Computational Social Science Analytics Project

An Insight from ***Hong Kong*** Protests

Brought to you by Team **The Supreme Press**
SMU IS434 Social Analytics and Applications
Date of Submission: 12 Nov 2019

BU Wende, JIANG Hanyu, QI Haodi, ZHANG Chengzi

Contents

Abstract

The 2019 Hong Kong Protests sparked off by the Extradition Law Amendment Bill introduced by the Special Administrative Region (SAR) government in March this year. This issue has gone viral online and led to an increasing number of protestors and a series of planned demonstrations across the island. Due to the newsworthy and contentious nature of the issue, a detailed analysis integrating social science research methodology and data analytics methods will benefit many people in the society, ranging from news media agency to sociologist researching on political issues.

We have scraped data from social media platforms like Twitter and Reddit using real time listener on AWS cloud to maximise the amount of data collected. After data manipulation and data cleansing, we have generated data sets to perform network analysis and text & sentiment analysis.

For network analysis, we have visualised the interactions between twitter users using Gephi and calculated the network measures at both macro and micro level. At the micro level, we are able to identify different types of influencers in the network based on the centrality measures. At the macro level, we are able to observe the changes of network structure and characteristics over time.

Our text analysis consists of sentiment analysis and topic extractions. For sentiment analysis, we have used vader polarity score to observe overall sentiment of twitter users in different regions over time. To further analyse users' sentiment towards each topic, we have first extracted a list of popular topics, and under each topic, we have conducted sentiment analysis as well as common words related to the topic which can be representative of a user's opinion.

The consolidated findings were visualised them into a Tableau dashboard. The dashboard displays location-specific sentiment scores and can display the information filtered by time. Together with the network graphs and word cloud generated, these visualisations allow individuals who are interested in the issue to gain a better understanding.

1. Our Clients

Our potential clients are news media and journalists who are interested in Hong Kong Protests, such as ChannelNewsAsia, Mediacorp and The Straits Times. Another group of potential clients are researchers who want to analyse the Hong Kong Protests using computational social sciences.

Traditionally, newsagents only interview key stakeholders or representative figures on the topic at hand, as it is impossible to interview everyone from the general public. However, as more and more people are using social media to express their views, it becomes more and more valuable and newsworthy to capture, analyse and report the public sentiments online as a whole. Newsagents who want to stay relevant and competitive in today's world will have to leverage computational social science analytics to understand the public sentiment and get insights from it.

For social science researchers, it becomes essential to perform an in-depth analysis of public sentiments, identification of key stakeholders and community network involved in the research topic. It becomes more and more difficult to manually trace everyone in the network as the networks grow larger. In addition, statistical and scientific evidence are required in their research to testify to a hypothesis. Therefore, computational social science analytics is the solution to their problem by providing accurate and scientific analysis on the topic at hand.

2. Twitter Data Analysis

2.1 Data Collection and Data Description

The collection of tweet data leverages on Twitter API; due to the limit of 180 requests per 15-minute window of Twitter Search API, we used its Streaming API to crawl real time tweets. We spent on average two to three hours a day to crawl tweets; the real-time listener is also deployed on AWS Cloud to continuously crawl data. The data collection period is from 24 September 2019 to 30 October 2019.

To ensure the tweet data crawled is relevant to the topic of Hong Kong Protest, we only retrieve the tweets with commonly used, Hong Kong protest related hashtags, such as “#standwithhongkong” and “antiELAB”.

Due to the limited computational power of our hardware, we are unable to process all the tweet data crawled online. As such, we randomly sampled at most 5000 tweets from tweets created between 7pm to 11pm Hong Kong time (UTC+08) and 7pm – 11pm (UTC-05) for each day. The reason why we chose this time period of 7pm to 11pm is that we believed people would tweet more after work and school before sleep. In total, we have 150,889 tweets for our analysis.

The tweet data crawled using Twitter API is in the format of nested JSON objects; each tweet JSON object contains three main categories of information: information of this tweet, information of the user who tweeted this tweet and the original tweet that this tweet retweeted or quoted. For our analysis, we extracted the following information from the tweet JSON objects:

1. Tweet id, as the unique identifier of each tweet
2. Datetime when the tweet is created
3. Text of the tweet
4. Country code in which the tweet is created
5. Language in which the tweet is written
6. Hashtags in the tweet
7. User of the tweet and his/her screen name
8. The user of the original tweet and his/her screen name, if there is any

The sampling of tweets is based on the created datetime of tweet and we ensure there is no duplicated tweets in our data by validating against the unique tweet id.

We carried out two types of analysis with tweet data: text analysis with items 1 to 6 and network analysis with item 7 and 8.

2.2 Tweet Data Processing and Text Analysis

Our text analysis consists of two main parts: common word extraction and sentiment analysis. Common word extraction allows our clients to identify the hot topics about Hong Kong protests Twitter users are discussing and the keywords related to each topic; sentiment analysis enables our clients to have not only a general understanding of Twitter users attitudes towards Hong Kong protest, but also their sentiment towards each topic.

Tweet Structure

A tweet can contain any of the following elements:

1. Words in the tweet content
2. Non-alphabetic characters in the tweet content (e.g. numbers and punctuations)
3. URLs
4. Emojis (e.g. 😊) and Emoticons (e.g. :D)
5. Hashtags (e.g. #hashtag)
6. Mentions (e.g. @user)

The standard practice of text data pre-processing involves removal of non-alphabetic characters including emojis and emoticons and common procedure of cleaning tweet content include removal of hashtags, URLs and mentions. However, after looking through some tweet, we identified a few issues with such a practice and procedure regarding hashtags, and emoji, emoticons and punctuation.

Hashtags

As many hashtags represent an entity or a topic, many users directly use them as the subject or the object in their tweets. One example is *"#HongKongPolice has been calling #HongKongProtester "cockroach", and now "yellow object". #racism #YellowObject*", where hashtags #HongKongPolice and #HongKongProtester are used as the subject and the object of the sentence respectively. Therefore, removing hashtags will destroy the sentence structure and potentially lose certain insights from the tweet.

On the other hand, as we crawled data with hashtags as filters, all tweets will contain at least one hashtag in our filter list; based on our observation, one tweet often contains about 2 hashtags on average (1.73 hashtags on average for our sample tweets). Therefore, keeping hashtags or simply removing the "#" symbol will cause them to be the top commonly used words and phrases in our common word extraction, which does not provide much value to our clients.

Therefore, we need to deal with hashtags separately for different analysis and test which approach offers greater values.

Emoji, Emoticons, Punctuation

Similar to hashtags, emoji and emoticons may affect common word extractions and thus should be removed. However, along with punctuation, they reflect users' sentiment more accurately. For example, when comparing the negativity in the sentiment of tweets, "I hate you", "I hate you!" and "I hate you! 😠", clearly the negativity increases with "!" and "😠". As such, we should not ignore and remove punctuation, emojis and emoticons when conducting sentiment analysis of the tweets which reflects the users' sentiment.

Language

As Twitter is a global social media platform and Hong Kong protest has been a popular topic across the globe, our sample contains tweets written in different languages. For our sample tweets, about 72% of tweets (108617 tweets) are written in English and our analysis will be focused on them.

Data Processing

The general steps for tweet content processing are the following:

1. Pre-processing using the clean function in tweet-preprocessor
2. Deal with hashtags
3. Deal with punctuation, emojis and emoticons
4. Lemmatize words

We imported an open source library, tweet-preprocessor, for pre-processing of the tweets. Its clean function allows us to specify aforementioned items 3 – 5 to be removed from a tweet.

Knowing the potential issues with removing hashtags, through our observation, we realize that many hashtags are in the form of camel casing, such as "#HongKongPolice". Therefore, we split up it up into individual words after removing the "#" symbol. For hashtags that are not in such a form, we leave it as a term by itself.

It is easy to deal with punctuation as we can choose not to remove them from the tweet. For emojis and emoticons, we imported an open source library, emot, to convert emojis and emoticons to English words. In this way, the use of emojis and emoticons to express one's sentiment can be captured in the model used in our sentiment analysis.

Lemmatization of words is carried out to convert words into their root form in English is a standard procedure for text pre-processing. We did not choose to stem the words because the stemmed words may not be a proper English words and thus may cause confusion to our clients. We did not remove the stop words at this stage and utilized the part-of-speech technique with the NLTK library so that the words are tagged and then lemmatized correctly.

The detailed approach of the tweet cleaning for different analysis will be discussed in each of the following sections.

Word Cloud

The purpose of the word cloud analysis is to identify and visualize common words and phrases that appear in the tweets. Visualizing the common words and phrases in the form of word cloud is more visually appealing and attention capturing, providing greater values to our clients.

As for the data pre-processing, we removed punctuations, emojis and emoticons, and attempted both approaches with regards to hashtags, removing them and splitting them up. To further enhance the visualization and after discovering the presence of non-English words in the tweets, we used the Rui Zi Yun font and Hong Kong map as the mask for our word clouds.

Splitting up Hashtags

From Fig. 2.2.1 where hashtags are split up, we can see that the words standing out in the word cloud mostly contains “hong” or “kong” or both. This is not surprising as we found that hashtags containing “hong” or “kong” or both takes up about 41% of all hashtags used in our sample tweets. Since the overall topic is Hong Kong protest, “hong” and “kong” should not be considered as common topics and thus should be excluded.

