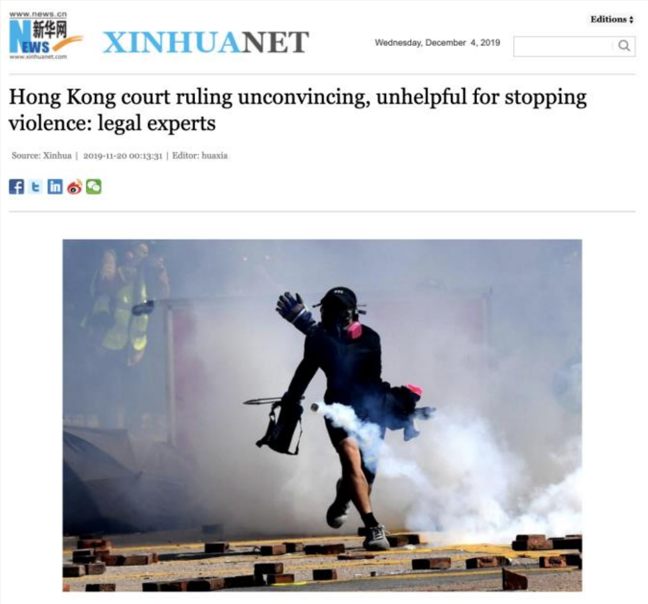
AROUND THE WORLD COVERAGE

Classifying the Differences in News Coverage of the Hong Kong Protests

Michael Siebel
Junchi Tian
Bixuan Huang

Background

- 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.

- Can we build a model that will understand these differences?

Overview

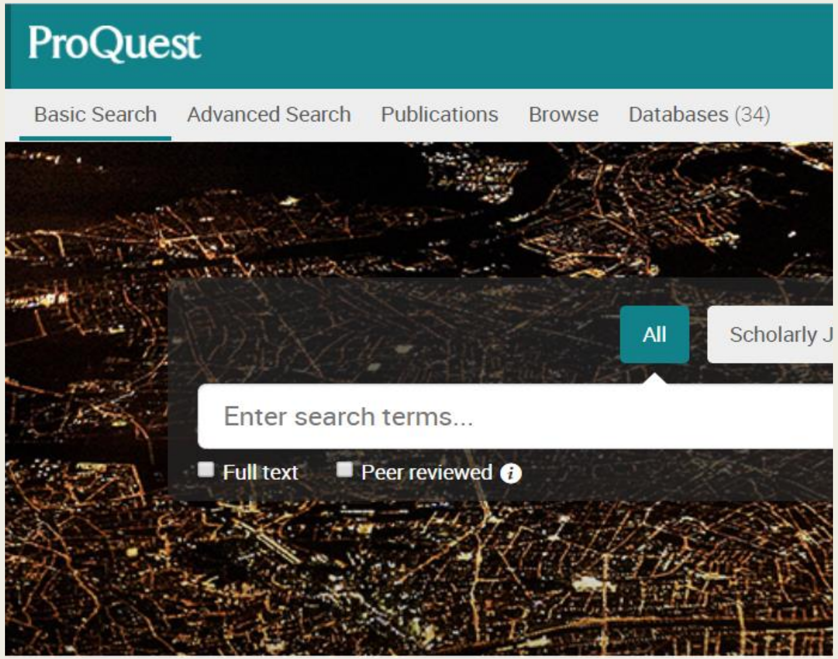
- Figuring out different journalistic tones between U.S. newspapers and Chinese newspapers could help us get better understanding of U.S. and Chinese commend view points about Hong Kong Protest.

