Sentiment Analysis and Topic Trending Analysis with Weibo Data

CMPT733 Project Report

Chu Chu, Minyi Huang, Valerie Huang, Yinglai Wang

Abstract

As one of the most popular online social gathering platforms for microblogging in China, “Sina Weibo” (“新浪微博”) has become a rich database to collect Chinese text and has attracted extensive attention from academia and industry. Netizens express their feelings through Weibo, thus generating massive text messages. Through data collection, data processing, model selection, sentiment analysis, and visualization, our project created an extended analysis of the emotional status of netizens on certain topics, opinions on a social phenomenon, and preferences, which not only has a certain commercial value, but also helps to understand societal changes.

1. Motivation and Background

The sentiment trend of the society, which presents what people care about over time and how they feel about the hot topics, can not only give a data analytical solution for business operation or marketing decision, but also reveal the emotional changes and attitudes of social groups and official accounts, as well as user behaviors across different time periods.

a. Data Source: “Weibo”

Weibo, short for “Sina Weibo[1]” (“新浪微博” in Chinese), is an information sharing/microblogging website, similar to Twitter. It is one of the biggest social media platforms in China, with over 600 million monthly active users as of Q3 2019. Users can post information within 140 words or share or repost instantly. It gives Internet users more freedom and convenience to communicate information, express opinions, and record events.

b. Sentiment Analysis

Sentiment analysis refers to the analysis of the emotional state implied by the speaker in conveying information, attitude, judgment or evaluation of his or her opinions. Sentiment analysis on massive text data on Weibo helps us understand the Internet public opinion trend, expand companies’ marketing capabilities, and predict cases of emergency situations.

[1] https://www.weibo.com

c. Motivation

Currently, research on sentiment analysis of Chinese microblogs is just starting. There are many explorations on sentiment analysis of Twitter and other English-language social platforms, but the specific application in Chinese language has certain limitations from a natural language processing perspective, where grammatical rules and language habits are very different.

We are interested to apply the data science techniques we have learned to understand people's opinions, and how it changes and reflects social mood. Our project integrates sentiment analysis and topic modelling into several parts: text preprocessing, information extraction and emotion classification. Text preprocessing includes word segmentation, part-of-speech tagging, and stopping word formation, etc,. Emotional information extraction is based on certain rules extracted from Weibo Unit elements of propensity characteristics; sentiment classification is extracted from the underlying sentiment information. The emotional information extracted is divided into words, topics and relationships, and comes down to sentiment calculation with semantic dictionary and classification based on machine learning.

2. Problem Statement

a. Sentiment analysis

Given a message, how we can decide whether it is of positive, negative, or neutral sentiment.

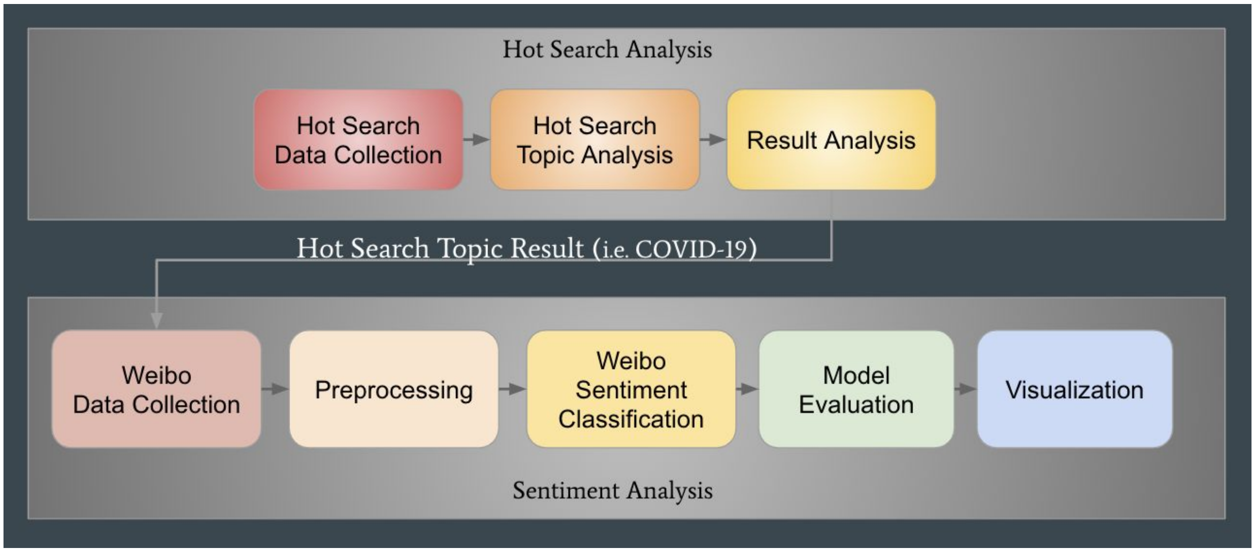
b. Topic Modelling

How people's attitude and focus about certain topic change during a period of time

c. Model Selection

How well will the segmentation and sentimental analysis model perform?

