# Around the World Coverage

## Classifying the Differences in News Coverage of the Hong Kong Protests

The George Washington University
Natural Language Processing - DATS 6450
Michael Siebel Junchi Tian Bixuan Huang

### 1 Introduction

This summer, protests broke out in Hong Kong, at first related to an extradition treaty that the protesters viewed as an encroachment of mainland China into the special administrative region of Hong Kong. U.S. and Chinese newspapers largely take a different view on these protests. U.S. newspapers are more likely to frame them in terms of protests for democracy, while Chinese newspapers are more likely to focus on the violence of the protesters.

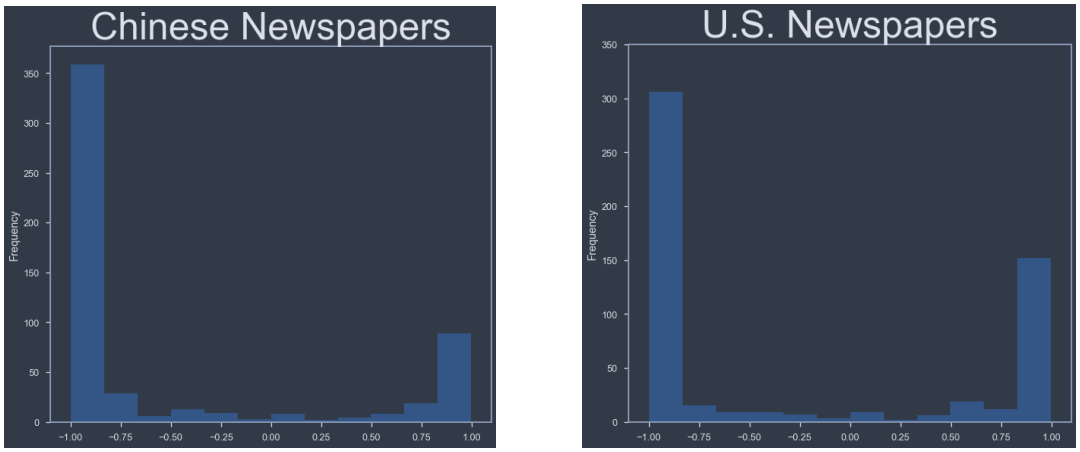
Our goal is to determine what natural language processing (NLP) techniques can best distinguish between U.S. and Chinese reporting on the 2019 Hong Kong protests. We evaluate our attempts to make this distinction by setting up a classification model in which we predict newspaper articles' country of origin, half of which originate from U.S. newspapers and half of which originate from Chinese newspapers (English-language versions). Our evaluation will center around three key parameters within a bag of words methodology: minimum sparsity thresholds, terms weights, and n-grams. In attempting to build a successful model, we will assess several model configurations that vary these parameters and describe any meaningful improvements that these parameters provided. Before we conduct such analysis, we will provide extensive exploratory data analysis in order to provide context on our data.

### 2 Corpus

We collected 1,101 news articles on the Hong Kong protests; 552 were from U.S. newspapers and 549 were from Chinese newspapers. U.S. newspapers included the Wall Street Journal, Washington Post, and New York Times. Chinese newspapers included China Daily, People's Daily, and Xinhua Agent. While our articles were mostly balanced in terms of quantity between newspapers, U.S. articles were longer, averaging 557 words compared to the Chinese average of 306.

We conducted some exploratory data analysis after term case transformation, stemming, and stopword removal. We start by analyzing all Chinese and U.S. newspapers as two separate documents in our corpus. The first noticeable difference between the two countries' newspapers are the overall sentiment. Chinese newspapers are largely more negative, which is expected as Chinese newspapers were hypothesized to focus more on the chaos created from

the protests. Chinese newspapers contained a mean score of -0.51 and U.S. newspapers contained a mean score of -0.28, on a scale of -1 to 1.




Note: Analysis run through the rule-based sentiment analysis engine VADER

Figure 1. Sentiment Scores

Table 1 displays the most frequently occurring words in each country's newspapers that we collected, and Table 2 displays the most frequently occurring words that do not appear in the other country's newspapers. Perhaps most notable is that "violence" (1,098), "violent" (945), "radical" (531), and "order" (511) are words commonly found in Chinese newspapers but are found not in U.S. newspapers. This may imply that the content of Chinese articles is more focused on the protestors and less on the politics surrounding the issue and fits with the sentiment scores in Figure 1. By comparison, U.S. newspapers commonly wrote "Beijing" (1,645), "mainland" (894), and "Trump" (707), which do not appear in Chinese newspapers. This provides evidence that the content of U.S. articles is focused on the politics surrounding the protests rather than coverage on the unfolding of daily events.