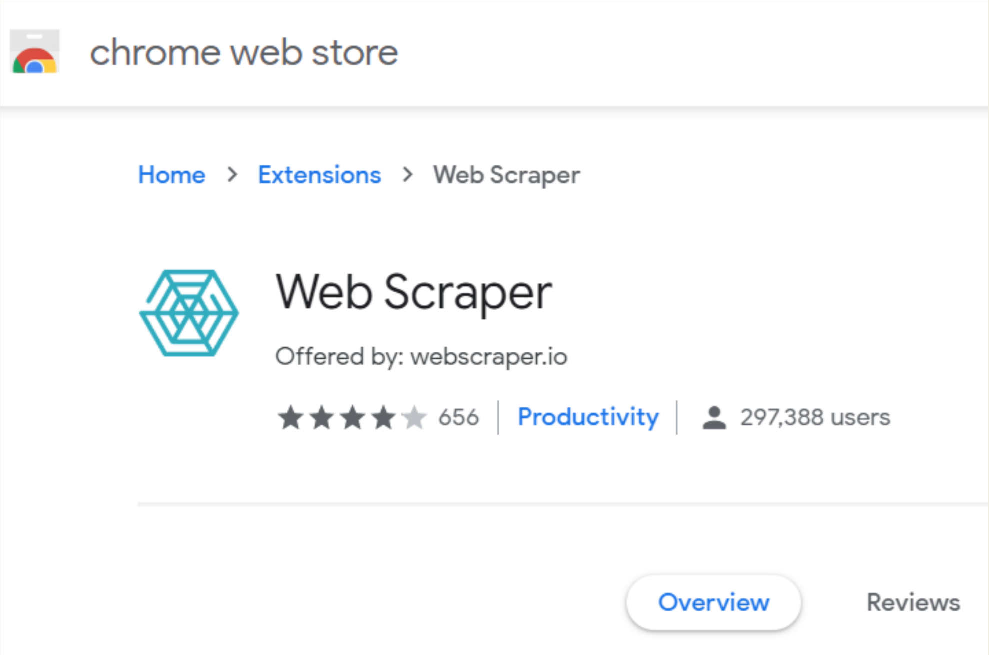
- Our goal is to determine what natural language processing (NLP) techniques can best distinguish between U.S. and Chinese reporting on the 2019 Hong Kong protests.

- Process:
  - *Pre-processing*
  - *Exploratory Data Analysis*
  - *Modelling*
  - *Assessment*

Corpus

- 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.
- We use web Scraper to scrape corpus from ProQuest.





Data Preprocessing Steps

- Tokenized words
- Removed punctuation and numbers
- Set words to lowercase
- Stemmed words
- Removed stop words
- Split into training and testing
- Prepared for a Bag of Words (BoW) modeling

China

U.S.

WORD CLOUD

Word Frequency

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

	Chinese Newspapers		U.S. Newspapers	
	Word	Freq	Word	Freq
10	violent	945	would	1451
11	offic	784	one	1296
12	citi	703	citi	1234
13	public	682	year	1044
14	countri	670	polit	956
15	chines	651	demonstr	955
16	one	617	mainland	894
17	hksar	546	time	861
18	Two	538	use	845
19	Act	537	lam	830

TF-IDF

	Chinese Newspapers		U.S. Newspapers	
	Word	Weight	Word	Weight
0	said	0.087133	said	0.075097
1	polic	0.063606	police	0.064404
2	china	0.049195	protesters	0.057372
3	govern	0.040110	chinese	0.050596
4	law	0.040074	mr	0.046603
5	ha	0.039674	beijing	0.046157
6	violent	0.035292	people	0.039267
7	wa	0.032080	government	0.037620
8	violenc	0.031111	city	0.035355
9	peopl	0.031015	lam	0.032596

	Chinese Newspapers		U.S. Newspapers	
	Word	Weight	Word	Weight
10	hksar	0.028894	trump	0.028166
11	public	0.028253	mainland	0.025260
12	offic	0.028036	xi	0.024914
13	airport	0.026426	party	0.023303
14	chines	0.026026	democracy	0.023188
15	lam	0.025312	trade	0.022204
16	intern	0.025171	law	0.021734
17	sar	0.023906	political	0.020953
18	order	0.023900	year	0.020823
19	act	0.023037	pro	0.019826