3. Data Science Pipeline


Hot Search Analysis
Hot Search Data Collection
Hot Search Topic Analysis
Result Analysis
Hot Search Topic Result (i.e. COVID-19)
Weibo Data Collection
Preprocessing
Weibo Sentiment Classification
Model Evaluation
Visualization
Sentiment Analysis

4. Methodology

a. Hot search analysis

i. Data collection

Hot Search Topic data collection is done by sending HTTP requests using Weibo API. To align with the weibo post data collection, the time period was also set from Jan 1, 2020 to Mar 26, 2020. We collected roughly 6000 Hot Search Topics in JSON format and extracted information we needed (i.e., time_stamp, content, views, start_time, end_time).

ii. Data preprocess

We used the start time and end time to calculate the alive time in seconds and also extracted start date and end date in the form of 'YYYY-MM-DD'. Duplicated and null hot search contents were removed. When plotting the distribution of the length of alive time and the number of views over time, we normalized these two sets of data to make them more compact.

iii. Hot search words frequency

Unlike English sentences, Chinese sentences are character by character and have no space in between. Topic content was tokenized by using the segmentation tool 'jieba'. We also created certain rules on filtering stop words and meaningless vocabularies.

b. Sentiment analysis

i. Data collection

We crawled the search result of the keywords we obtained from the Hot Search Analysis under domain s.weibo.com/, and scraped the HTML elements as needed using BeautifulSoup.

In order to avoid the user verification disruption, we used Firefox webdriver and implemented a random break mechanism which auto reconnects after time-out. In this way, we made the crawler run on the EC2 instance for a week and gathered about 90,000 weibo posts raw data. Considering the time limit of this project and the amount of work of data labeling, we only used data from Jan 10, 2020 to Mar 26, 2020.

Sentiment mark labeling has to be done before we pre-process the data, otherwise it would be hard to read tokenized content. Four of us spent nearly a month finishing labeling 45,000 weibo posts with sentiment marks.

ii. Data preprocess

Text preprocessing techniques include word segmentation, part-of-speech tagging, syntactic analysis, and other natural language processing. These technologies are relatively mature. Though resources for Chinese NLP are limited, there are several packages and libraries for us to use which include a complete set of XML-based Chinese language processing modules. The application of these has laid a good foundation for our sentiment analysis.

Current open source Chinese word segmentation tools or modules mostly have some comparison data on the closed test set, but this can only show the effect of these word segmentation models on a certain closed test set, and cannot fully explain its performance. According to the characteristics of the microblog text, the link address in the microblog text,

"@" Character (for responding to or communicating with other users) and "#" character (for topics categorization) needs to be filtered.

Even the data was scraped properly from HTML elements, they were still highly unstructured and could not be used to fit in a classification model. All forwarded posts are treated as duplicates and we managed to locate the origin post and kept only that one in the data set.

We also created 26 regular expression rules to filter the post’s content and extract attributes we needed.

Before:

```
729,0,新冠,2020-03-10,//@GonozQvQ://@JaneMere://@诛砂:是有人良心被狗吃了！//@紫飞SAMA:
//@杨林-杨家枪法第六十七代传人:浙江还给新冠患者做肺移植，两例了。这医疗救治强度不敢想象
//@维稳先锋卡菊轮:74万，换算下来就是10万刀多点，再一算，在美国也就够35人次的核算检测。,
03月10日 14:13,0,0,0
```

```
127,2,新冠,2020-02-27,#天津爆料# 【2月27日6时至18时 天津新增1例新冠肺炎确诊病例 累计确诊病
例136例 治愈出院6人】记者从市疾控中心获悉，2月27日6时至18时，天津新增1例新冠肺炎确诊病例
，累计确诊病例136例。今日治愈出院6人，累计治愈出院102例。      第136例患者，男，41岁，为
天津市海河医院呼吸与危重症医学科副主  展开全文c,02月27日 22:59,8,16,18
```

After:

```
729,__label__negative,新冠,2020-03-10,是 有人 良心被狗吃 了 浙江 还给 新冠 患者 做 肺 移植 两例
了 这 医疗 救治 强度 不敢 想象 七十四 万 换算 下来 就是 十万 刀 多点 再 一算 在 美国 也 就 够 三十
五 人次 的 核算 检测,2020-03-10 14:13:00,0,0,0
```

```
127,__label__neutral,新冠,2020-02-27,天津 爆料 二月 二十七日 六时 至 十八 时 天津 新增 一例 新冠
肺炎 确诊 病例 累计 确诊 病例 一百 三十六 例 治愈 出院 六人 记者 从市 疾控中心 获悉 二月 二十七
日 六时 至 十八 时 天津 新增 一例 新冠 肺炎 确诊 病例 累计 确诊 病例 一百 三十六 例 今日 治愈 出院
六人 累计 治愈 出院 一百零二 例      第一百 三十六 例 患者 男 四十一岁 为 天津市 海河 医院 呼吸
与 危重症 医学科 副 主 ,2020-02-27 22:59:00,8,16,18
```

iii. Sentiment classification

We used fastText as the model for efficient learning of word representations and sentence classification. Its major advantages are: it supports multilingual word vectors and multiprocessing during training. Some preliminary results[2] seem to show that fastText embeddings are better than word2vec at encoding syntactic information. The original word2vec model seems to perform better on semantic tasks. By comparing their performance in some other downstream supervision tasks, it will be interesting to see the portability of these two models for different types of tasks.