Moreover, Chinese newspapers use the term "HKSAR" (546), which references Hong Kong as a special administrative region—the legal term for Hong Kong's status in China. Words like "region" (526) and "special" (431) further emphasize Hong Kong's legal status. This is compared to U.S.'s use of capitals and each country's head of state, which emphasize a larger geopolitical background.

Finally, U.S. newspaper's use of "people" (1,720), "politics" (956), "party" (762), and "support" (702) may imply that the U.S. is not just interested in hard geopolitics, but also the soft politics of global perceptions.

| | Chinese Newspapers | | U.S. Newspapers | |
|---|---|---|---|---|
| | Word | Freq | Word | Freq |
| 0 | hong | 5721 | hong | 6067 |
| 1 | kong | 5674 | kong | 5825 |

| | Chinese Newspapers | U.S. Newspapers |
|---|---|---|
| | Unique Words | Unique Words |
| 0 | violenc | beij |
| 1 | violent | year |

| | | | | |
|---|---|---|---|---|
| 2 | said | 2859 | protest | 4575 |
| 3 | polic | 2015 | said | 3601 |
| 4 | protest | 1954 | china | 3525 |
| 5 | govern | 1340 | chines | 2245 |
| 6 | peopl | 1299 | polic | 2171 |
| 7 | china | 1289 | **peopl** | 1720 |
| 8 | **violenc** | 1098 | **beij** | 1645 |
| 9 | law | 1046 | govern | 1613 |
| 10 | **violent** | 945 | would | 1451 |
| 11 | offic | 784 | one | 1296 |
| 12 | also | 706 | citi | 1234 |
| 13 | citi | 703 | year | 1044 |
| 14 | public | 682 | **polit** | 956 |
| 15 | countri | 670 | demonstr | 955 |
| 16 | chines | 651 | **mainland** | 894 |
| 17 | one | 617 | time | 861 |
| 18 | **hksar** | 546 | use | 845 |
| 19 | two | 538 | also | 844 |
| 20 | act | 537 | lam | 830 |
| 21 | **radic** | 531 | compani | 817 |
| 22 | **region** | 526 | state | 809 |
| 23 | intern | 524 | leader | 794 |
| 24 | **order** | 511 | mani | 768 |
| 25 | administr | 510 | **parti** | 762 |
| 26 | affair | 486 | week | 756 |
| 27 | forc | 483 | offici | 756 |
| 28 | support | 476 | offic | 744 |
| 29 | us | 474 | say | 714 |
| 30 | demonstr | 463 | two | 712 |

| | | |
|---|---|---|
| 2 | public | **polit** |
| 3 | countri | **mainland** |
| 4 | **hksar** | time |
| 5 | act | compani |
| 6 | **radic** | state |
| 7 | **region** | leader |
| 8 | intern | mani |
| 9 | **order** | **parti** |
| 10 | administr | week |
| 11 | affair | offici |
| 12 | forc | say |
| 13 | us | **trump** |
| 14 | nation | like |
| 15 | **special** | call |
| 16 | right | includ |
| 17 | system | could |
| 18 | central | bill |
| 19 | airport | new |

| | | | | |
|---|---|---|---|---|
| 31 | nation | 441 | **trump** | 707 |
| 32 | **special** | 431 | **support** | 702 |
| 33 | lam | 425 | like | 698 |
| 34 | use | 424 | call | 697 |
| 35 | right | 418 | includ | 695 |
| 36 | system | 410 | could | 694 |
| 37 | central | 406 | law | 684 |
| 38 | airport | 405 | bill | 677 |
| 39 | would | 402 | new | 676 |

Note: Bolded, blue words indicate possibly meaningful words based on the authors' interpretation of the subject matter

Table 1. Most Frequent Words Table 2. Unique Words

Next, we conducted information extraction to investigate the people and organizations each countries' newspapers discuss. We found that Chinese newspapers rarely discuss individuals. In fact, the top mention garnered only 37 mentions. "Xi" and "Xi Jingping" were written a total of 10 times. U.S. President Trump was never mentioned, nor were notable protestors. Most names appear to be article authors and not individuals involved in the protest.

Meanwhile, US newspapers mention individuals frequently with "Lam" or "Carrie Lam" being mentioned a total of 785 times. Heads of state are mentioned frequently with "Xi" or "Xi Jinping" mentioned a total of 549 times and "Trump" mentioned 279 times. Notable protesters are also mentioned such as "Wong" or "Joshua Wong" mentioned a total of 141 times.