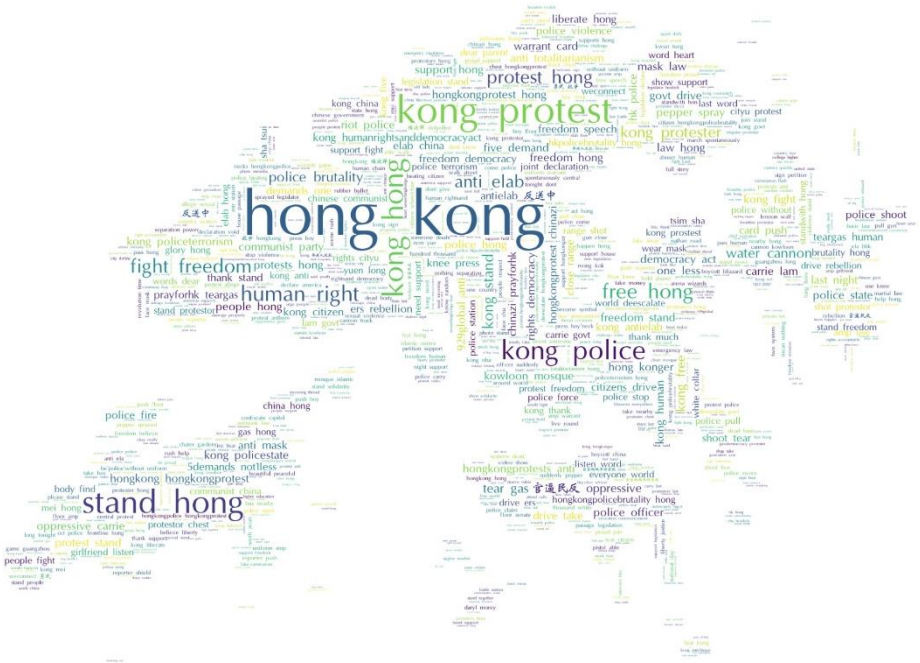

Fig. 2.2.1 Word cloud with hashtags spit up

Fig. 2.2.2 is the word cloud after we remove phrases containing “hong” or “kong. It shows much better results compared to Fig. 2.2.1 as phrases representing Twitter users’ concerns and interests stand out in the word cloud, such as “human right” and “police brutality”. However, we also found out that some phrases, such as “anti elab” and “929global anti”, are from hashtags “AntiELAB” and “929GlobalAntiTotalitarianism”.

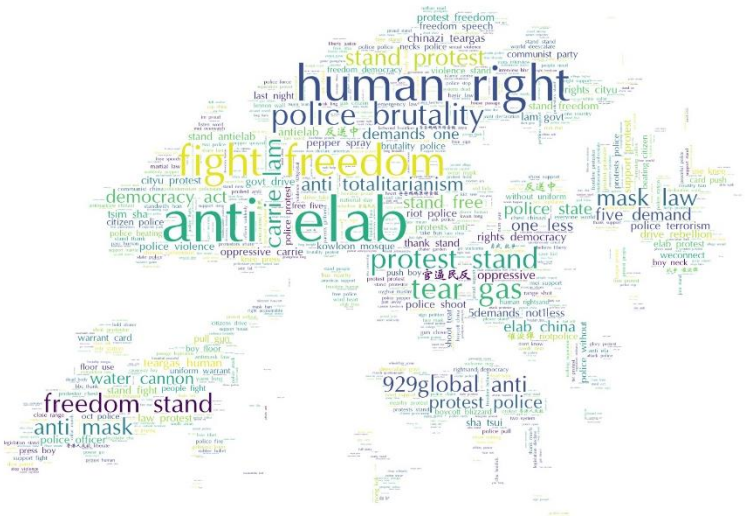

Fig. 2.2.2 Word cloud with hashtags spit up and “hong” and “kong” removed

Before moving on to remove the hashtag, we attempted to create word cloud by using only noun phrases in the tweets, with the TextBlob library (Fig. 2.2.3). The word cloud shows topics that are not seen in Fig. 2.2.2, such as “tear gas” and “mask law”, which are related to critical incidents in the whole Hong Kong protest.

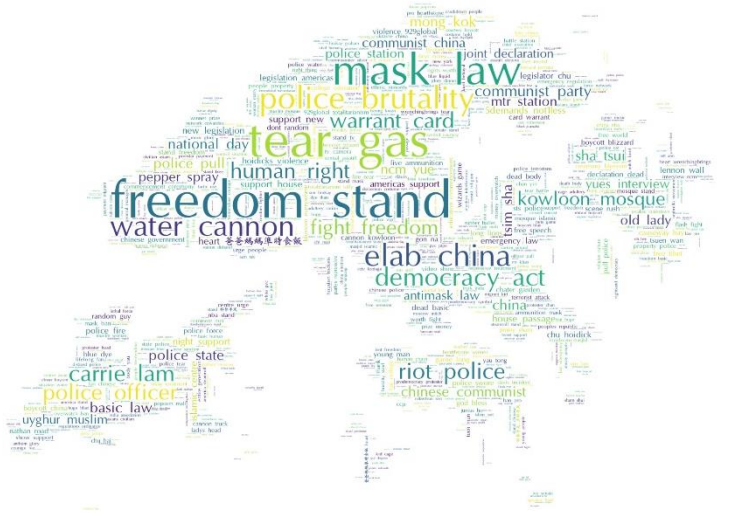

Fig. 2.2.3 Word cloud using noun phrases with hashtags spit up

Remove Hashtags

Lastly, we removed the hashtags from the tweet and generated two-word clouds, Fig. 2.2.4 and Fig. 2.2.5, for tweets without hashtags and noun phrases in tweets without hashtags respectively. Similar to Fig. 2.2.2 and Fig. 2.2.3, there are words appearing in both word clouds and unique words appearing only in one-word cloud.

One interesting observation is that the term “police brutality” is not shown on Fig. 2.2.4 and Fig. 2.2.5. We discovered that the term actually appears in many of the hashtags and thus is removed in our pre-processing for Fig. 2.2.4 and Fig. 2.2.5. We recognized that such terms are valuable as it shows users’ sentiment towards the police force and therefore, we should consider not removing hashtags when conducting sentiment analysis.

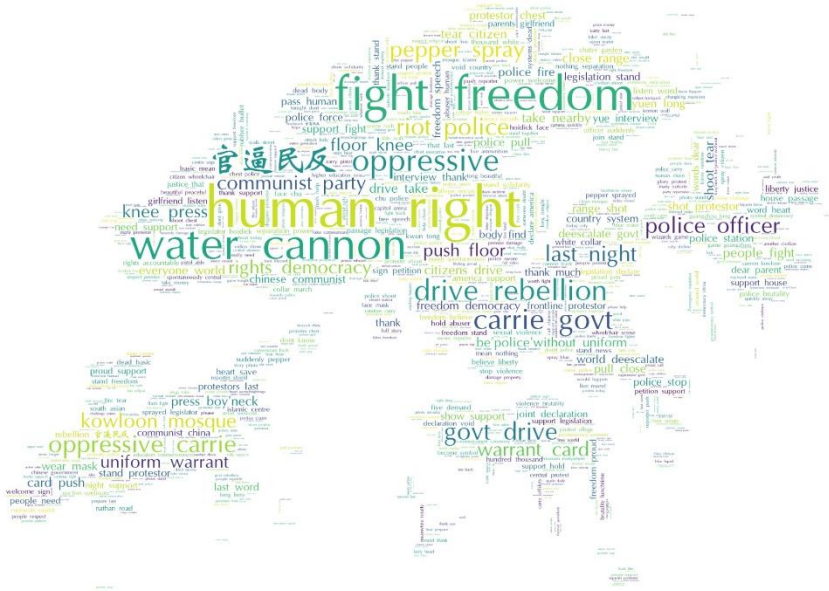

Fig. 2.2.4 Word cloud with hashtags removed

Fig. 2.2.5 Word cloud using noun phrases with hashtags removed

Word Cloud with Unigram

Furthermore, as we used bigrams when generating for Fig. 2.2.1 to Fig. 2.2.5, we also attempted unigram to identify popular single words (Fig. Fig. 2.2.6). The word cloud is not very informative as a single word such as “people” do not give much information about users’ opinions.

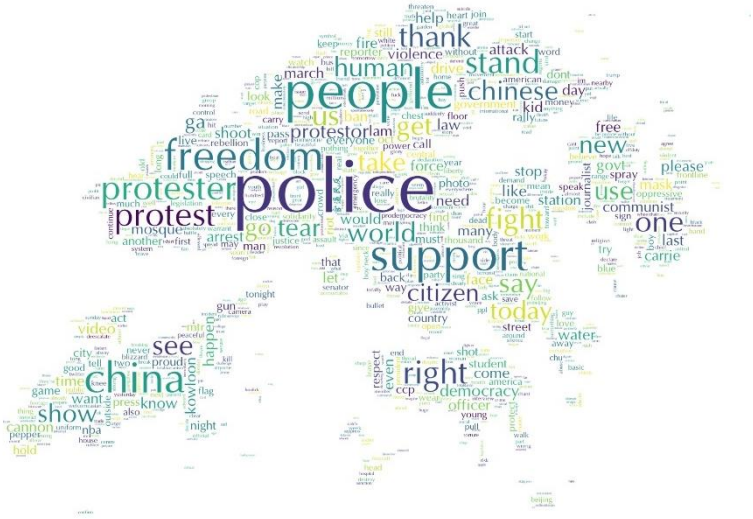

Fig. 2.2.6 Word cloud using unigram with hashtags removed

Findings and Insights

In conclusion, we discovered that Twitter users in our sample have many opinions about Hong Kong Police Force, especially with regards to how they deal with the protesters, such as using water cannons, pepper sprays and tear gas. Moreover, regarding the overall issue, Twitter users in our sample believed the protest is about human rights and the Hong Kong government led by Carrie Lam is oppressive and thus drives the rebellion from the public.

Popular Topic Extraction

One drawback about the word cloud is that it includes any bigram phrases that appear frequently in the tweets and many phrases are common phrases, such as “last night”, and do not give information about the topics of discussion. Moreover, certain phrases do not capture the main entity or event, such as “shoot tear”, which should be “tear gas”. As such, we moved on to identify the popular topics from the tweets to deal with those issues.

As the outcome of this part of the analysis is similar to that of word cloud, we continued using the cleaned tweets without hashtags.

Assumptions

Our topic extraction is built upon the assumption that, if a tweet contains a particular phrase, the phrase is a topic of that tweet. The rationale behind the assumption is that as tweets are rather short compared to an article, the phrases in the tweets would be representative of tweet.

Overall Popular Topics

We carried out popular topic extraction for both unigrams and bigrams. Similar to word cloud generation, we removed words or phrases containing “hong” and “kong” and other location words. To deal with the issue of “weird phrases”, we came up with a list of reserved pairs, such as “carrie lam” and “communist party”. When a word in a phrase is in a reserved pair, the phrase must be the pair, or it will be excluded.