In our scenario, fastText as a N-gram classification model, has its reputation by Facebook, supports all unicode languages and multi-labeling. It's also fairly easy to tune.

iv. Topics cluster

[2] https://markroxor.github.io/gensim/static/notebooks/Word2Vec_FastText_Comparison.html

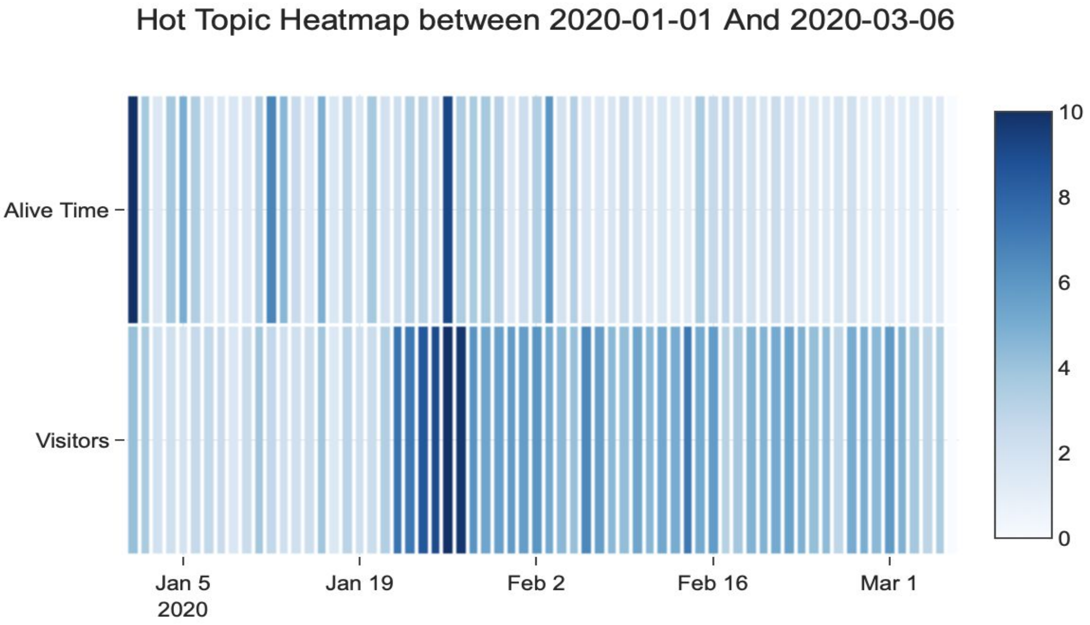
The topics of the Weibo texts were extracted by the Latent Dirichlet Allocation(LDA) model. LDA is a Bayesian probability model to identify the semantic topic information. NLTK is also used for data preprocessing.

After removing the Chinese stopwords, we created the dictionary and corpus needed for topic modeling. In order to get an optimal number of topics for the model, coherence value is used to evaluate the quality of a given topic model. We also built a number of LDA models with different values of the number of topics k and picked the one that gives the high coherence value. Since the coherence score seems to keep increasing, we chose the model that gave the highest CV before flattening out.

