Weibo-COV: A Large-Scale COVID-19 Social Media Dataset from Weibo

Yong Hu†, Heyan Huang†, Anfan Chen‡, Xian-Ling Mao†
†Beijing Institute of Technology
{huoyong, hhy63, maoxl}@bit.edu.cn
‡University of Science and Technology of China
caf16@ustc.edu.cn

Abstract

With the rapid development of COVID-19, people are asked to maintain “social distance” and “stay at home”. In this scenario, more and more social interactions move online, especially on social media like Twitter and Weibo. People post tweets to share information, express opinions and seek help during the pandemic, and these tweets on social media are valuable for studies against COVID-19, such as early warning and outbreaks detection. Therefore, in this paper, we release a large-scale COVID-19 social media dataset from Weibo called Weibo-COV, covering more than 30 million tweets from 1 November 2019 to 30 April 2020. Moreover, the filed information of the dataset is very rich, including basic tweets information, interactive information, location information and retweet network. We hope this dataset can promote studies of COVID-19 from multiple perspectives and enable better and faster researches to suppress the spread of this disease.

1 Introduction

At the beginning of writing, COVID-19, an infectious disease caused by a coronavirus discovered in December 2019, also known as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has infected 4,517,399 individuals globally with a death toll of 308,515 (Doctor, 2020). Under the circumstances, the physical aspects of connection and human communication outside the household among people are limited considerably and mainly depend on digital device like mobile phone or computers (Abdul-Mageed et al., 2020). In this kind of scenario, people will stay at home and spend more time on the social media communication. The social media plays an important role for people sharing information, expressing opinions and seeking help (Lopez et al., 2020), which makes social media platforms like Weibo, Twitter, Facebook and Youtube a more crucial sources of information during the pandemic.

In the previous studies, social media was considered as a valuable data source for studies against disease, like uncovering the dynamics of an emerging outbreak (Zhang and Centola, 2019), predicting the flu activity and disease surveillance (Jeremy et al., 2009). For example, some studies facilitate better influenza surveillance, like early warning and outbreaks detection (Kostkova et al., 2014; De Quincey and Kostkova, 2009), forecast estimates of influenza activity (Santillana et al., 2015) and predict the actual number of infected cases (Lampos and Cristianini, 2010; Szomszor et al., 2010). Therefore, it is necessary to make the relevant social media datasets freely accessible for better public outcomes to facilitate the related studies of COVID-19.

In this paper, we release a large-scale COVID-19 social media dataset from Weibo, one of the most popular Chinese social media platform. The dataset is named Weibo-COV and covers more than 30 million tweets from 1 November 2019 to 30 April 2020. Specifically, unlike the conventional API method to retrieve data, which limit large data access, we first build a high-qulity Weibo active user pool with 20 million active users from over 250 million users, then collect all active users’ tweets during the time period and filter out tweets related to COVID-19 by selected 179 keywords. Moreover, the fields of tweets in the dataset is very rich, including basic tweets information, interactive information, location information and retweet network. We hope this dataset can promote studies of COVID-19 from multiple perspectives and enable better and faster researches to suppress the spread of this disease.
2 Data Collection

2.1 Collection Strategy

At present, given specified keywords and a specified period, there are two methods for constructing Weibo public opinion datasets: (1) advanced search API given by Weibo; (2) Traversing all Weibo users, collecting all their tweets in the specified period, and then filtering tweets with specified keywords.

However, for method (1), due to the limitation of the Weibo search API, the result of once search contains up to 1000 tweets, making it difficult to build large-scale datasets. As for method (2), although we could build large-scale datasets with almost no omissions, traversing all billions of Weibo users requires very long time and large bandwidth resources. In addition, a large number of Weibo users are inactive, and it makes no sense to traverse their homepages, because they may not post any tweets in the specified period.

To alleviate these limitations, we propose a novel method to construct Weibo public opinion datasets, which can build large-scale datasets with high construction efficiency. Specifically, we first build and dynamically maintain a high-quality Weibo active user pool (just a small part of all users), and then we only traverse these users and collect all their tweets with specified keywords.