The popular topics extracted with unigram match with the word cloud generated with unigram (Fig.2.2.6), such as “police” being the top followed by “people” and “support”.

The bigram popular topics are more informative than the word cloud, because of removal of the weird pairs. Most of bigram popular topics are captured by the word cloud, such as “human rights” and “water cannon”. Some words that are not highlighted in the word cloud stand out, such as “carrie lam” and “tear gas”.

We can repeat the same process or use word cloud to identify words related to a particular topic to further understand Twitter users’ opinion regarding the topic. For

instance, “human rights” as the most popular phrase is often associated with “democracy act” and “new legislation”, indicating that the new legislation that Hong Kong protesters are striving for is about democracy with the purpose to protect their human rights.

These topics enable our clients to direct their focus of investigations so that they can be more cost efficient and effective on their reporting and research.

Popular Topics over Time

To identify the popular topics of discussion for each day, we carry out the same procedure for unigram and bigram for tweets of each day. While the unigram results are very similar across different days and thus not provide much information, we can discover that certain topics are popular across days and some only becomes popular on certain dates. For example, “national day” is the most popular topic on 30 September as it is the China National Day the next day.

When we look at the number of tweets containing a particular popular phrase and the number of days where the phrase is among the top 10 popular phrases, we can see that some words are only appear as popular topics for a few days, there are many tweets about it, such as “water cannon”. The popularity of a topic over time can be seen from Fig.2.2.7. There is a sudden increase in the number of tweets about water cannon on 20th October, as Hong Kong police started to use that around that date.

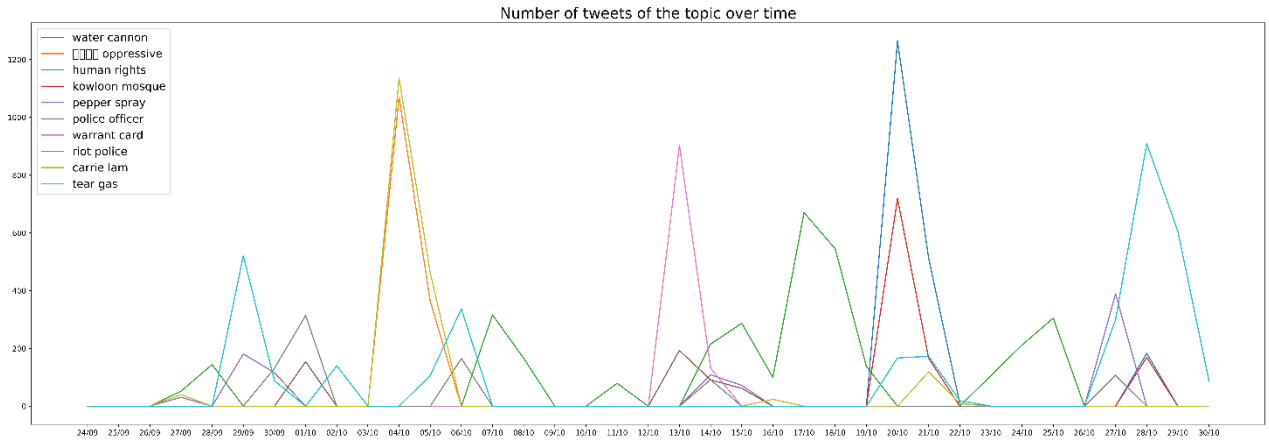

Fig. 2.2.7 Number of tweets of topics over time

Such trend can alert our clients that there may be a popular topic that they should take a look at since there is a sudden surge in the number of tweets about it, indicating certain incidents must have happened.

Topic Association

Besides identifying and extracting the popular topics, we also explored the associations between topics by finding the number of tweets that two particular topics appearing together.

We visualize them using the Networkx library to see the entire topic network with the edge weight indicating the scaled number of tweets containing both topics, followed by leveraging on the Community library to find out the communities of topics. Fig. 2.2.8 shows the topic network from the top 20 popular bigram topics. It can be seen that “官逼民反 oppressive” and “carrie lam” have a stronger association since they are often mentioned together, indicated by a much thicker edge.

In this way, our clients will be able to not only be able to identify popular topics alone but also see the overall picture of what Twitter users are discussing and how their topics of discussion link from one to another.

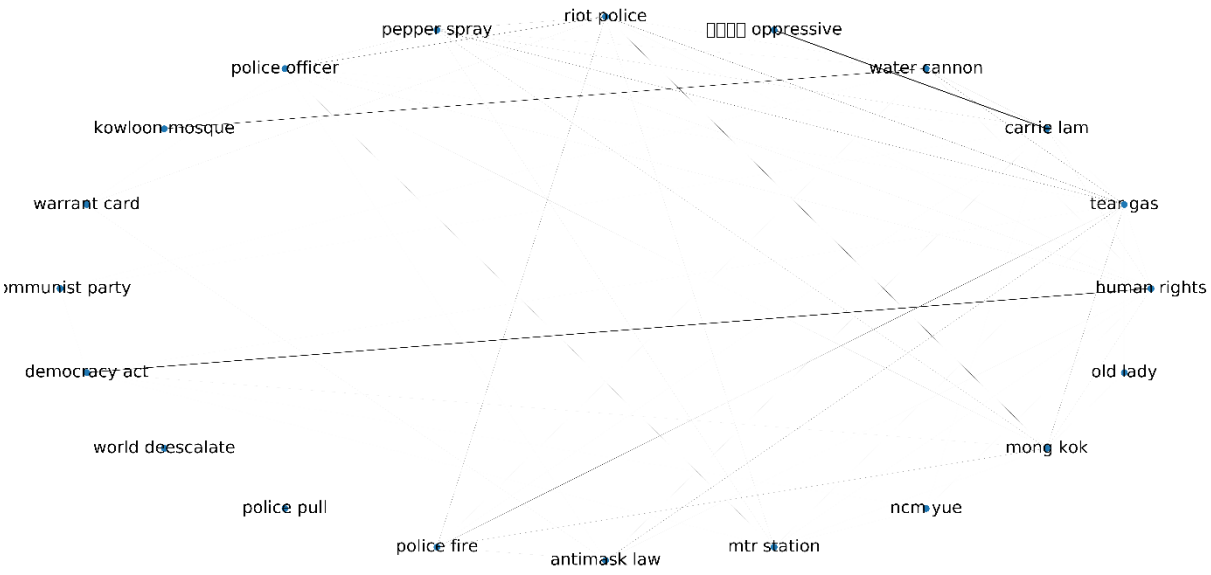

Fig. 2.2.8 Topic Network of the top 20 bigram topics

Sentiment Analysis

VadBesides the popular topics of discussion, we understand that our clients also need to understand the users' feelings and attitude towards the topics and Hong Kong protest in general and sentiment analysis is conducted to measure that with a quantitative indicator.

Sentiment Analysis Tool

Many libraries contain pre-trained sentiment analysis models; in our analysis, we used the Valence Aware Dictionary and sEntiment Reasoner (VADER) sentiment analysis, which is a lexicon and rule-based tool that specifically attuned to sentiments expressed in social media. We also discovered that VADER takes into account of the punctuations and converted emojis and emoticons and gives different sentiment scores.

VADER sentiment analysis provides four scores with its polarity_scores function: positive, negative, neutral and compound. Based on its documentation, a tweet with a compound score more than 0.05 is considered one with positive sentiment, below -0.05 is one with negative sentiment and between -0.05 and 0.05 is considered a neutral one. It is said that the compound score is "the most useful metric if you want a single unidimensional measure of sentiment for a given sentence".

Therefore, as sentiment analysis does not concern common topics and some hashtags reflect users' sentiments, our text pre-processing will involve splitting up the hashtags, not removing the punctuation and converting the emojis and emoticons.

Aggregation of Sentiment

With VADER compound score, we can categorize a tweet to be with positive, neutral or negative sentiment. However, when having a list of tweets about a topic or in a day with their respective sentiment scores, there are multiple ways to aggregate the sentiment of tweets to conclude the overall sentiment.

In our analysis, we had three methods for the aggregation: mean compound score, median compound score and the sentiment with the highest proportion of tweets.

The mean compound score method is the simplest method given a list of tweets with their compound scores. However, the problem with mean is always the presence of outliers. In our case, it is likely that someone express views with extreme sentiments on Twitter, due to the anonymity of social media.

As such, possibly using median score is a better method in gauging the overall sentiment. However, it is likely that the median score will fall in the neutral sentiment range, thus unable to take into account of the tweets with positive and negative sentiment.

Last but not least, another approach is to find the sentiment with the proportion of proportion of tweets. This approach totally disregards the compound scores and the extent of negativity or positivity, and only considers the category of sentiments the tweets fall into. This can give us the proportion of the user community of different sentiments and may provide extra insights as mean and median score is only a single-value measure.

Overall Sentiment

From Table. 2.2.9, we discovered that based on both mean compound score and median compound score, the overall sentiment of Twitter user sentiment in our sample is neutral. However, from the proportion of tweets with negative and positive sentiment, it is observed that there are about an equal number of tweets for positive and negative sentiments. Moreover, from Fig. 2.2.10, the distribution of compound score for tweets with positive and negative sentiments seems symmetric. It can also be observed that there are indeed a number of tweets with extreme positive and negative sentiment. A similar proportion and distribution leads to the almost-zero mean score and zero median score.

Thus, we can conclude that the sentiment of Twitter users in our sample towards Hong Kong protest is split rather equally towards negativity and positivity.