A look at organizational terms used shows that U.S. newspapers seemed to link the protests with other events. "NBA", "ESPN", "Rockets", "Disney", and "TikTok" all reference arguably unrelated organizations that had some type of banned content in China or US public controversy. For example, NBA Rockets's general manager expressed support for protestors and against the Hong Kong police, leading to several Chinese businesses are suspending ties with the team and NBA preseason games banned for broadcast in China. Moreover, lead actress of Disney's Mulan voiced support for the Hong Kong police, leading to public outcry in the U.S. and pressure on Disney to remove the actress from the upcoming film.

Finally, U.S. newspapers discussed a variety of countries, capitals, and regions, such as Taiwan, Washington, Britain, Japan, America, Russia, Australia, Europe, and Iran. Meanwhile, Chinese newspapers mentioned Hong Kong 5,194 times and the next country, the U.S., only 747 times.

From this, we gathered that Chinese newspapers seemed to focus on the protests as they were unfolding and rarely discussed parallel issues. Meanwhile, U.S. newspapers were clearly linking the stories to issues such as censorship, boycotts, and the larger issue of Hong Kong's status in China. Protesters like Joshua Wong are named and discussed. These people could continue this as a larger movement beyond current protests. This is speculation on our part but

discussing a variety of parallel and possibly off-topic issues suggests that U.S. newspapers view the protests as a larger, long-term trend.

Cosine similarity metrics seem to bear this out. We updated our corpus to convert each newspaper, and all articles from that newspaper as a document, and then compared the cosine similarity of the three Chinese newspapers to each other, the three U.S. newspapers to each other, and then all Chinese articles to U.S. articles. Chinese newspapers have a high cosine similarity with the three newspapers possessing a cosine similarity between 0.64-0.72. This suggests that they are fairly similar to each other, except for China Daily and Xinhua being somewhat different. U.S. newspapers contained less similarity, with a lower and tighter cosine similarity range of between 0.62-0.63. The difference might be due to the wider range of topics and issues U.S. newspapers seem interested in discussing. More importantly, Chinese and U.S. newspapers contain a substantively lower cosine similarity of 0.55, indicating that they are quite different. This difference will likely make it easier to classify between the two countries' newspapers.

From this exploratory phase we conclude classification between Chinese and U.S. newspaper articles on the Hong Kong protests should be possible as the difference between each country's articles appear quite vast. The three main points we found are:

1) There are many meaningful words that are unique between the two countries.
2) The focus areas are quite different from each other with Chinese newspapers seemly choosing to focus more on the narrower, day-to-day events unfolding, whereas U.S. newspapers seemly choosing to focus more on the broader, possibly-related global issues and long-term trends associated with the protests.
3) Chinese newspapers are fairly homogenous with each other and heterogenous with U.S. newspapers on the topic.

## 3 NLP Techniques

We turn our attention to how we can model a corpus in which each document is a news article and the outcome variable is an article's country of origin. We will use a bag of words (BOW) methodology, which is foundational method for text analysis. Under BOW, pre-processing steps usually include term case transformation (e.g., setting all terms to lowercase), stemming (cutting off the end of terms, e.g., removing "ing" from "jumping" to make "jump"), spelling correction, and "stopword" removal (removing terms that occur infrequently or are not relevant to the analysis such as articles and prepositions) (Satapathy et al., 2017). Next, the unstructured text is transformed into a structured form known as a "document term matrix," in which documents are treated as rows (i.e., observations) and the items of text are treated as columns (i.e., features).

Although stemming and stopword removal reduces the total amount of terms, sparsity remains a common problem. Often, a document term matrix's columns can rival the length of its rows (Çano and Maurizio, 2019). Setting a minimum threshold of documents for term to appear alleviates this issue. It assumes that terms used infrequently across documents may not provide meaningful variation for modeling. However, this is a balancing act as a low minimum sparsity threshold may help modelling efforts by enabling more context (i.e., controlling for commonly used terms).

Removal of terms is not the only method for improving text analysis. Determining the relevance of terms in a document term matrix can be done using term weights. Terms can be weighted as

binary occurrences indicating presence (1) or absence (0) of a term within a document, integer counts for the number of occurrences in the document (i.e., term frequencies), or term frequency-inverse document frequency values (TF-IDF) which determines relevance by the frequency of a term within a document multiplied by the rarity of the term across documents (Silge & Robinson, 2018).

Finally, some word context can be gleaned from combining words into n-grams (Brownlee, 2017) —such as combining the unigrams (“I”, “am”, “well”) into a bigram (”I am”, “am well”) or a trigram (“I am well”). For example, a BOW implementation that uses bigrams pairs “not” and “awful” together as the bigram “not awful”, would better identify underlying sentiment than a method that would treat “not” and “awful” as separate words. However, n-grams are limited in their benefit as they do not identify semantic relationships between words and can add sparsity to the document term matrix.