2.2 Weibo Active User Pool

As shown in Figure 1, based on initial seed users and continuous expansion through social relationship, we first collect more than 250 million Weibo users. Then we define that Weibo active users should meet the following 2 characteristics: (1) the number of followers, fans and tweets are all more than 50; (2) the latest tweet is posted in 30 days. Therefore, we can build and dynamically maintain a Weibo active user pool from all collected weibo users. Finally, the constructed Weibo active user pool contains 20 million Weibo users, accounting for 8% of the total number of weibo users.

2.3 COVID-19 Tweets Collection

According to the collection strategy described in section 2.1, we set the time period from 0:00 December 1, 2019 (GMT+8, the date of the first diagnosis) to 23:59 April 30, 2020 (GMT+8), and design a total of 179 COVID-19 related keywords. These keywords are comprehensive and rich, covering related terms such as coronavirus and pneumonia, as well as specific locations (e.g., “Wuhan”), drugs (e.g., “remdesivir”), preventive measures (e.g., “mask”), experts and doctors (e.g., “Zhong Nanshan”), government policy (e.g., “postpone the reopening of school”) and others (see Appendix.1 for the complete list).

Specifically, based on 20 million Weibo active user pool, we first collect a total of 569,829,866 tweets posted by these users in the specified period. Subsequently, we filter these tweets by keywords and finally obtain 33,519,644 tweets. These tweets constitute our final dataset.

3 Data Properties

3.1 Data Structure

As shown in Table 1, fields of tweets in the dataset is very rich, covering the basic information (_id, crawl_time, content), interactive information (like_num, repost_num, comment_num), location information (origin_weibo, geo_info), and others.
formation (geo_info) and retweet network (origin_weibo). Therefore, various aspects of studies related to infectious diseases can be conducted based on this dataset, such as the impact on people’s daily life, the early characteristics of the disease and government anti-epidemic measures.

### 3.2 Basic Statistic

As shown in Table 2, Weibo-COV contains a total of 33,519,644 tweets, including 895,012 tweets with geographic location information and 6,586,969 original tweets, and the number of deduplication users in the entire dataset is 8,876,036.

### 3.3 Daily Distribution

The distribution of the number of tweets by day is shown in Figure 2. It can be found that from December 1, 2019 to January 18, 2020, the number of COVID-19 related tweets is very small (less than 5K) and may include some noise data. Since January 19, 2020, the number of COVID-19 related tweets increase rapidly and maintain at least 200,000 per day.

Note that the data on April 4, 2020 is particularly striking and the number of tweets on that day exceeds 1.6 million. Because that day is Chinese Tomb Sweeping Festival, and a national mourning was held for the compatriots who died in the epidemic. People posted or reposted a lot of mourning tweets on Weibo on that day.

### 3.4 GEO Distribution

As shown in Figure 3, we plot location distribution of tweets with location information on April
It can be seen that the distribution of tweets spreads all over the world, including major countries in Asia, Europe, Australia and America. Because although Weibo is a Chinese social media platform, with the development of economic globalization, more and more Chinese people go abroad and more and more foreigners start to use Weibo.

Therefore, our dataset can study the impact of the disease on the whole world, not only limit to China.

3.5 Word Cloud

As shown in Figure 4, we select four days of tweets data at different stages of the epidemic development and draw word clouds. It can be seen that people did not know the characteristics of the virus and the government began to take preliminary actions in the early days (e.g. “unexplained pneumonia” and “health committee” in 2019-12-31), then people learned that the virus is a new coronavirus and studied prevention methods and medicines (e.g. “new coronavirus”, “N95 musk” and “ShuangHuangLian” in 2020-01-31), and then the control of COVID-19 became a problem that the whole world needs to face and governments took strict prevention measures (e.g. “isolated at home” and “American COVID-19” in 2020-03-31), and right now virus has not been effectively controlled globally and has had many impacts on people’s lives (“Cirque du Soleil in Canada” in 2020-04-30).