SENTIMENT	PROPORTION OF TWEETS WITH THE SENTIMENT	MEAN COMPOUND SCORE OF TWEETS	MEDIAN COMPOUND SCORE OF TWEETS
POS	42.47		
NEU	15.43	0.000006	0
NEG	42.11		

Table. 2.2.9 Overall sentiment statistics of the tweets

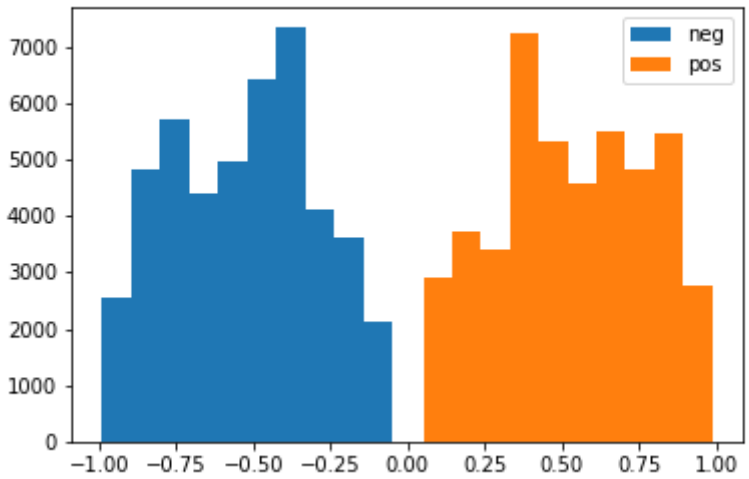

Fig. 2.2.10 Distribution of the compound scores of tweets

Change in Sentiment Over Time

Similar to the overall sentiment analysis, we calculated the mean and median compound score of tweets and the proportion of tweets with different sentiments for each day. Fig. 2.2.11 and Fig. 2.2.12 shows the trend of the change over time.

The mean and median sentiment scores follow a similar trend, but they are fluctuating over time. The proportion of tweets exhibit similar fluctuations. Moreover, it can be further confirmed that the sentiment drawn from the highest proportion of tweets is not necessarily the same as that drawn from the mean or median score. These two figures enable our clients to identify how events and incidents on certain dates affect Twitter user sentiment.

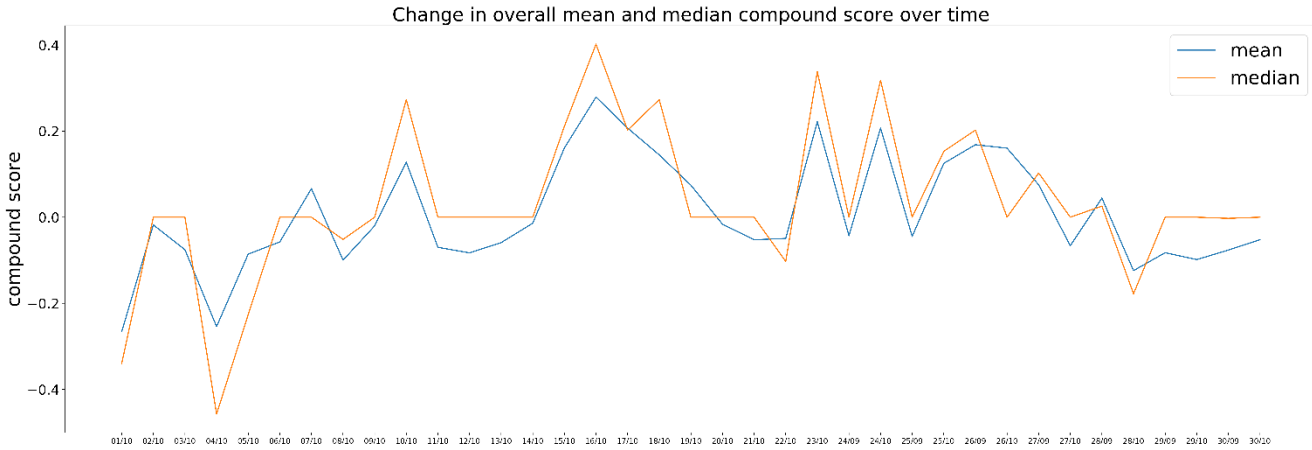

Fig. 2.2.11 Change in mean and median compound score of all tweets over time

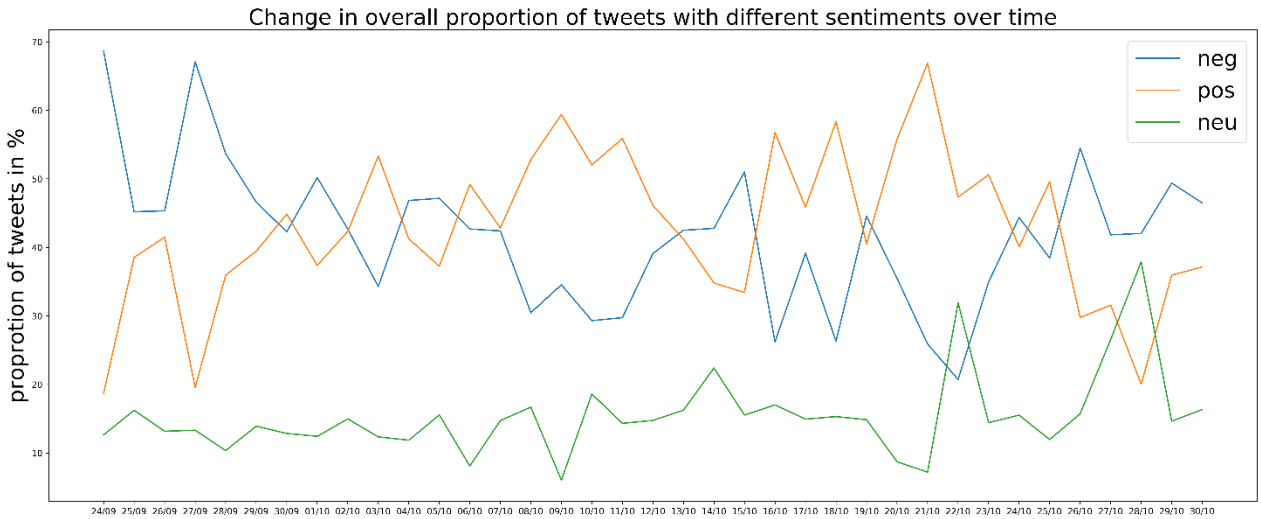

Fig. 2.2.12 Change in proportion of with different sentiments over time

When it comes to the overall sentiment for each day, we can see from Table. 2.2.13 that the number of days with overall positive and negative sentiment by mean compound score is similar to that by proportion. However, the majority of days that have an overall negative sentiment by mean scores or highest proportion fall under the neutral sentiment with median scores (Fig. 2.2.14). It is discovered that there are 10 days with an overall negative sentiment based on mean scores and the proportion have a neutral sentiment based on median scores; this result is very similar to the overall sentiment result, further confirming that users have a clear stand of whether they are supportive or not.

SENTIMENT	DAY COUNT BY MEDIAN SCORE	DAY COUNT BY MEAN SCORE	DAY COUNT BY HIGHEST PROPORTION
POS	6	16	20
NEU	21	8	0
NEG	10	13	17

Table. 2.2.13 Number of days with overall sentiment by different measures

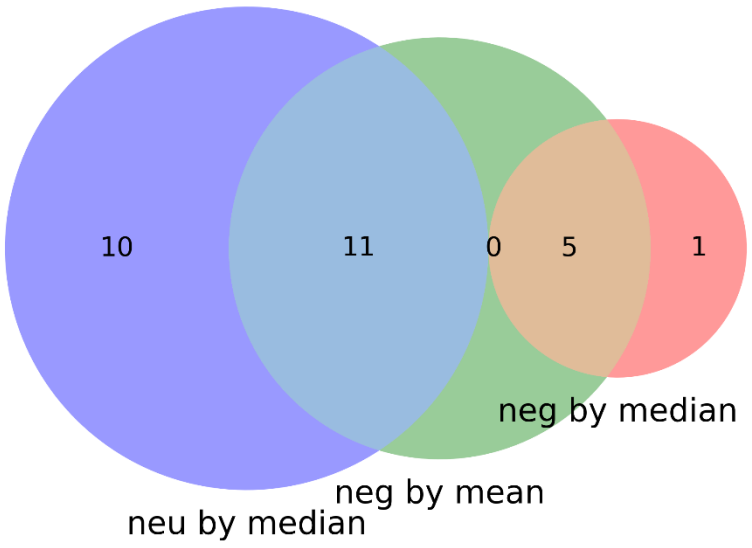

Fig. 2.2.14 Venn Diagram of the days with negative sentiment by mean score and those with neutral and negative sentiment by median score

Sentiment of Topics over Time

With the popular topics generated from Popular Topics segment, we can further analyse the Twitter user sentiment towards each of the topic. Using the same three measures, the number of topics with positive sentiments from the Twitter users in our samples is the most, and similar to that with negative sentiments. There are little discrepancies between the different measures, indicating that the Twitter users in our sample have a similar sentiment towards a topic.

After looking through the list of positive and negative topics, we did not discover patterns from the positive topics, but we found out that there are a number of negative topics related to Hong Kong police force and the Hong Kong government. This shows that while users may be positive about a range of topics, they are generally negative about the Hong Kong police and government.

Moreover, when the sentiment towards a topic across different dates, we can see its general sentiment as well when the sentiment changes. For example, from Fig. 2.2.14, the general sentiment towards “riot police” is very negative; however, on 14 Oct it suddenly becomes positive. It is worth investigating what actually happens that day.

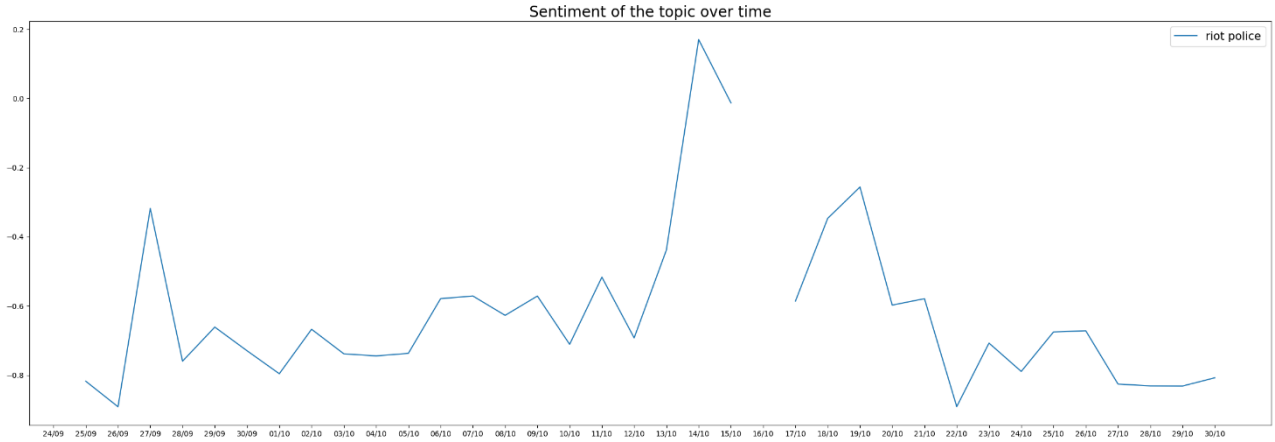

Fig. 2.2.15 Sentiment with mean scores of the topic “riot police” across dates

Combine Sentiment and Word Cloud

After the sentiment analysis, we attempted to generate word clouds with colours of words reflecting the sentiment of the users towards the phrase or word. The green colour represents positive sentiment, red colour represents negative sentiment and the default colour is grey.