Information Extraction - Persons

China

```
[('Lam', 271),
 ('Chan', 111),
 ('Lee', 95),
 ('Wong', 72),
 ('Carrie Lam', 60),
 ('Yang', 59),
 ('Carrie Lam Cheng Yuet-ngor', 52),
 ('Lau', 48),
 ('Wan Chai', 47),
 ('Hua', 45),
 ('Tse', 40),
 ('Lai', 39),
 ('Geng', 34),
 ('Ho', 33),
 ('Yuen Long', 30),
 ('Zhang', 27),
 ('Xinhua', 27),
 ('Albertson', 27),
 ('Chow', 25),
 ('Tse Chun-chung', 25)]
```

U.S.

```
[('Lam', 576),
 ('Xi', 375),
 ('Trump', 279),
 ('Carrie Lam', 209),
 ('Xi Jinping', 174),
 ('Hong Kongers', 157),
 ('Crédito', 113),
 ('Facebook', 111),
 ('Morey', 107),
 ('Wong', 92),
 ('Li', 86),
 ('Mao', 83),
 ('Chan', 61),
 ('Lai', 60),
 ('Ho', 58),
 ('Natasha Khan', 54),
 ('Joshua Wong', 49),
 ('Daryl Morey', 45),
 ('Leung', 40),
 ('Yuen Long', 37)]
```

Information Extraction - Locations

China

```
[('Hong Kong', 4045),
 ('China', 1151),
 ("Hong Kong's", 586),
 ('US', 398),
 ('U.S.', 218),
 ('HONG KONG', 206),
 ('the Hong Kong Special Administrative Region', 195),
 ('Taiwan', 153),
 ('the United States', 131),
 ('Beijing', 115),
 ('Hong Kong Special Administrative Region', 79),
 ('Sept.', 69),
 ('Britain', 67),
 ('Macao', 66),
 ('BEIJING', 54),
 ('Shenzhen', 47),
 ('Washington', 46),
 ('Hong Kong Island', 46),
 ('Shanghai', 44),
 ('Twitter', 36)]
```

U.S.

```
[('Hong Kong', 4265),
 ('China', 3406),
 ('Beijing', 1722),
 ("Hong Kong's", 747),
 ('U.S.', 645),
 ('the United States', 346),
 ('Taiwan', 332),
 ('Hong Kong’s', 326),
 ('HONG KONG', 159),
 ('Washington', 138),
 ('Britain', 134),
 ('Japan', 126),
 ('Shenzhen', 104),
 ('America', 84),
 ('Twitter', 82),
 ('Asia', 80),
 ('Russia', 79),
 ('Shanghai', 76),
 ('Australia', 71),
 ('Europe', 65)]
```