Therefore, our dataset runs through the whole development of COVID-19, and includes impacts of the disease on all aspects of the society.

4 Related Work

Several works have focused on creating social media datasets for enabling COVID-19 research. (Chen et al., 2020), (Lopez et al., 2020) and (Abdul-Mageed et al., 2020) have already released datasets collected from Twitter. However, these datasets are mainly in English. Chinese tweets are also valuable and can provide additional supplements for researches.

Only one dataset proposed by (Gao et al., 2020) includes tweets from Weibo, but their method based on Weibo advanced search API, so they can not collect large-scale tweets from Weibo. Compared with our dataset, their overall size (less than 200K), time span (from January 20, 2020 to March 24, 2020), and number of keywords (only 4 keywords) are all much smaller.

5 Conclusion

In this paper, we release Weibo-COV, a first large-scale COVID-19 tweets dataset from Weibo. The dataset contains over 30 million tweets covering from 1 November 2019 to 30 April 2020 and each tweet with rich field information. We hope this dataset could promote and facilitate related studies on COVID-19.
References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Dinesh Pabbi, Kunal Verma, and Rannie Lin. 2020. Mega-cov: A billion-scale dataset of 65 languages for covid-19.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Covid-19: The first public coronavirus twitter dataset.

Ed De Quincey and Patty Kostkova. 2009. Early warning and outbreak detection using social networking websites: The potential of twitter. In International Conference on Electronic Healthcare, pages 21–24. Springer.

Dingxiang Doctor. 2020. Covid-19 global pandemic real-time reports. https://ncov.dxy.cn/ncovh5/view/pneumonia.

Zhiwei Gao, Shunsharo Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Naist covid: Multilingual covid-19 twitter and weibo dataset.

Jeremy, Ginsberg, Matthew, H, Mohebbi, Rajan, S, Patel, Lynnette, and Brammer and. 2009. Detecting influenza epidemics using search engine query data. Nature.

Patty Kostkova, Martin Szomszor, and Connie St. Louis. 2014. # swineflu: The use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. ACM Transactions on Management Information Systems (TMIS), 5(2):1–25.

Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In 2010 2nd international workshop on cognitive information processing, pages 411–416. IEEE.

Christian E. Lopez, Malolan Vasu, and Caleb Gallowmore. 2020. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset.

Mauricio Santillana, André T Nguyen, Mark Dredze, Michael J Paul, Elaine O’Nsoesie, and John S Brownstein. 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS computational biology, 11(10).

Martin Szomszor, Patty Kostkova, and Ed De Quincey. 2010. # swineflu: Twitter predicts swine flu outbreak in 2009. In International conference on electronic healthcare, pages 18–26. Springer.

Jingwen Zhang and Damon Centola. 2019. Social networks and health: new developments in diffusion, online and offline. Annual Review of Sociology, 45:91–109.
### A Appendices