After initial trials of word cloud generation with recolouring based on sentiments of the topics, we realized that most words are coloured in grey as they are not part of the positive or negative topics. However, some of the words do carry a positive or negative connotation. As such, we sourced a list of positive and negative words from two papers, Mining and Summarizing Customer Reviews by Minqing Hu and Bing Liu and Opinion Observer: Analyzing and Comparing Opinions on the Web by Bing Liu, Minqing Hu and Junsheng Cheng.

Fig. 2.2.15 is the recoloured word cloud generated from all tweets using bigrams without hashtags.

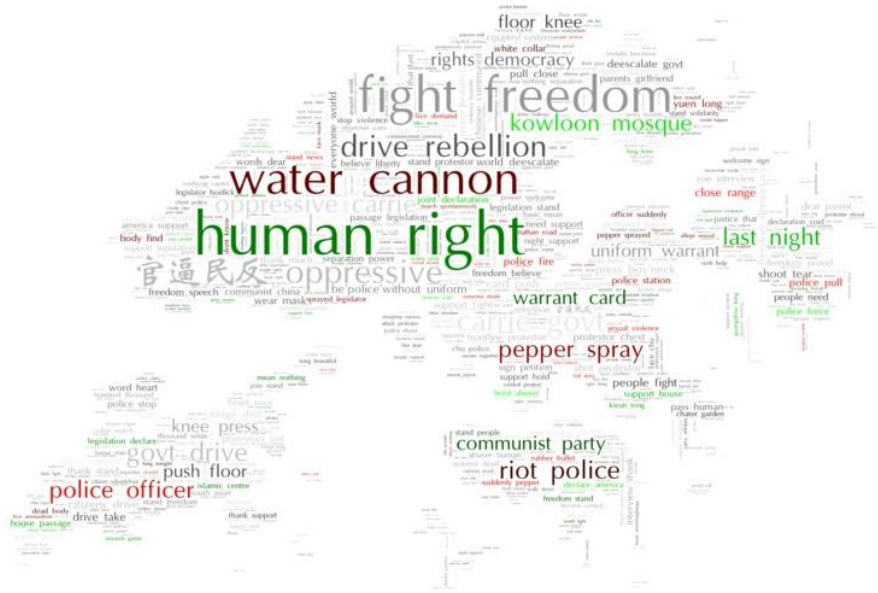

Fig. 2.2.15 Re-coloured word cloud based on sentiment of words and phrases (green: positive sentiment; red: negative sentiment; grey: default)

2.3 Network Analysis

Data Collection and Cleaning

Due to the nature of network, random sampling will disrupt the user retweet network. In order to have a more representative network of how Twitter users interact with one another through retweeting, the team used Twitter retweet data of suitable size on a daily basis from 24 Sep 2019 to 30 Oct 2019.

Due to the constraint of the availability of the data from Twitter and manpower of the team, relevant tweets on Twitter were collected by two of the team members so as to have a fuller picture of the daily tweets with regards to Hong Kong Protests.

For the purpose of network analysis, the two sets of data from the two team members were split, combined and cleaned to remove duplicates. The cleaned daily Twitter data were then used to generate GML files with python scripts. (Appendix X)

Network Analysis Using Gephi

Gephi is the primary tool the team used to analyse the networks from the 34 GML files. The size of the networks varies vastly because of the availability of data and relevant events happened in Hong Kong.

Gephi Algorithms Used

For smaller networks, more complicated algorithms such as Fruchterman-Reingold were used. For larger networks, simple and fast algorithms such as OpenOrd and ForceAtlas 2 were used.

Network Graph

Please refer to the table below to understand the network graph.

Node size and label size	Proportional to the degree centrality of the Twitter user.
Node colour	Modularity the nodes belong to, with the largest modularity being coloured red #E33A34