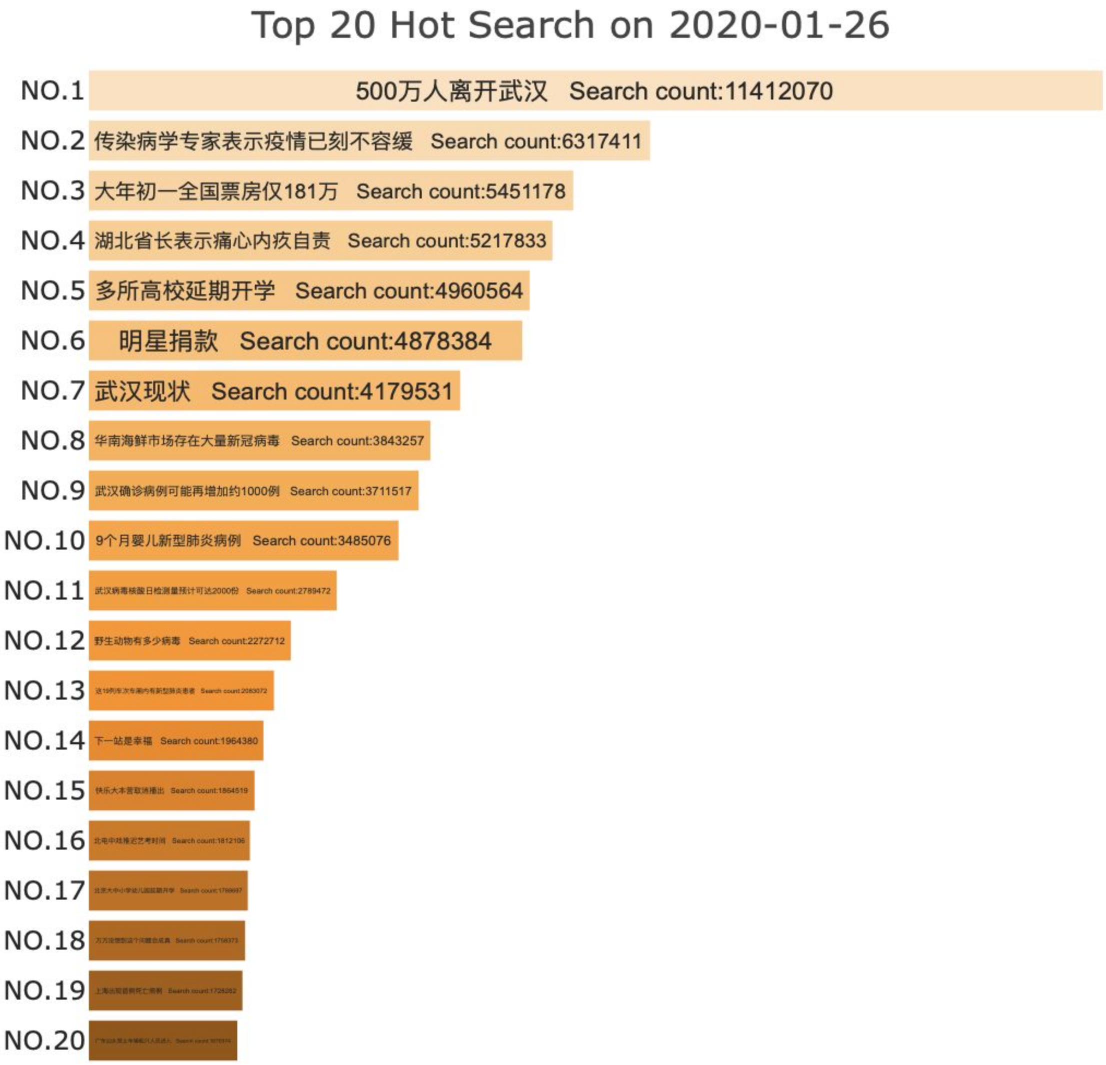
c. Visualization

i. Hot Search

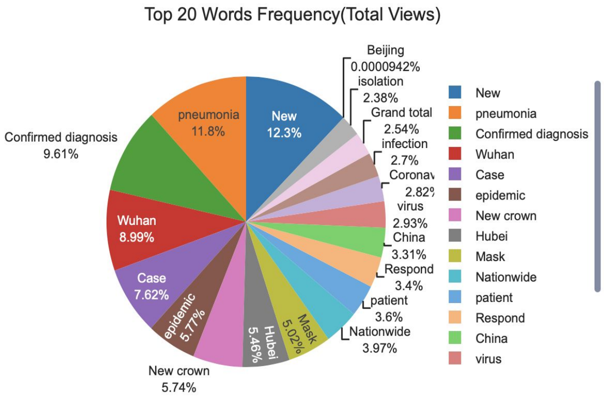
Time series analysis of the Weibo Hot Search was crucial to investigate what people care about most in a period of time. We summed up the total views and alive time every day among three months and plotted a heat map as follows. It is obvious to see that the number of views on the 26th January are much higher than other days, and the total alive time on that day is also relatively long.

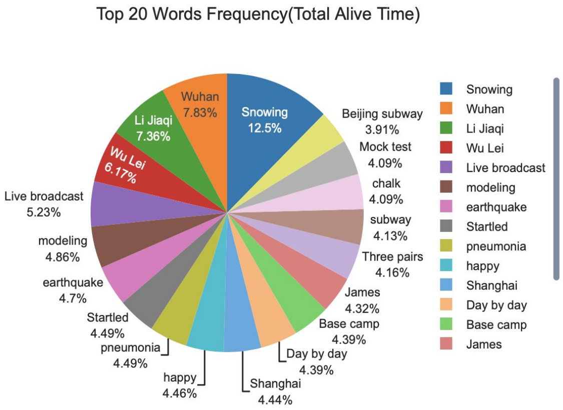


Let us take a close look at what happened on that day. We picked the top 20 hot searches on the 26th of January. Surprisingly, there are 18 out of 20 records directly related to or caused by "coronavirus".



In addition, we segmented the hot search, and selected 20 higher frequency words to make two pie charts by the number of views and the length of alive time. It is not surprising that when the measurement is total views in the first pie chart, all the words are related to "coronavirus".