#### A.1 Covid-19 Related Keywords

Table 3: The list of selected keywords related to COVID-19

| Keywords      | Translations                                                                 |
|---------------|-----------------------------------------------------------------------------|
| 冠状          | Coronavirus                                                                  |
| Cov-19        | Cov-19                                                                       |
| 新冠          | Coronavirus                                                                  |
| 感染人数      | Infected cases                                                               |
| N95           | N95 Mask                                                                     |
| 大众畜牧野味店 | Dazhong wildlife restaurant                                                  |
| 华南野生市场  | South China wild market                                                      |
| 管轶          | Guan Yi                                                                      |
| 武汉病毒所     | Wuhan Institute of Virology                                                  |
| CDC           | Center for Disease Control and Prevention                                   |
| 中国疾病预防控制中心 | Chinese Center for Disease Control and Prevention                              |
| 疫控中心       | Center for Disease Control and Prevention                                   |
| #2019nCoV     | #2019nCoV                                                                    |
| 双黄连 AND 抢购 | Shuanghuanglian AND Rush to buy                                              |
| 双黄连 AND 售罄 | Shuanghuanglian AND Sold out                                                 |
| 武汉卫健委     | Wuhan Municipal Health Committee                                            |
| 湖北卫健委     | Health Commission of Hubei Province                                          |
| #nCoV         | #nCoV                                                                        |
| PHEIC         | PHEIC                                                                        |
| 疫情          | Epidemic outbreak                                                            |
| 火神山         | Huoshen Shan hospital                                                        |
| 雷神山         | Leishen Shan hospital                                                       |
| 钟南山         | Zhong Nanshan                                                                |
| Coronavirus    | Coronavirus                                                                  |
| Remdesivir     | Remdesivir                                                                   |
| 瑞德西韦        | Remdesivir                                                                   |
| 感染 AND 例     | Infected AND cases                                                           |
| 武汉 AND 封城   | Wuhan AND Lockdown                                                          |
| 高福           | George Fu Gao                                                                |
| 王延轶         | Wang Yanyi                                                                   |
| 舒红兵         | Shu Hongbing                                                                 |
| 协和医院       | Xiehe Hospital                                                               |
| 武汉 AND 隔离   | Wuhan AND Quarantine                                                         |
| 李文亮 AND 医生 | Doctor AND Li Wenliang                                                       |
| 云监工         | Supervising work on cloud                                                    |
| 武汉仁爱医院   | Wuhan Ren’ai Hospital                                                       |
| 黄冈 AND 感染者 | Huanggang AND Infected cases                                                 |
| 孝感 AND 感染者 | Xiaogan AND Infected cases                                                   |
| 居家隔离       | Isolated at home                                                             |
| 防护服         | Protective Clothing                                                          |
| 隔离14天       | Isolation AND 14 days                                                        |
| 潜伏期 AND 24天 | Incubation period AND 24 days                                                |
| 潜伏期 AND 14天 | Incubation period AND 14 days                                                |
| 国际公共卫生紧急事件 | International Public Health Emergencies                                       |

Continued on next page
| Keywords                                      | Translations                                                                 |
|----------------------------------------------|-----------------------------------------------------------------------------|
| 方舱医院 AND 武汉                              | FangCang Hospital AND Wuhan                                                  |
| 一省包一市                                   | one province gives a hand to one Hubei city                                 |
| 晋江毒王                                     | Super spreader of COVID-19 in Jinjiang                                        |
| 超级传播者                                   | Super spreader                                                               |
| 湖北 AND 王晓东                              | Hubei AND Wang Xiaodong                                                      |
| 蒋超良                                       | Jiang Chaoliang                                                              |
| 李文亮                                       | Li Wenliang                                                                  |
| 千里投毒                                     | Spread Virus from a thousand miles                                           |
| 武汉病毒研究                                  | Virology research in Wuhan                                                  |
| 武汉 AND 李医生                              | Wuhan AND Li Wenliang                                                       |
| 国家疾控中心                                 | Chinese Center for Disease Control and Prevention                             |
| 武汉 AND 疫苗                                | Wuhan AND Vaccine                                                            |
| 武汉 AND 征用宿舍                             | Wuhan AND Requisitioned students’ dormitory                                 |
| 周佩仪                                       | Zhou Peiyi                                                                   |
| 武汉中心医院                                 | The Central Hospital of Wuhan                                                |
| 张晋 AND 卫健委                               | Zhang Jin AND Health Commission                                              |
| 张晋 AND 卫生将康委员会                      | Zhang Jin AND Health Commission                                              |
| 刘英姿 AND 卫健委                            | Liu Yingzi AND Health Commission                                             |
| 刘英姿 AND 卫生健康委员会                    | Liu Yingzi AND Health Commission                                             |
| 王贺胜 AND 卫健委                            | Wang Hesheng AND Health Commission                                          |
| 王贺胜 AND 卫生健康委员会                    | Wang Hesheng AND Health Commission                                          |
| 复工                                         | Enterprise work resuming                                                     |
| 中小企业 AND 困境                             | Small and medium-sized enterprise AND Dilemma                               |
| 武汉 AND 死亡病例                            | Wuhan AND Death cases                                                       |
| 武汉 AND 感染病例                            | Wuhan AND Infection cases                                                   |
| 湖北 AND 死亡病例                            | Hubei AND Death cases                                                       |
| 湖北 AND 感染病例                            | Hubei AND Infected cases                                                    |
| 中国 AND 死亡病例                            | China AND Death cases                                                       |
| 中国 AND 感染病例                            | China AND Infected cases                                                    |
| 潜伏期                                       | Incubation Period                                                           |
| 北京 AND 病例                                | Beijing AND Cases                                                           |
| 天津 AND 病例                                | Tianjin AND Cases                                                           |
| 河北 AND 病例                                | Hebei AND Cases                                                             |
| 辽宁 AND 病例                                | Liaoning AND Cases                                                          |
| 上海 AND 病例                                | Shanghai AND Cases                                                          |
| 江苏 AND 病例                                | Jiangsu AND Cases                                                           |
| 浙江 AND 病例                                | Zhejiang AND Cases                                                          |
| 福建 AND 病例                                | Fujian AND Cases                                                            |
| 山东 AND 病例                                | Shandong AND Cases                                                          |
| 广东 AND 病例                                | Guangdong AND Cases                                                         |
| 海南 AND 病例                                | Hainan AND Cases                                                            |
| 山西 AND 病例                                | Shanxi AND Cases                                                            |
| 内蒙古 AND 病例                              | Inner Mongolia AND Cases                                                    |
| 吉林 AND 病例                                | Jilin AND Cases                                                              |
| 黑龙江 AND 病例                              | Heilongjiang AND Cases                                                      |
| 安徽 AND 病例                                | Anhui AND Cases                                                             |
| 江西 AND 病例                                | Jiangxi AND Cases                                                           |
| 河南 AND 病例                                | Henan AND Cases                                                             |