## 4 Modeling

Using this BOW framework, we ran a total of 27 models, in which we varied three parameters: the minimum sparsity threshold, the term weights, and the n-grams. We attempted three variations of each of the parameters. Each model was run using a gradient-boosted decision tree with a logistic classifier and a log-likelihood cost function. Table 3 displays each model configuration.

| Parameters | Variations | | |
|---|---|---|---|
| **Minimum sparsity** | 1% threshold | 5% threshold | 10% threshold |
| **Term weights** | Binary occurrence | TF | TF-IDF |
| **N-grams** | Unigram | Bigram | Trigram |

Table 3. Model Configurations

These models were preprocessed using conventional methods such as term case transformation, stemming, stopword removal, and setting a maximum sparsity threshold of 90%—meaning that words that appeared in 90% or more of documents were removed under the same logic as stopword removal.

We evaluated our models based on their F1-scores. Using this evaluation metric, we found that classification proved easy for our bag of words technique. A 1% minimum sparsity threshold proved better than a 5% or 10%, but only by a little. Moreover, term weights and n-grams made little difference. Figure 2 shows the nine models run with a 1% minimum sparsity threshold—all containing similar F1-scores ranging from 0.955 to 0.979.

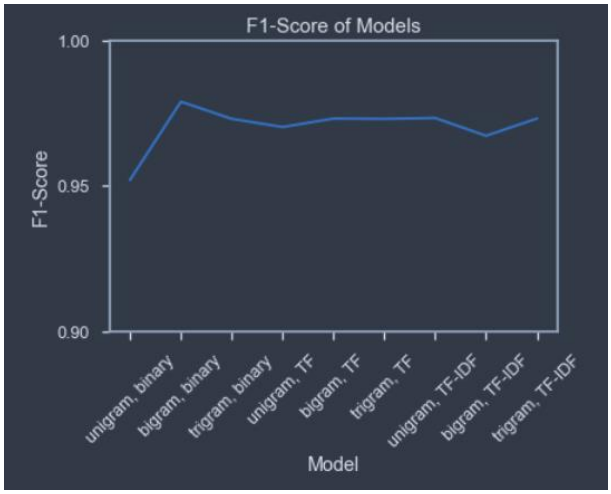


Figure 2. F1-Scores of Models with 1% Minimum Sparsity

Our best model configuration included a bigram with binary term weights (i.e., one-hot-encoding) and a 1% minimum sparsity threshold. It contained an F1-score of 0.979. On our test set, comprise of 331 news articles (30% of our corpus), we had only 7 misclassifications—3 false positives and 4 false negatives.

## 5 Conclusion

Chinese newspapers focused attention to the violence of the protesters and the Hong Kong government's response. In doing so, they tended to talk about narrow, day-to-day events and did not link the protest story to wider global politics. By contrast, U.S. newspapers paid more attention to wider issues and democracy. Because they linked the protest story to wider US-China relations and global politics, their news contained more variance in content.

Overall, the similarity between Chinese newspapers is high, which likely improved our ability to distinguish it from U.S. newspapers. As a result, the model configurations did not vary in their effectiveness. We find that this corpus cannot meaningfully distinguish between BOW parameters. This is most likely due to the amount substantively important terms that are unique between the articles' country of origin.

A BOW methodology does not appear to need to strike a balance between parameters when key words are distinct within the classification. The minimum sparsity threshold likely did not matter as the substantive terms were not common across the countries of origin and therefore not common enough for thresholds between 1% and 10%. Term weights likely did not matter as distinctions at the document-level did not appear important. Instead, the distinctions between country of origin were strong, overriding document-level term relevance. Finally, n-grams did not appear important. This is likely due to each country discussing the Hong Kong protests under vastly different topics. Contextual information is important when the meaning is difficult to grasp from individual words. Because the discussion was vastly different, causing many of the key terms to be unique between country of origin, this contextual information was unnecessary.

Overall, our research shows that a BOW methodology is more than sufficient on corpora where the subject matter takes vast different topics leading to many unique words between classification outcomes.

## Bibliography


Brownlee, J. (2017). A gentle introduction to the bag-of-words model. Retrieved from https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Çano, E., and Maurizio, M. (2019.) Word embeddings for sentiment analysis: A comprehensive empirical survey. ArXiv abs/1902.00753.

Satapathy, R., Guerreiro, C., Chaturvedi, I., and Cambria, E. (2017). Phonetic-based microtext normalization for twitter sentiment analysis. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE: 407-413.

Silge, J., and Robinson, D. (2018). Analyzing word and document frequency: TF-IDF. In Text Mining with R: A Tidy Approach. Retrieved from https://www.tidytextmining.com/tfidf.html