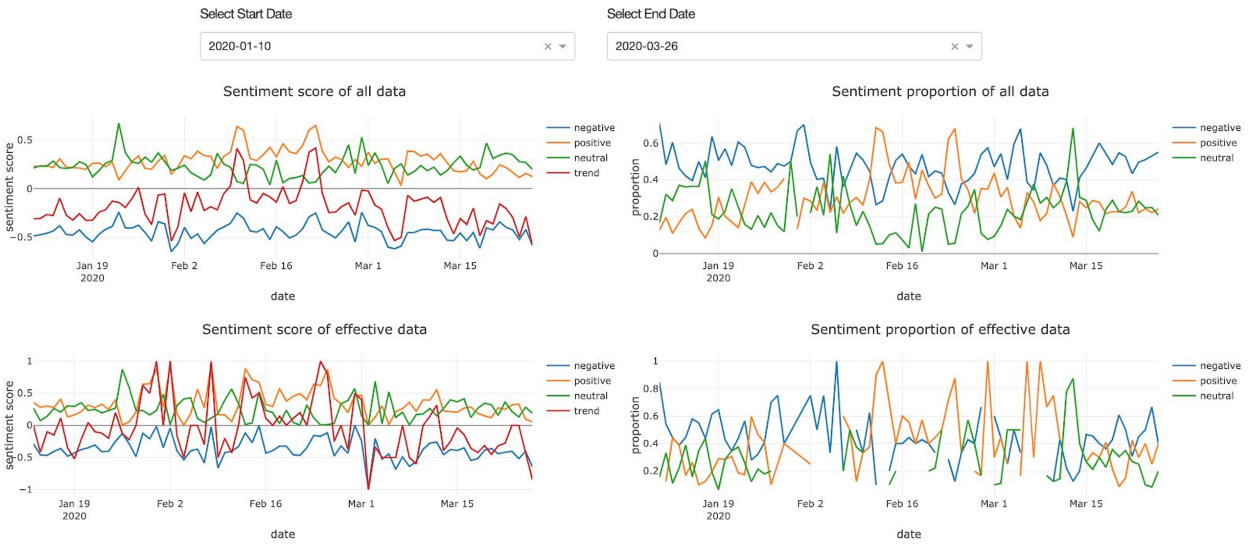


Therefore, we decided to choose "coronavirus" as our sentiment keywords. We could not only have a comprehensive understanding of the development of this epidemic, but also see how people’s emotion has changed over time.

ii. Sentiment Analysis and Trend

Sentiment score and proportion are used to measure the sentiment trend of Weibo posts. Sentiment score is the average of the label value,with negative label as -1, neutral as 0, positive as 1. Proportion presents the faction of each labeled class. Considering the difference between official account and personal account and wondering what kind of post people prefer, we took posts with more that 50 like/forward/comment as effective data.

From our result, people’s attitude on bad things is not always negative, which can be shown from the red line in two pictures on the left. During the outbreak of COVID-19, not only negative impact of the events is transferred on social media, encouragement is also passed on with positive energy. At the same time, people have a preference toward positive posts, which can be shown in the two figures below.



Select Start Date: 2020-01-10

Select End Date: 2020-03-26

iii. WordCloud

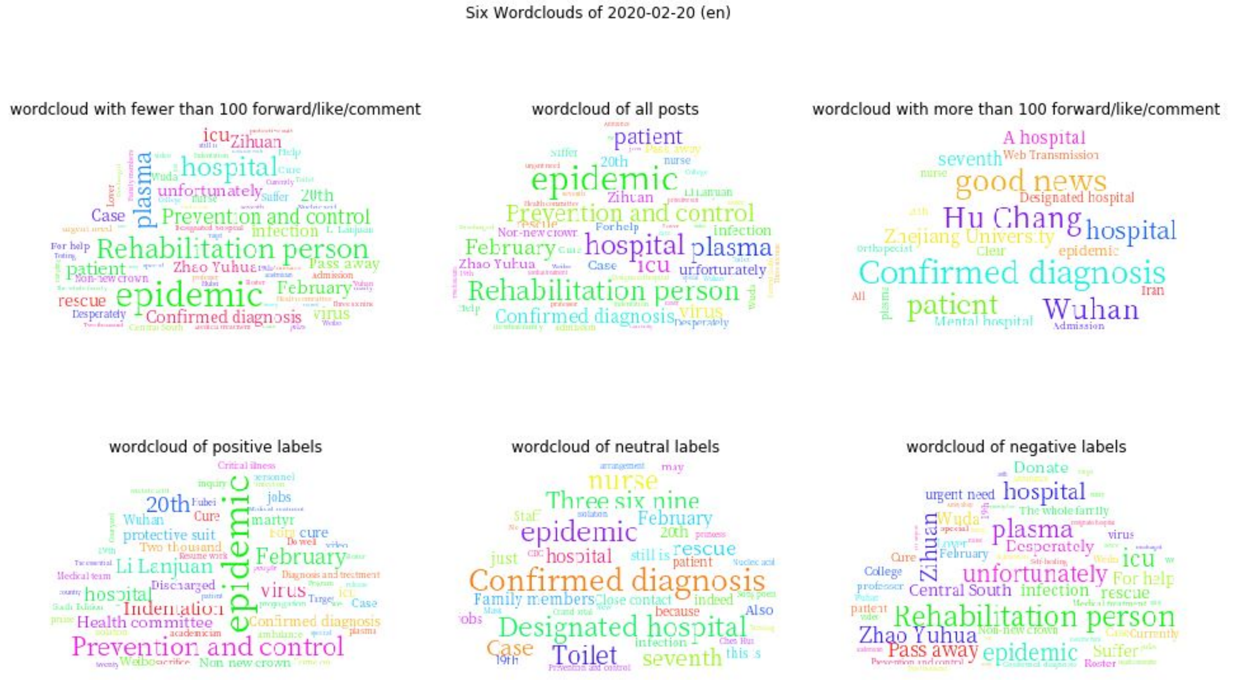
WordCloud is a data visualization technique to vividly represent text data and is widely used for analyzing data from social network websites. The size of each word indicates its frequency or importance. It is a great tool to show our results for most frequent words and topics by highlighting and coloring.