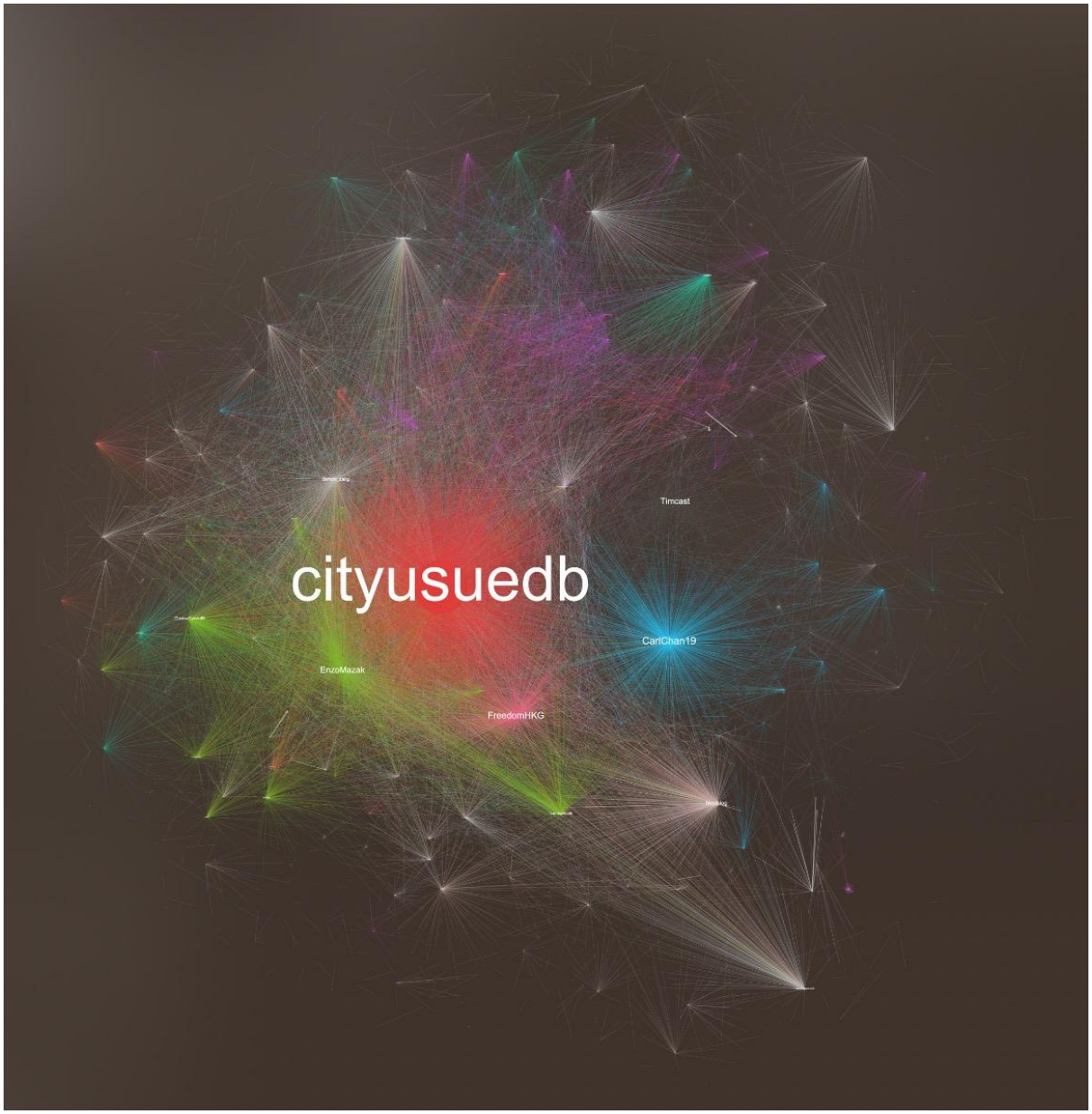

Figure 2.3.1 Network Graph

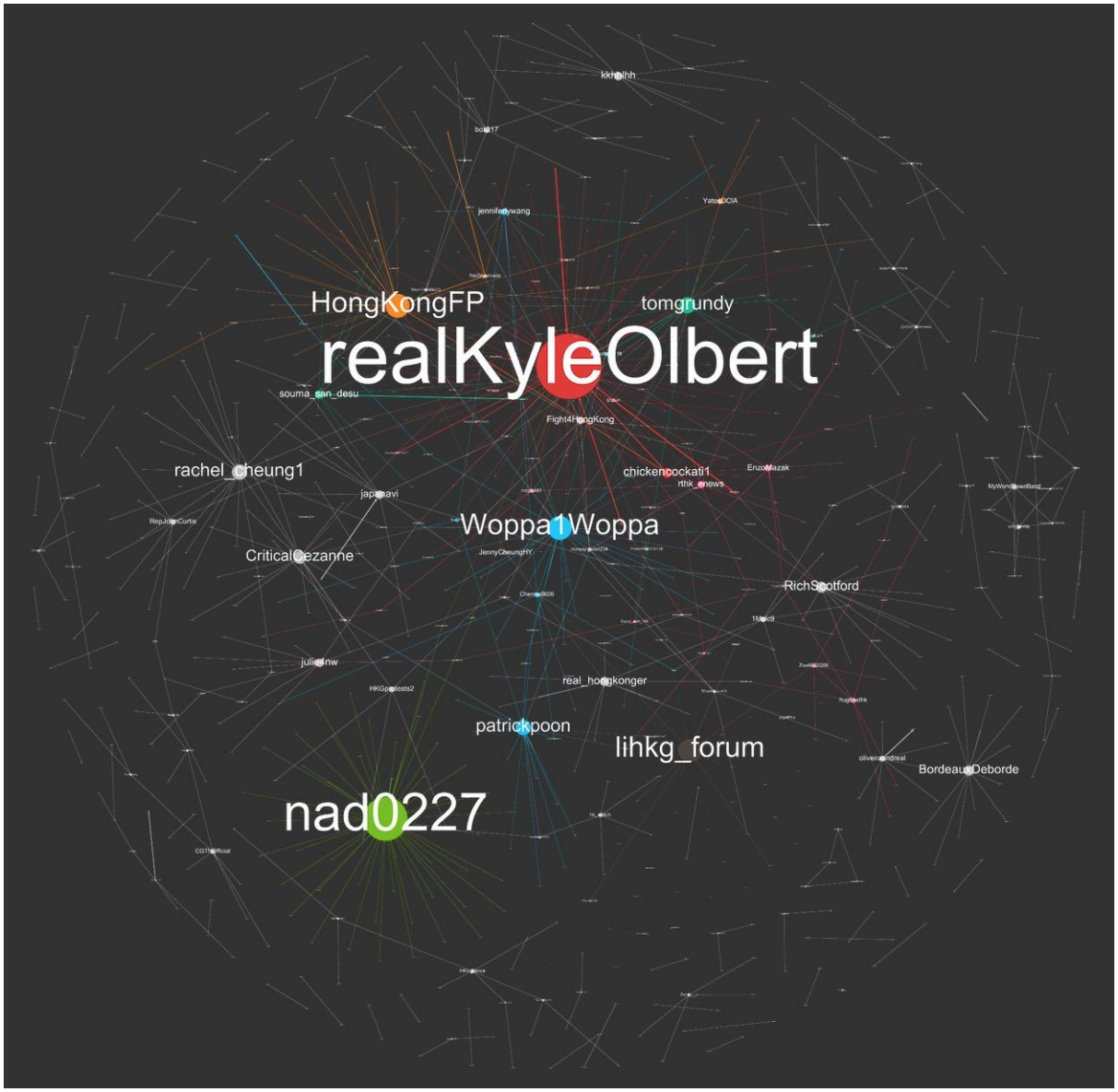

Figure 2.3.2 Network Graph

Network Analysis Using Networkx and Pandas Library on Python

Important network measures such as degree centrality, closeness centrality and betweenness centrality are calculated using the networkx librarying on python. These data are stored in DataFrame in pickle files for fast processing. (Appendix X)

Our Findings

From Network Graph

Looking at the pattern of network graphs generated from 24 September to 30 October, we can see that there will always be one or a few nodes with extremely high degree centrality, regardless of the size of the network graph. This illustrates that the general network structure remains the same, with a few prominent nodes at the centre of the graph. There are some nodes that are constantly being one of the top nodes in every network graph like @HongKongFP and @Woppa1Woppa. These twitter users are some of the most influential users in the Hong Kong protests related issues' network. The user @Woppa1Woppa for example, is a Hong Konger-both-Canadian covering the Hong Kong protests issue in English, he serves as an "insider" for the Hong Kong issues to the English-speaking western world. Furthermore, his personal content regarding the Hong Kong issues can be very influential to people from the west.

However, the most prominent nodes in each of the network graph are not always the same, we can see that some prominent nodes in one of the network graphs may disappear in the next day's graph. We have picked some of the most prominent nodes that only appears for a short period of time in our network. For example, in the network graph of 13 October, the most prominent node is @Abrelialia with a high degree centrality of 0.643, however the user has a drastic drop in its degree centrality the day after and disappear afterwards. We went on to look at the twitter profile of this user and found that it has been suspended. There are multiple possibilities to this finding, the user can be less active and stop tweeting on Twitter after 13 October, or it has been suspended quickly after 13 October due to the violation in the twitter rules.

The network is also always divided into different communities as represented by different colours. Within each community, there will be one local influencer as shown by the largest size node.

From Network Measures

We calculated the three important network measures such as degree centrality, closeness centrality and betweenness centrality for the 34 networks individually. The normalised centrality measures are then combined together to compare the networks. From the aggregated centrality measures sorted in descending order with respect to each network centrality measures shown below, it is worth noting that top five Twitter users across all three measures are almost the same individuals. For example, Twitter user named @*cityusuedb* and @*SolomonYue* scored top two in all three measures. Twitter user @*cityusuedb* is the media run by students from City University of Hong Kong, and it has less than five thousand followers. Twitter user @*SolomonYue* is located in the US and he describes himself as '*vice Chairman & CEO at Republicans Overseas, RNC Member since 2000, Co-founder of RNC Republican National Conservative Caucus & Conservative Steering Committee'.* With 89.8 thousand followers, the individual has great influencing power over his followers as his tweets and retweets can be seen by many.

```
data.sort_values(by=['degree centrality'],ascending =False)
```

	User name	degree centrality	betweenness centrality	closeness centrality
221158	cityusuedb	0.845648	0.691779	0.467904
139168	SolomonYue	0.805688	0.581093	0.427188
196287	alexangel8577	0.727803	0.469120	0.290815
216338	Catherinecat826	0.669312	0.464046	0.419341
47669	lihkg_forum	0.662256	0.474655	0.416609

```
data.sort_values(by=['closeness centrality'],ascending =False)
```

	User name	degree centrality	betweenness centrality	closeness centrality
221158	cityusuedb	0.845648	0.691779	0.467904
139168	SolomonYue	0.805688	0.581093	0.427188
214140	icomefrommars1	0.003821	0.004334	0.419831
216338	Catherinecat826	0.669312	0.464046	0.419341
214582	May41204747	0.002939	0.004677	0.418566

```
data.sort_values(by=['betweenness centrality'],ascending =False)
```

	User name	degree centrality	betweenness centrality	closeness centrality
221158	cityusuedb	0.845648	0.691779	0.467904
139168	SolomonYue	0.805688	0.581093	0.427188
196294	abigailafhk	0.510045	0.561232	0.342164
43262	Catherinecat826	0.657658	0.494205	0.333695
47669	lihkg_forum	0.662256	0.474655	0.416609

Figure 2.3.3 Degree Centrality Table of Network Graph

3. Reddit Data Analysis

3.1 Data Collection and Data Description

Data Collection

The social media website Reddit is a popular hub for online discussion. A Subreddit in the website is a community or forum where people of similar interests could discuss about a specific topic. It has posts as well as corresponding comments associated with it. In our project, to collect Hong Kong Protest related data, we would choose Subreddit Hong Kong (https://www.reddit.com/r/HongKong/) as our source.

To scrape data from Reddit, we use Reddit API and Python Reddit API Wrapper, PRAW. Reddit API allow us to get user submitted and rated content on the website and data returned is JSON object. PRAW is a python package that give us access to Reddit API.

One key feature of Reddit is that all the content shown on the home page are sorted. They could be sorted using 5 methods: new, hot, controversial, rising and top. Reddit API also return data sorted in these 5 methods. Because Reddit API as a limit of 1000 posts per scraping, to increase our data size, we use all the 5 methods to sort and scrape.

Data in HK subreddit is scraped daily from 26 Sep to 29 Oct, based on the 5 sorting methods mentioned above. Since the data amount is already limited by Reddit API, it won't take lots of computational power to run the code. Thus, data scraping for reddit is done locally. It takes around 1 hour to scrape data from Reddit every day.

Although data returned from Reddit API is JSON object, for easier processing, we create a submission instance from HK subreddit, iterate through it to get data needed, store them in Pandas data frame and export as CSV files.

Data Description

Data from Reddit includes following fields and all the csv files generated daily were combined to one static CSV file and stored locally.

1. Author
2. Title of submission
3. Score of submission
4. Id of the submission
5. URL of submission
6. Number of comments under this submission
7. Time created
8. Body of the submission
9. Comments of the submission

Comments under each submission can be scraped by looping through the *comment* attribute. PRAW provides a CommentForest and it could return all the top-level comments, followed by second level and third level, etc. While we could use breadth-first traversals to go through these data, it would cost too much computing power. Thus, we simply use the *list* method and store comment text in a list.

3.2 Data Cleaning and Processing

Compared to twitter data, Reddit data is much cleaner. Thus, we take the following step to clean and process our data.

1. Remove duplicate
2. Remove data with empty submission body and empty comments.
3. Transfer date created to date time value
4. Transfer emoji to text
5. Remove URL.

To clean Reddit data, firstly we remove all duplicate posts based on “id” and number of comments. Because of reddit’s sorting mechanism, we don’t have the freedom to select data in Subreddit HK based on our own criteria. For each sorting method, only the top 1000 submissions would be returned. Duplicate of submissions is

common among data scraped from these 4 sorting methods. However, same posts scraped at different time could have different score and different number of comments. For same submission, we would choose the one with highest number of comments. “id” is used to identify unique submissions and remove duplicates.

In reddit, submissions sorted by New or Rising sometimes contain only submission titles. Submission bodies and comments would be empty. These data won’t provide much value for our analysis, so they are removed.

Time created is represented in UNIX Time, so we transfer it to date time value for future easier analysis. Same as twitter’s cleaning method, emojis are transferred to text for sentiment analysis.

Comments, post title and post body could contain lots of URL that link readers to other articles or post on other popular social media websites such as twitter, YouTube, Facebook and Instagram. All the URL in submission titles, bodies and comments are removed.

3.3 Data analysis and insights

Sentiment analysis

To analyse the sentiment for reddit posts, we use the same analysis method and calculate the Vader score for submission title, submission body and each comment under the submission.

Each submission would represent a fact or opinion regarding specific events or news related to Hong Kong Protest. We are interested to know, in general, how people feel about these events and topics. Thus, we aggregate the average sentiment for each submission and calculated the mean compound score to represent the sentiment for that submission.

Sentiment Changes over Time

Below are the changes of average comment sentiment and title sentiment over time. We further aggregate the score to get the mean of each day. Data scraped from

Reddit are from June till October. However, since we only started scrapping in September, data size before September are small. So, we drop these data because they won't be representative.

From the Fig. 3.1 and 3.2 we can see that, average sentiment and title sentiment fluctuated over time from September till October. Key event would significantly affect the public sentiment.