Continued on next page
| Keywords                | Translations                     |
|------------------------|----------------------------------|
| 湖北 AND 病例           | Hubei AND Cases                  |
| 湖南 AND 病例           | Hunan AND Cases                  |
| 广西 AND 病例           | Guangxi AND Cases                |
| 四川 AND 病例           | Sichuan AND Cases                |
| 贵州 AND 病例           | Guizhou AND Cases                |
| 云南 AND 病例           | Yunnan AND Cases                 |
| 西藏 AND 病例           | Tibet AND Cases                  |
| 陕西 AND 病例           | Shanxi AND Cases                 |
| 甘肃 AND 病例           | Gansu AND Cases                  |
| 青海 AND 病例           | Qinghai AND Cases                |
| 宁夏 AND 病例           | Ningxia AND Cases                |
| 新疆 AND 病例           | Xinjiang AND Cases               |
| 香港 AND 病例           | Hong Kong AND Cases              |
| 澳门 AND 病例           | Macau AND Cases                  |
| 台湾 AND 病例           | Taiwan AND Cases                 |
| ECOM                   | Extracorporeal Membrane Oxygenation |
| sars-cov-2             | sars-cov-2                       |
| 复学                   | Resumption of schooling           |
| 护目镜                 | Goggles                          |
| 核酸检测               | nucleic acid testing (NAT)       |
| COVID-19                | COVID-19                          |
| 2019-nCoV               | 2019-nCoV                         |
| 疑似 AND 病例           | Suspicious cases                 |
| 无症状                  | Asymptomatic Patients            |
| 累计病例                | Cumulative confirmed cases       |
| 境外输入                | imported cases of NCP            |
| 累计治愈                | Cumulative cured cases           |
| 绥芬河                 | Sui Fenhe                        |
| 舒兰                    | Shu Lan                          |
| 健康码                  | Health QR code                   |
| 出入码                  | Community Access Code            |
| 返校                    | Back to Camp                     |
| 美国 AND 例             | USA cov-19 AND Cases             |
| 西班牙 AND 例           | Spain cov-19 AND Cases           |
| 新加坡 AND 例           | Singapore cov-19 AND Cases       |
| 加拿大 AND 例           | Canada cov-19 AND Cases          |
| 英国 AND 例             | UK cov-19 AND Cases              |
| 印度 AND 例             | India cov-19 AND Cases           |
| 日本 AND 例             | Japan cov-19 AND Cases           |
| 韩国 AND 例             | South Korea cov-19 AND Cases     |
| 德国 AND 例             | Germany cov-19 AND Cases         |
| 法国 AND 例             | France cov-19 AND Cases          |
| 意大利 AND 例           | Italy cov-19 AND Cases           |
| 新增 AND 例             | New cov-19 AND Cases             |
| 人工膜肺                | Extracorporeal Membrane Oxygenation |
| 双盲测试                | Double Blind Test                |
| 疫苗                    | Vaccine                          |
| 小区出入证              | Community Entry card             |