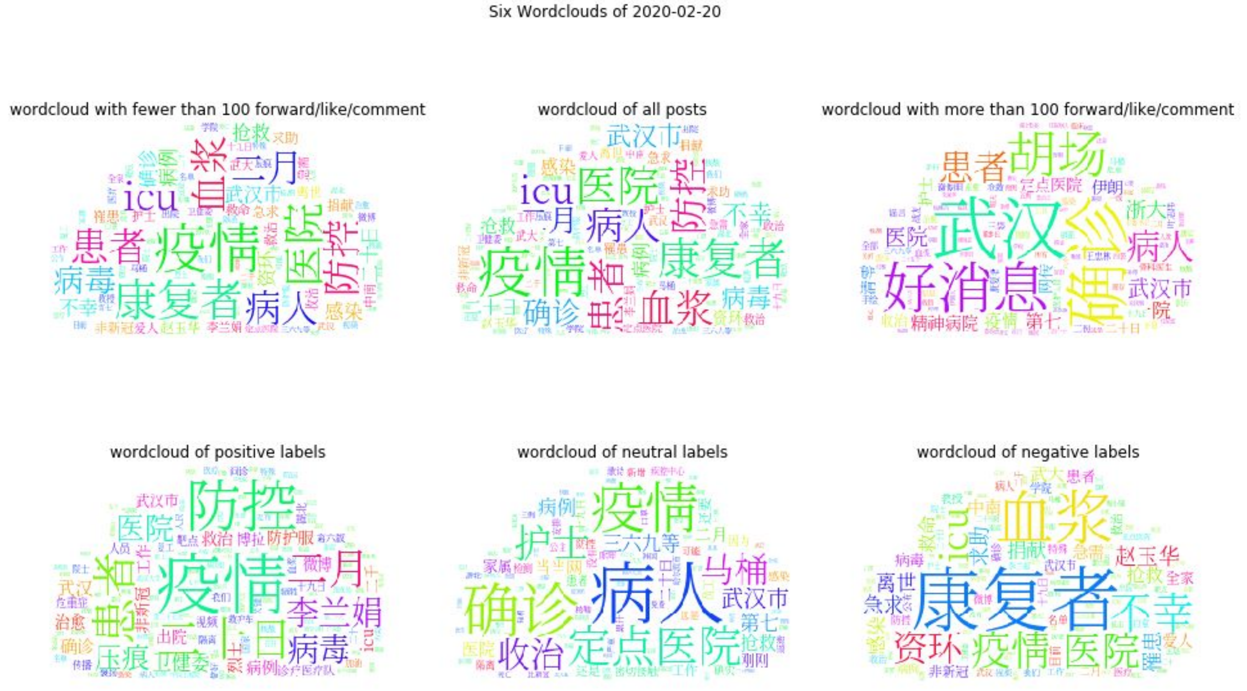
We generated our WordCloud in Python with the following modules: matplotlib, pandas and WordCloud. From the segmented word list from pre-processing, we aggregated all post contents on each day, calculated the word frequency list and generated the WordCloud object.

For each day, we generated six different WordClouds, to show the difference of:

①all posts with fewer than 100 comments/likes/reposts, ②all posts, ③all posts with more than 100 comments/likes/reposts, ④all posts with negative labels, ⑤all posts with neutral labels, and ⑥all posts with positive labels.
Then we translated the six WordClouds to English.

Take the day of 2020-02-20 as an example:





5. Difficulty and Challenge

a. Data collection

Most webs only present real-time hot search but not historical data. It is time-consuming to find a proper API to get hot search data.

The way we collect our data is through target searching, which might trigger the robot detection of the website and ask for identity verifications. We have to overcome this issue using random break times and keep the crawler running for weeks.

Tokenizing Chinese sentences is also quite challenging, we have to fine-tune and evaluate six different segment models to find the appropriate one for this project.

Locating original post among tons of forwarded posts.

b. Pre-processing

Dealing with unstructured and non-grammatical texts.

Aligning different numerical and date formats over multiple languages.

c. Data Labeling

People express opinions in complex ways, including rhetorical devices such as sarcasm, irony, implication, etc. Not to mention the extensive usage of acronyms. All these indicated that this job can only be done manually. In order to improve the performance of our model, we used 30 days labeled over 45,000 Weibo posts.

d. Model Selection

i. Segmentation Model

We evaluated 6 different text segmentation models (jieba > hanlp > pkuseg > thulac > snownlp > foolnltk) and picked one that performs best with Chinese character sentences.

ii. Word Representation Model

We compared Word2Vec and FastText models on Chinese character sentences and found FastText gives a better result that is more compatible with our sentiment tagging procedures.