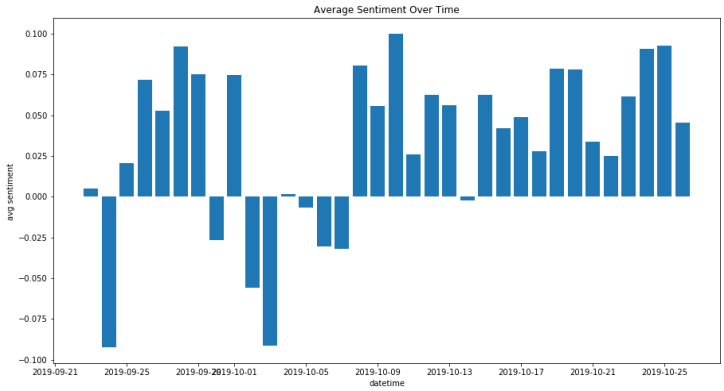

Fig 3.1 Average comments' sentiment changes over time

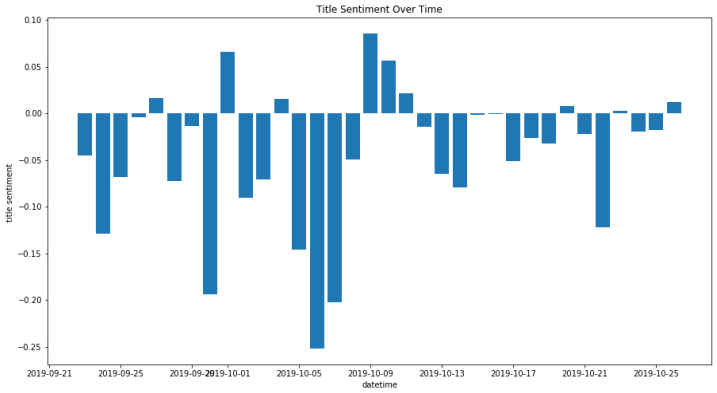

Fig 3.2 Title sentiment changes over time

For example, Average sentiment on 1st October and 2nd October are extremely negative. 1st October is the national day of mainland China and it is also the 70th anniversary of the founding of the People's Republic of China. In contrast to the celebration in Beijing, Protesters call it "Day of mourning", which possibly explains their negative sentiment. Thousands of protesters marched to protest the local

administration and Chinses government. Later in the day, Volleys of tear gas was used during the clash and 3 students were shot by police, which lead to another protest on the next day. Discussions and condemning online also continued for the following several days.

This result could help our clients to measure the effect of certain events. Besides numbers like turn out rate of march or number of people attended, sentiment score would also be a data source to support news articles.

One interesting finding is that, in contrast to large proportion of negative title sentiment, most of the average comments' sentiments are positive. Based on our observation of posts and comments on reddit, posts are usually about negative events such as violence, clash, police shooting protests, usage of weapons, etc. However, a large proportion of comments under these posts are still positive and people would express support.

We also calculated the correlation between body sentiment, title sentiment and average comment sentiment. Graph 3.3 shows the scatter plot of body, title and average comment sentiment. As the correlation coefficients show in graph3.4, there is nearly no correlation between these 3 sentiment scores. This suggests that sentiment of the post itself won't significantly affect the overall comment sentiment. No matter what issues and opinions are stated in the posts, people's stand and sentiment still vary.

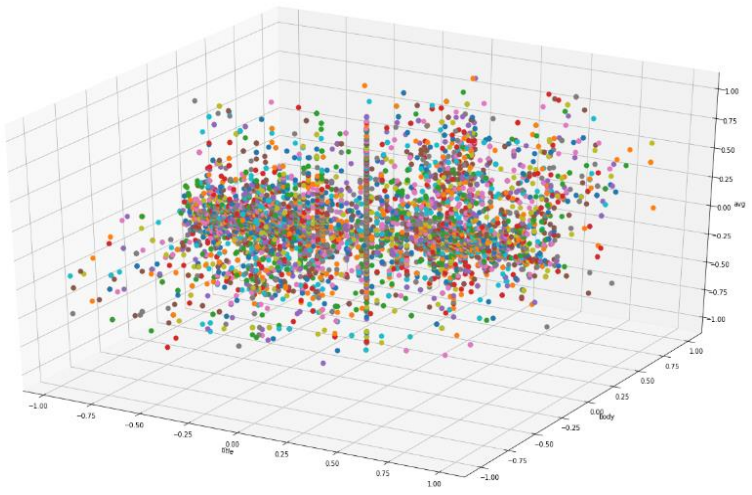

Fig 3.3 title sentiment over body sentiment and comment sentiment

	title semtiment	body_pol_list	avg_pol_list
title semtiment	1.000000	0.111356	0.250348
body_pol_list	0.111356	1.000000	0.115988
avg_pol_list	0.250348	0.115988	1.000000

Fig 3.4 Correlation Coefficient

Topic Sentiment

Besides overall sentiment analysis on Reddit data, we also want to extract key topic and analyse on these topics. Topic sentiment would help our clients identify key topics and have a more detailed understanding of people's attitude regarding certain topics.

We would extract common topics and their negative, positive and neutral scores. To find key topics, we use text in all the comments. After removing stop words from nltk's English stop words list, comments texts are broken into sentences, followed by part of speech tagging. Then we search for noun and noun phrases as our aspects. If the tag of the word is NNP (Proper Noun) or NN (common noun), and the word before it is also NNP or NN, they are considered as a noun phrase. If the word itself is a noun but the word before it is not, it's considered as a noun word. Aspects that occurred less than 50 times are eliminated. Then we identify opinion words that are related to these aspects within the sentence. JJ (adjective), JJR (adjective comparative), JJS (adjective, superlative), RB (Adverb), RBR (adverb, comparative), RBS (adverb, superlative) would be opinion words. If an opinion word is negative, the negative score for the corresponding aspect plus one. To handle negations, we create a list of negative words. If the opinion is negative relation, its sentiment score is reversed. If an opinion word is negative, the negative score for the corresponding aspect plus one, and vice versa. Then, we divide the negative, positive and neutral scores by number of total counts to get percentage for each sentiment category. Charts and tree maps are generated using Tableau.

Gf3.5 is the overall sentiment for key aspects. Red represents negative sentiment; green represents positive and grey represents neutral. From the graph we can see

that, most of the comments towards key aspects are still neutral. In general, positive proportion s larger than negative proportion. This could be due to the characteristic of the Reddit platform and their user type.

As a discussion forum, the comments made on reddit are more thoughtful and rational. People are more careful about their words and expressions. There are also rules for posting and making comments under Hong Kong Subreddit. For example, sensationalized contents are restricted.

