

IS434 Social Analytics and Applications

**Computational Social Science Analytics Project**

(Insights from *Hong Kong* Protests)

**GROUP 18**

Bu Wende

Jiang Hanyu

Qi Haodi

Zhang Chengzi

**INTRODUCTION**

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, news media agencies and reporters worldwide are looking for meaningful and sustainable sources on public sentiment in order to come up with eye-catching articles and even posts with in-depth analysis. Online sentiment and different communities involved in the issue are also valuable to social science researchers and academia for deeper analysis. Hence, our team sees the value of doing an analytics project that focuses on the social media analysis of the 2019 Hong Kong Protests and seeks to find out deeper insights for the potential clients.

**CLIENTS AND BUSINESS PROBLEMS**

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.

**DATA & PLATFORMS**

In our project, we have chosen a list of social media platforms that have an ample amount of data for our analysis. The following are lists of data we are going to scrape from the social media platforms

* **Twitter**
  + Tweets data like user location (coordinates), retweet & reply data (including users who retweeted), created time, the text of the tweets, followers and/or following counts, etc.
  + Use Twitter API to scrape tweets under some popular hashtags related to the issue. Data will be collected on a daily basis with a real-time listener, especially when there is major event outbreak related to Hong Kong. Standard search API will also be used to scrape the tweets of the past week to capture data that is missed out.
* **Reddit**
  + Number of likes, number of comments, comments, replies to comments, user data, timestamp
  + Use Reddit API to scrape data under the Sub-Reddit category of “Hong Kong”, data of the Reddit submissions and their comments will be scrapped.
* **YouTube**
  + Data like comments, reply, number of likes (to select the popular videos of the specific topic)
  + Using Google developer account to scrape comment data from YouTube Video. The plan is to first identify the most popular videos regarding the HK protest, and goes to each video’s comment section to scrape the comment data.

**ANALYSIS & INSIGHTS**

Currently, we have identified several data analysis methods to handle the datasets we are going to obtain. Through different methods of analysis, we hope to draw insights that are valuable for our clients from the issue. The list of methods is non-exhaustive:

* Sentiment Analysis
  + The general sentiment of the user population;
  + Changes in sentiment over time with key incidents;
  + Sentiment of different regions and how does it differ, how geographical location or people with different cultural backgrounds affect their perspectives on the Hong Kong issue.
* Network Analysis
  + Identify key stakeholders in the social networks based on different centrality measures;
  + Identify communities of users on different platforms, especially twitter and their common opinions;
  + Analyse the network structure and characteristics.
* Clustering Analysis
  + Text clustering using cosine similarities to categorize tweets, Reddit submissions and Youtube video descriptions, and to uncover hidden topics;
  + Word Cloud (with Time Series).

**TOOLS**

To achieve the above-mentioned analysis and get the desired information & insights, our team will be utilising Python as our main programming language and leverage on different libraries for data analysis purposes. Visualisation software and Application Programming Interface (APIs) will also be used to manipulate and visualise data to generate insights.

*Platforms*

* Jupyter Notebook and Visual Studio Code as the Programming IDEs
* Amazon Web Services

*Data Sourcing*

* APIs such as Twitter API and Reddit API
* Data scraping libraries for Python such as Beautiful Soup (BS4), Selenium, etc.

*Text Analysis*

* NLTK, Textblob and Gensim libraries

*Network Analysis*

* Networkx library

*Visualisation*

* Gephi (for Network Analysis & Visualisation)
* Tableau (Information Visualization Dashboard)
* Wordcloud library