.csdn.net/qq_41185868/article/details/84864905