e. Visualization

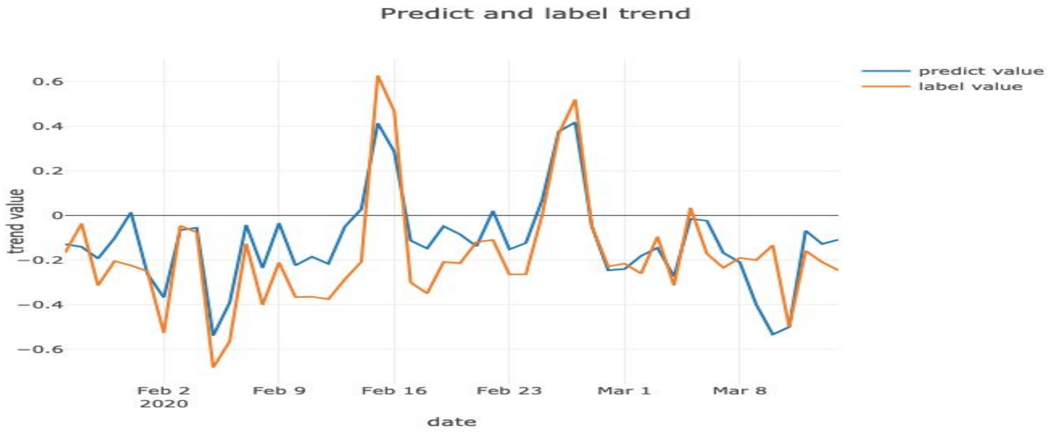
Mandarin-English Conversion

The hard limit for google translate API requests is 30K Unicode characters (code points). To overcome this issue, we split the vocabulary of over 60,000 unique words into 15 chunks and distributed 15 tasks using virtual machines. Then merged all the results locally to be used for the translation mapping.

6. Model Evaluation

a. Sentiment trend value

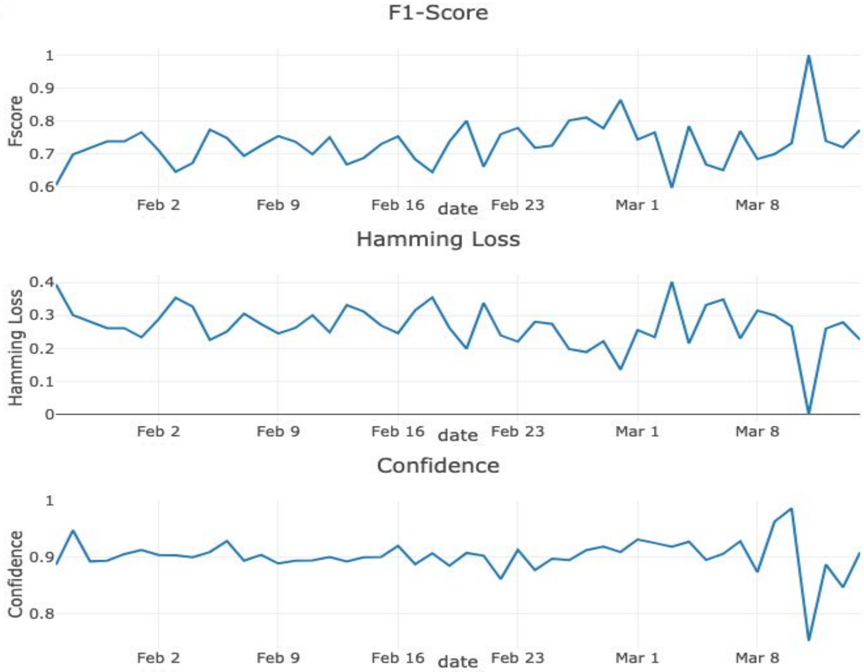
Sentiment trend value is designed to present the total trend of data. It measures the average classified result for negative as -1, neutral as 0, positive as 1. The result of prediction is closed to the label result, which is as following:



b. F1-score, hamming loss and confidence

These three scores were used to evaluate the classification accuracy. F1-score is a weighted harmonic mean of precision and recall. Higher values of the F1-score indicate that the classification method is more effective. Hamming loss is the fraction of labels that are incorrectly predicted. Confidence is the predicted probability of test data that predicts correctly. The definition of precision(P), recall(R), F1-score and the result as following:

