Linking Public Opinion to Direct Action

2019 Hong Kong Protests

Winfield Chen, Xiangpeng Hao, John Sweeney

Simon Fraser University

{wca88,xha62,sweeney}@sfu.ca

Abstract

When Facebook Messenger finds intent in a message, it suggests helpful prompts like Location Sharing and Money Transfer. For example, sending the message "I'm here" reveals a temporary "Send Location" button. Can similar analysis be done on the public opinion level for news events like protests? What language is linked with the public opinion surrounding an event? How do movements grow and organize? Intent and sentiment analysis has many applications beyond Messenger chats. With recent protests, strikes, and other forms of direct action around the world, researchers, policy makers, news organizations, and activists find it useful to gauge public opinion in times of crisis and link trends in language in online conversation with real-world events. The rise of instant messaging has led to decentralized and leaderless movements where protesters collaborate by jointly brainstorming and voting on strategies. Despite media coverage, little research has been devoted to studying the data on social media generated by these movements. We use sentiment analysis and a document vector model on social media posts and news articles to study the Hong Kong protests. We find that both techniques are able to link online trends to events occurring in the real world.

Introduction

The Hong Kong protest movement has been primarily using two platforms to organize: Telegram and LIHKG, a Hong Kong version of Reddit. We used posts from these platforms in combination with news articles as inputs to sentiment analysis and a document vector (doc2vec) model to better understand the connections between protest-related events and the posts. For sentiment analysis, we first identified trends in "hot words" for monthwise sentiment analysis using a Baidu API. For the doc2vec model, we trained a 50-dimensional document vector representation for each post and attempted to count the number of similar posts given a query using cosine similarity. Thus, if the query is protest-related, then the count would be the number of protest-related posts, an estimate of the amount of public discourse surrounding the protest. By selecting posts of different levels of passion and position on the protests, we would then be able to gauge public perspective. With this information, we can link posts to real-world events and compile a summary of the important events of the protest and gain insight into the movement's motivations.

Data

We used 156 computers in the Computing Science Instructional Labs (CSIL) to scrape LIHKG. This was a challenging task that took several days and attempts. Fortunately, LIHKG assigns thread IDs by incrementing from one. Coincidentally, thread 1000000 was posted on January 28 2019, just before the Hong Kong extradition bill gained notoriety. This left about 770000 threads, not accounting for deletions, posted from then until December 4 2019 when our scraping began. This big dataset required a distributed architecture, as using only one computer would require several days. We designed a script that would evenly distribute the web scraping across machines.

In our first attempt, the browser automation package Selenium was used with the Firefox. A browser automation approach was chosen due to LIHKG being reliant on dynamically page loading via Javascript and some technical limitations which restricted API access. This attempt extracted posts from the page source and resulted in HTML output. Not only was this slow, requiring several hours, but it also required HTML parsing to recover templated fields. The omission of some critical fields in this first attempt necessitated a second attempt.

In our second attempt, we queried LIHKG's API for one gigabyte of 7750 Zstandard compressed JSON files. We bypassed the technical limitations to API access by using Selenium to navigate Firefox to LIHKG's homepage to retrieve the session cookies and HTTP headers used by Firefox. The API queries were then conducted using Requests. Since the API was scraped rather than the site, this attempt only took one hour.

The dataset is nearly complete: two of the machines were shut off during the scraping, so the files for threads 1191216-1191315 and 1246234-1246333 were corrupted and 1191316-1196247 and 1246334-1246564 were lost. These intervals correspond to missing data in 2019-06-07T19:50:52 to 2019-06-09T18:34:03 and 2019-06-20T15:25:41 to 2019-06-20T18:33:57 UTC. We note that recovery of the Zstandard frames from the corrupted files may have been possible but was not attempted.

Methods

We moved the data to SFU's Hadoop cluster and used Apache Spark to filter out the two corrupted files, extract the first post of each thread, select interesting features, and flatten the JSON for compatibility with Pandas. The preprocessed JSON files were concatenated into a Pandas dataframe containing unsegmented text containing HTML tags.

The second preprocessing step took this dataframe and stripped all HTML tags from each post using Beautiful Soup and segmented the text. The word segmentation task was done using Jieba. Then, Gensim was used to train a document vector (doc2vec) model. By classifying two posts as "similar" if the cosine similarity of their corresponding vectors is greater than a, we select an affine subspace of the posts by document similarity, and can control the number and degree of similarity of such posts by tuning a.

We then aggregated the segmented text and counted word frequencies over time. We used Baidu's sentiment analysis API on the posts to detect positive and negative tone. While Baidu has the best Chinese-language cognitive service, its APIs are slow and expensive. To accomodate these limitations, we grouped the posts by month and submitted uniformly random samples of one thousand posts per month for sentiment analysis. This uniformly random sampling allowed us to process thousands of posts quickly while the analysis remained an unbiased estimator of actual sentiment.

Results

We found that our doc2vec model was able to, when queried with a text related to an event from the Hong Kong protests, return related posts from LIHKG with document cosine similarity score σ greater than a, where a is some threshold-of-similarity hyperparameter. Since $\sigma \in [-1, 1]$, an arbitrary $a = 0.5$ was chosen which resulted in a reasonable number of similar articles for queries of around one thousand characters. We also found that when the queried event text was smaller, greater values of a were needed to control the number of similar articles.