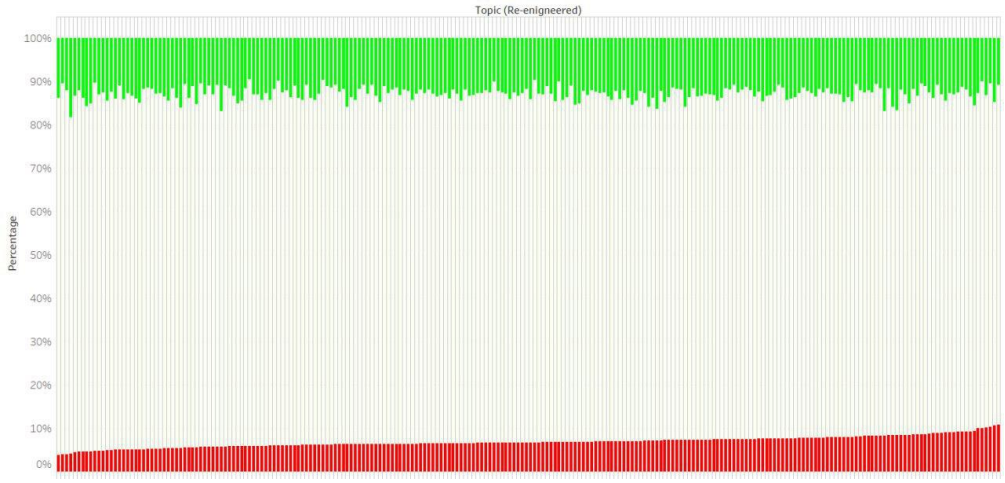

Fig 3.5 All topic sentiment

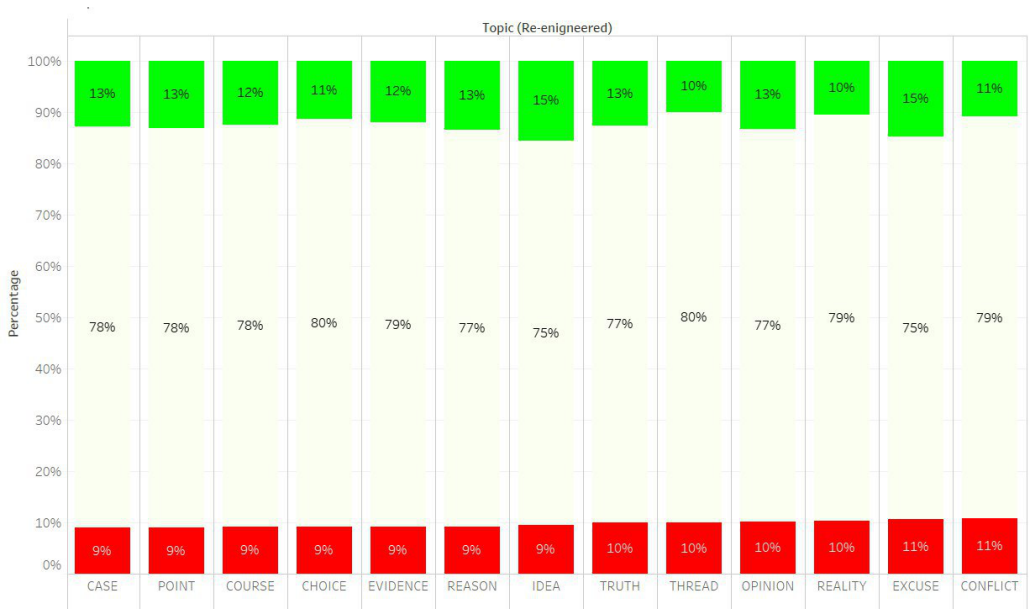

Fig 3.6 Most Negative Topics

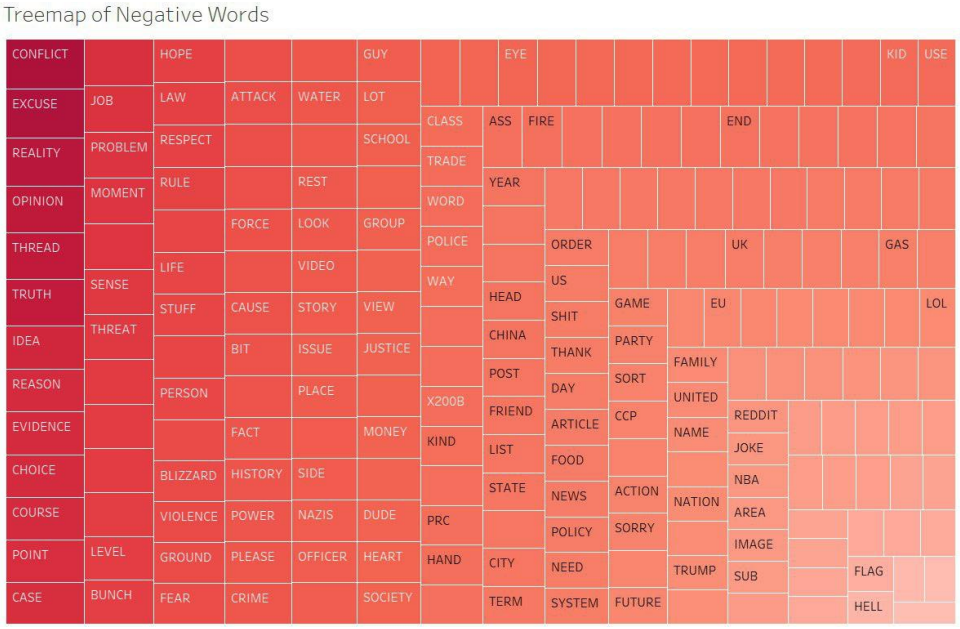

Fig 3.7 Tree map of Negative Words

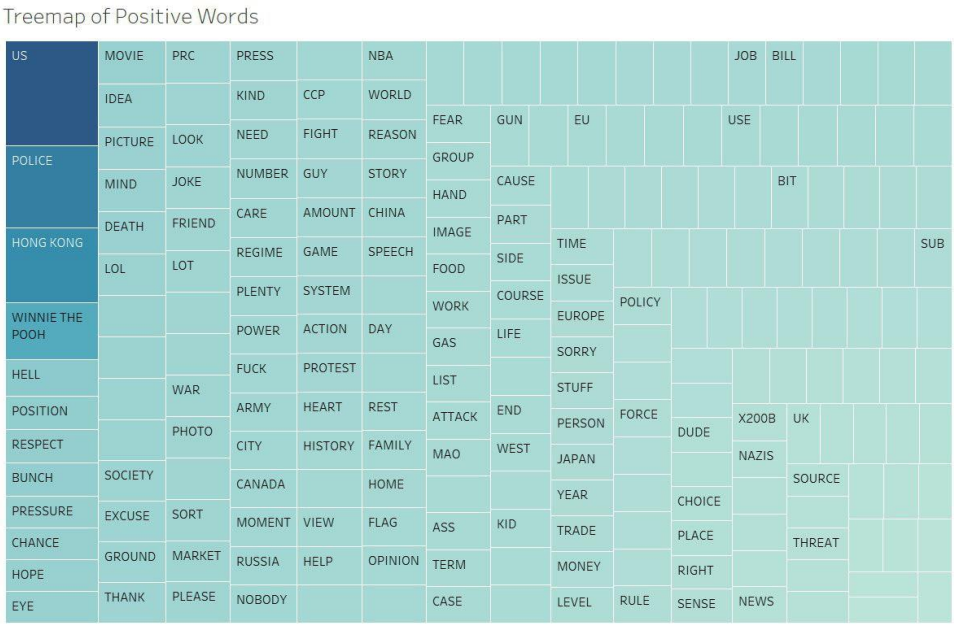

Fig 3.8 Tree map of Positive Topics

From the tree map of negative topics and most common negative topics, we find out that “reason”, “truth”, “evidence” and “reality” appear frequently. This shows that besides events or news themselves, people care about the ground-truth of these issues and want to figure out reasons behind them. They also talk about evidence to prove their opinions.

Positive topics includes “WINNIE the POOH”, which represent Chinese President Xi JInPing. Police and US also have high proportion of positive comments. This suggests that all these regimes have people supporting.

3.4 Word Cloud

Fig 3.9 Reddit Word Cloud

To generate word cloud from reddit data, we use all the text from post title, post body and post comments. We tokenize texts, remove stop words based on NLTK’s stop words list and correct spelling errors for words using the correct method in Text Blob. Same as the approached used in twitter analysis, “Hong”, “Kong” are also removed.

As we can see in the graph, words like “government”, “human right”, “CCP”, “US” are frequently mentioned by reddit users. This suggests that, people’s focus on Hong Kong protest is about the relationship between different powers. China government, Hong Kong police, protesters and US government are key players in this issue.

4. Limitation

4.1 Twitter Data Analysis

For our twitter data analysis, as we only sampled at most 5000 tweets during a specific period of time for each day, the insights and conclusions drawn from our results may not be representative to the whole Twitter user population. This is an inevitable consequence of our limited computational power, even though we scraped about 1.7 million tweets.

On top of that, as we used Twitter Streaming API, what tweets we crawl in the end heavily depends on the time and duration we called the API to scrape. Thus, the tweets we crawled may not be representative in the first place.

Moreover, as mentioned in the reason why we proceeded with popular topic extraction, the word cloud library does not allow us to have reserved pairs for bigrams, thus the word cloud visualization may include “weird phrases” and not highlight the key phrases. Furthermore, even if we sourced the list of positive and negative words, the word cloud seems not to incorporate them for bigrams as the re-colour function needs to match the exact phrases. Due to limited amount of time, we could not fully customize the word cloud as we wished.

4.2 Reddit Data Analysis

For our Reddit data analysis, due to the sorting algorithm of Reddit, the data is not randomly sampled, and we don’t have the freedom to choose data based on our own criteria. If we sorted by new and choose the latest posts, all the results returned could be within several hours before the time we scrap, and we cannot get data at other time over time. And while we get the top 1000 submissions daily, there could be overlap of data day by day since the hottest submissions could remain there for several days. Sorted by rising would return submissions that is getting a lot of activity right now and the data here could be very limited. The result will be influenced by the time that we scarp. Even though we try to scrap at the same time every day, we still cannot make sure data are randomly sampled.

Second limitation is that there is no network analysis on reddit data. Because of the data structure of CommentForests. In PRAW, it's very complicated to get reply network in reddit. We must go breadth-first traversal which would cost a lot of time and computing power. Besides, the network of comments is not as representable as Twitter user network. Comments and replies are under each submission, so it's hard to get an overview of user network. As a discussion forum, reddit users care more about topics rather than key influencers. Thus, we didn't perform network analysis on reddit.

5. Team Contribution

The project was done in two major phases. Phase 1 Data Extractions & Manipulation and Phase 2 Analysis & Insights Generation. The team members have different responsibilities during the two phases to produce the final product.

5.1 Phase 1 Data Extractions & Manipulation

- Data Processing (hashtags, emojis, languages): Haodi
- Data Collection (Twitter): Haodi & Hanyu
- Data Collection & Processing (Reddit): Chengzi

5.2 Phase 2 Analysis & Insights Generation

- Text Analysis (Sentiment Analysis, WordCloud and Topic Extractions & Analysis on Twitter): Haodi
- Text Analysis and Network Analysis (Reddit): Chengzi
- Network Analysis (Twitter): Hanyu & Wende
- Tableau Data Visualisation (Twitter & Reddit): Wende

5.3 Final Deliverables

- Report: The Whole Team
- Presentation Slides: The Whole Team
- Final Poster & Abstract: Wende

The team had reached a common consensus that everyone gets the same score for this project. There are few justifications to this result. Firstly, the team had distributed the work fairly based on each other's strength & capabilities. Secondly, the work from each one of us are equally valuable to the project. Thirdly, the time and efforts contributed to the project are largely equal. Last but not least, all team members have been actively involved in the discussion which drew many insights that value-add to the project and our findings. In conclusion, we have decided to assign the same score to all of us.

Team Member	**Score**
BU Wende	5
JIANG Hanyu	5
QI Haodi	5
ZHANG Chengzi	5

6. Conclusion and Future Direction

In conclusion, several analyses we have drew many insights that can be beneficial to the clients. For text analysis, a list of popular phrases discussed by users on social media network are generated. The list of phrases is assigned with sentiment score and visualised using WordCloud and Tableau to show public sentiment towards each topic. Network analysis on the other hand provided us with vastly different types of findings, allowing us to look at the network visually and see the structural changes of network over time. It also allows us to identify different users that are prominent in the network. Overall, our research findings will be beneficial to all our clients who are working on this issue, providing scientific & statistical reference for researchers to do analysis and news media agency to look for prominent figures involved in the issues.

Future Direction

Due to the limitations of the project faced, our team could only show the “tip of the iceberg” for this issue that is full of insights. We believed that this computational social science analytics project can be more valuable in the future as it has the scalability and wide application to different topics of social sciences.

For scalability of the project, we believed that with more reliable sources that contain ample amount of data will provide us with even more and accurate insights. Computational power is also one of the issues we have faced that causes a bottleneck to our project, limiting the amount of our results produced. Hence, higher computational power (whether is local machines or cloud usage) will allow us to come up with more data for analysis.

For application of the project to other topics, our team feels that this project can be applied to any social science-related issues. This computational social science analytics project has been done in a very systematic manner, a standardised workflow of data extractions, manipulations, cleansing, analysis to visualisation allows us to look for interesting findings more conveniently. This workflow can be applied to other topics that allows different clients to know more insights about various social science issues.

- End of Report -