Continued on next page
| Keywords                              | Translations                                                                 |
|--------------------------------------|------------------------------------------------------------------------------|
| 战疫                                 | Anti-COVID-19                                                                |
| 抗疫                                 | Anti-COVID-19                                                                |
| 湖北卫健委 AND 免职                   | Health commission of Hubei Province AND Remove from the position             |
| 发热患者                             | Fever patients                                                               |
| 延迟开学                             | Postpone the reopening of school                                            |
| 开学时间 AND 不得早于                | The start time of school AND Not earlier than                               |
| 累计死亡数                            | Cumulative deaths                                                           |
| 疑似病例                             | Suspicious cases                                                             |
| 入户排查                             | Household troubleshoot                                                       |
| 武汉 AND 肺炎                        | Wuhan AND Pneumonia                                                          |
| 新型肺炎                             | Novel Pneumonia                                                              |
| 不明原因肺炎                        | Pneumonia of unknown cause                                                   |
| 野味肺炎                             | Wildlife pneumonia                                                           |
| 出门 AND 戴口罩                       | Going out AND Wear mask                                                      |
| 3M AND 口罩                           | N95 AND Mask                                                                 |
| KN95 AND 口罩                         | 3M AND Mask                                                                  |
| 新肺炎                               | Novel Pneumonia                                                              |
| #2019nCoV                             | #2019nCoV                                                                    |
| 新型肺炎 AND 死亡                     | Novel Pneumonia AND Death                                                    |
| 新型肺炎 AND 感染                      | Novel Pneumonia Infection                                                   |
| 武汉 AND 肺炎 AND 谣言                | Wuhan AND Pneumonia AND Rumors                                              |
| 8名散布武汉肺炎谣言                  | Eight people AND Spread rumors of Wuhan pneumonia                            |
| 黄冈 AND 新肺炎                      | Huanggang AND Novel Pneumonia                                                |
| 孝感 AND 新肺炎                      | Xiaogan AND Novel Pneumonia                                                  |
| 居家隔离                             | Isolated at home                                                             |
| 武汉中心医院 AND 新型肺炎             | The Central Hospital of Wuhan AND Novel Pneumonia                            |
| 武汉肺炎                             | Wuhan Pneumonia                                                             |
| 企业复工                             | Enterprise work resuming                                                      |
| 囤积口罩                             | Hoarding mask                                                                |
| 零号病人                             | Zero Patient                                                                 |
| 黄燕玲                               | Huang Yanling                                                                |
| 病毒源头                             | Oringin of Cov-19                                                            |
| 电子烟肺炎 AND 新型冠状病毒战          | E-cigarette Pneumonia AND Coronavirus                                        |
| 病毒战                               | Virus War                                                                   |
| 病毒 AND 实验室泄露                   | Virus AND laboratory leakage                                                 |
| 比尔盖茨 AND 疫苗牟利                | Bill Gates AND Vaccine for profit                                            |
| 美国细菌实验室                      | US Army Bacterial Laboratory                                                 |
| 确诊                                 | Confired Infected COV-19 cases                                               |
| pandemic                             | pandemic                                                                    |