The distribution of the weekly number of these related post queries was found to have peaks at the corresponding time of the queried event. The following is an example post and the weekly number of similar posts.

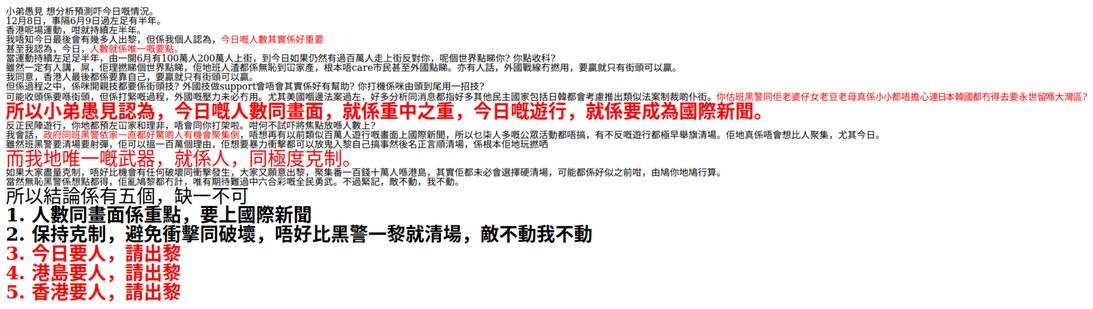

Figure 1: A protest-related post. The colour and emphasis are original.

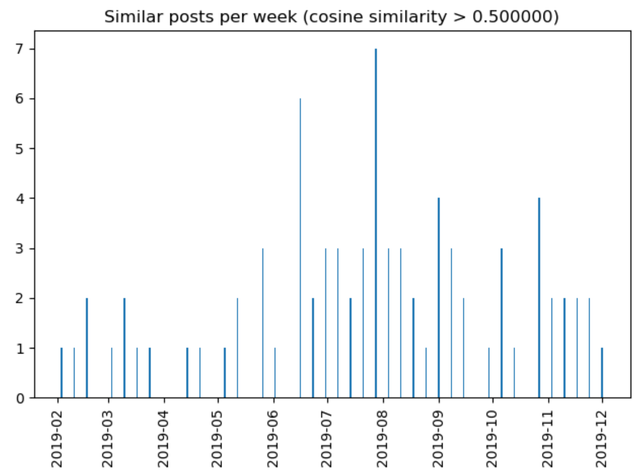

Figure 2: The weekly number of similar posts.

This post in particular was chosen for its length of about one thousand characters being appropriate for $a = 0.5$ and for being outside of the dataset. The distribution of similar posts peaks in the fourth week of July. In this time period there was a march on the 21st, the first airport sit-in on the 26th, tear gas, rubber bullets, and violent clashes on the 27th and 28th, and solidarity rallies on the 30th and 31st. The second peak occurs in mid-June, when organizers claimed 1 million people marched on the 9th to protest the extradition bill's second reading, the "612 Incident" occurred on the 12th when protestors surrounded the Legislative Council, when the bill was suspended on the 15th, and two million protestors assembled on the 16th. The alignment of these significant events with peaks in the distribution shows that our doc2vec model can succesfully link social media conversation to real-world events, and perhaps with a higher resolution and larger corpora predict such events before they happen. The following figures show the results of our hot words and sentiment analysis.

Figure 3: Hot words for (a) February-July (b) August-December.

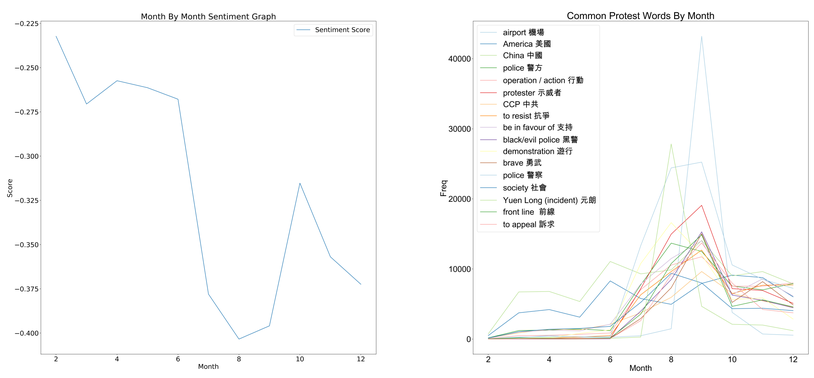

Figure 4: (a) Monthly sentiment (b) Monthly protest words.

Conclusions

- Data collected indicates significant changes in sentiment towards active resistance of the government and civil disobedience starting in July and August.
- Large-scale data collection and computational tools can be useful to help identify protest events and to extract those elements relevant to the study of social movements.
- Further research projects include automatically predicting future protests and analyzing protest data for novel causes.

References

[1] Hanna, A. (2017, January 10). MPEDS: Automating the Generation of Protest Event Data. https://doi.org/10.31235/osf.io/xuqmv