$$P = \frac{Tp}{Tp+Fp}\ ,\ R = \frac{Tp}{Tp+Fn}\ \ ,\ F1 = \frac{P*R*2}{P+R}$$

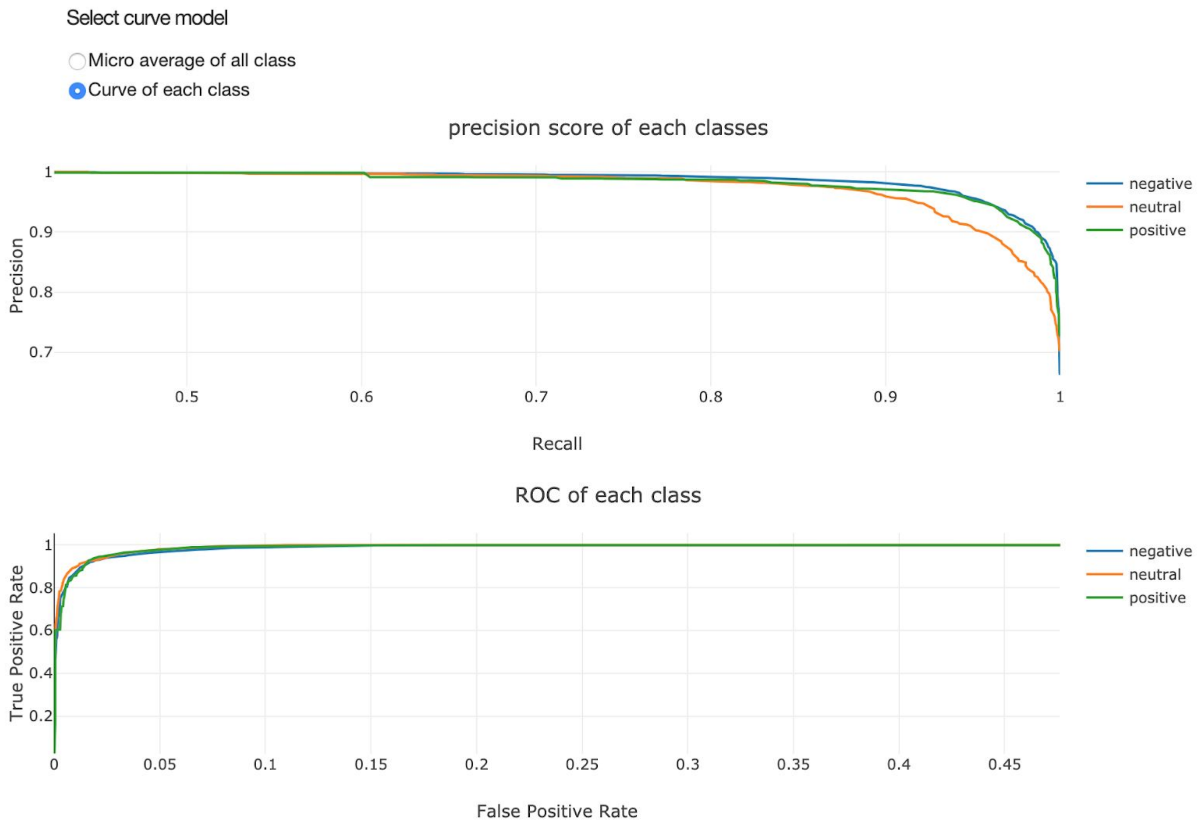


c. PR Curve and ROC Curve

PR/ROC Curve is used for further understanding of the model. PR Curve shows the relationship between precision and recall. ROC Curve summarizes the trade-off between the true positive rate(TPR) and false positive rate(FPR) for the model using different probability thresholds. TPR and FPR are defined as following:

$$TPR = \frac{Tp}{Tp+Fn}\ ,\ FPR = \frac{Fp}{Fp+Tn}$$

The indicators are for binary classification, but the model is for multi-classes problems. Therefore, we use micro-average to present the PR/PRC Curve of total data. Micro-average aggregates the contributions of all classes to compute the average metric. We also compute the PR/PRC Curve of each class respectively.



Select curve model

○ Micro average of all class

● Curve of each class

The above result is the model that trained and tested with 60,000 weibo posts. It performs better than we expected, the possible reason is the posts for training and testing all belong to the same topic, so that their contents could be similar.

Before training with full data, we also tried to train the model with only 1000 labelled data, but that performs poorly.

7. Data Product

Our Visualization result can be accessed by the following URLs:

Result of Hot Search Topics
http://ec2-18-218-241-59.us-east-2.compute.amazonaws.com:8051/
Result of Weibo Sentiment Analysis
http://ec2-18-218-241-59.us-east-2.compute.amazonaws.com:8050/

8. Lessons Learnt

a. Team Collaboration

Under this special quarantine time due to COVID-19, our team transitioned from off-line meeting to on-line meeting, and under different time zones. It can be challenging sometimes, but also served as a great opportunity to practice code-sharing, self-discipline and remote communication, which can be essential skills later in either academia or industry.

b. Data Science Techniques

i. Data Collection

Through the challenges we faced, we learned that in order to have a decent result, data collection serves as the most important first step and should be considered thoroughly. In our case, data should be collected randomly, and then classified or labels, instead of using target searching.

ii. Multi-language Results

We also learned how to efficiently prepare for multi-language versions of all results since our raw data is all in Chinese.

9. Summary & Future Work

Surprisingly, from the visualization we can tell, people’s attitude on bad things aren't always negative or dominated by negativity. That’s also why data science is important, things might not always be the way we thought.

Compared with traditional text sentiment analysis, our project on Weibo has its similarities and particularity. The basic structure of data science flow is similar, but the particularity is mainly reflected in dealing with Chinese language differences. As a new research direction of data science, Chinese Weibo sentiment analysis still has many areas worthy of in-depth exploration. Future sentiment analysis on Weibo should have the following directions:

a. Dealing with Spam Contents

Pre-process procedures need to be refined. The existence of spam information on Weibo will undoubtedly interfere with sentiment analysis, and currently there are very limited established filtering algorithms.

b. Labelling Data

We can find a better way to label data, like using crowdsourcing. Finish labelling over 50,000 content in a short time is overwhelming.

c. Online Language Usage

There is a lack of filtering and sentiment mining of “online language”. There are few related dictionaries or corpora available for use now.

d. Label Expansion

Weibo posts, as well as many texts have various emotions, and should not be limited to positive neutral and negative aspects. It can be extended to explore different emotions and their levels.

e. Live Updates

The data analysis product can be made from using historical data to live updates, which needs scalable tools and cloud computing.