Information Extraction - Organizations

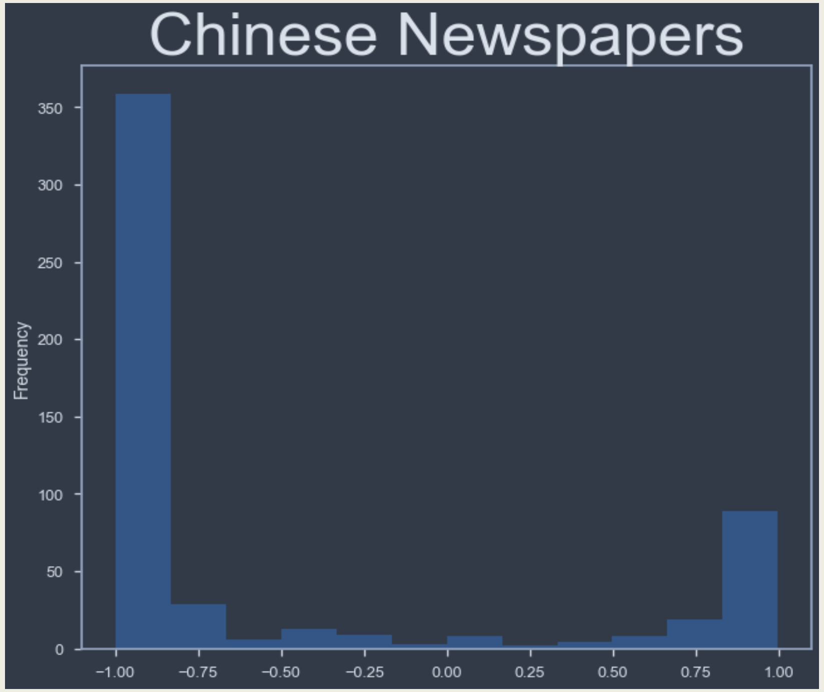
China

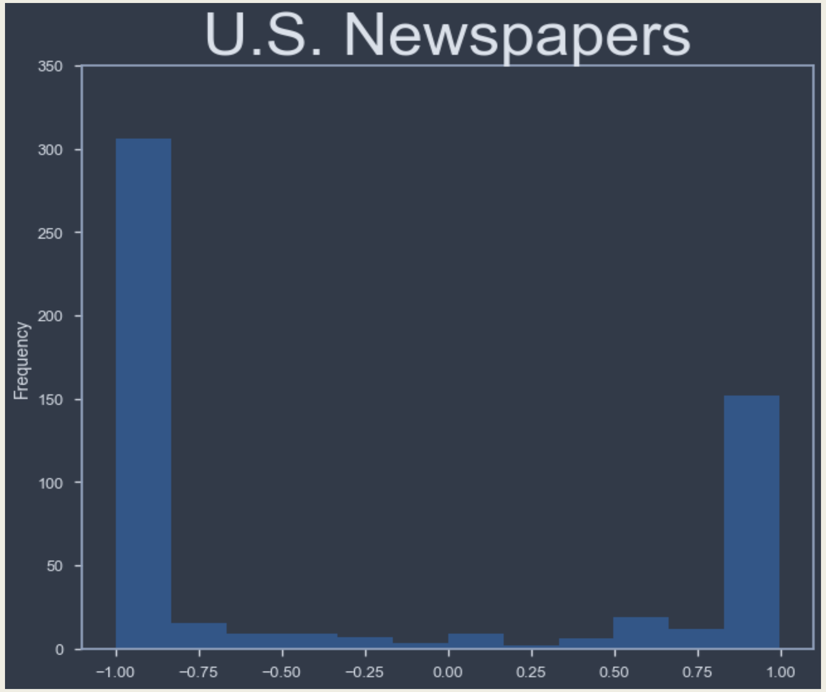
```
[('Xinhua', 297),
 ('SAR', 240),
 ('MTR', 77),
 ('China Daily', 60),
 ('CNN', 58),
 ('Macao Affairs Office', 52),
 ('the State Council', 49),
 ("the Liaison Office of the Central People's Government", 38),
 ('Foreign Ministry', 36),
 ('Legislative Council', 33),
 ('the Legislative Council', 33),
 ('the Hong Kong International Airport', 31),
 ('NBA', 28),
 ('EU', 28),
 ('the Global Times', 25),
 ('the Hong Kong Police Force', 23),
 ('Hong Kong Human Rights and Democracy Act', 20),
 ('YouTube', 20),
 ('LIHKG', 20),
 ('Telegram', 20)]
```

U.S.

```
[('Trump', 380),
 ('NBA', 248),
 ('Communist Party', 131),
 ('TikTok', 123),
 ('the Communist Party', 111),
 ('Cathay', 101),
 ('’s', 79),
 ('Rockets', 76),
 ('N.B.A.', 74),
 ('Telegram', 63),
 ('the Chinese Communist Party', 62),
 ('Congress', 60),
 ('Tiananmen', 56),
 ('Apple', 53),
 ('Times', 43),
 ('The Washington Post', 42),
 ('Cathay Pacific', 41),
 ('Legislative Council', 40),
 ('Disney', 39),
 ('the Chinese University of Hong Kong', 37)]
```

Sentiment analysis
(by rule-based sentiment analysis engine VADER)





- Chinese newspapers are largely more negative, which is expected as we expect Chinese newspapers 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.

Cosine Similarity

- Chinese newspapers: 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: 0.62-0.63.
  - *The difference might be due to the wider range of topics and issues U.S. newspapers seem interested in discussing.*

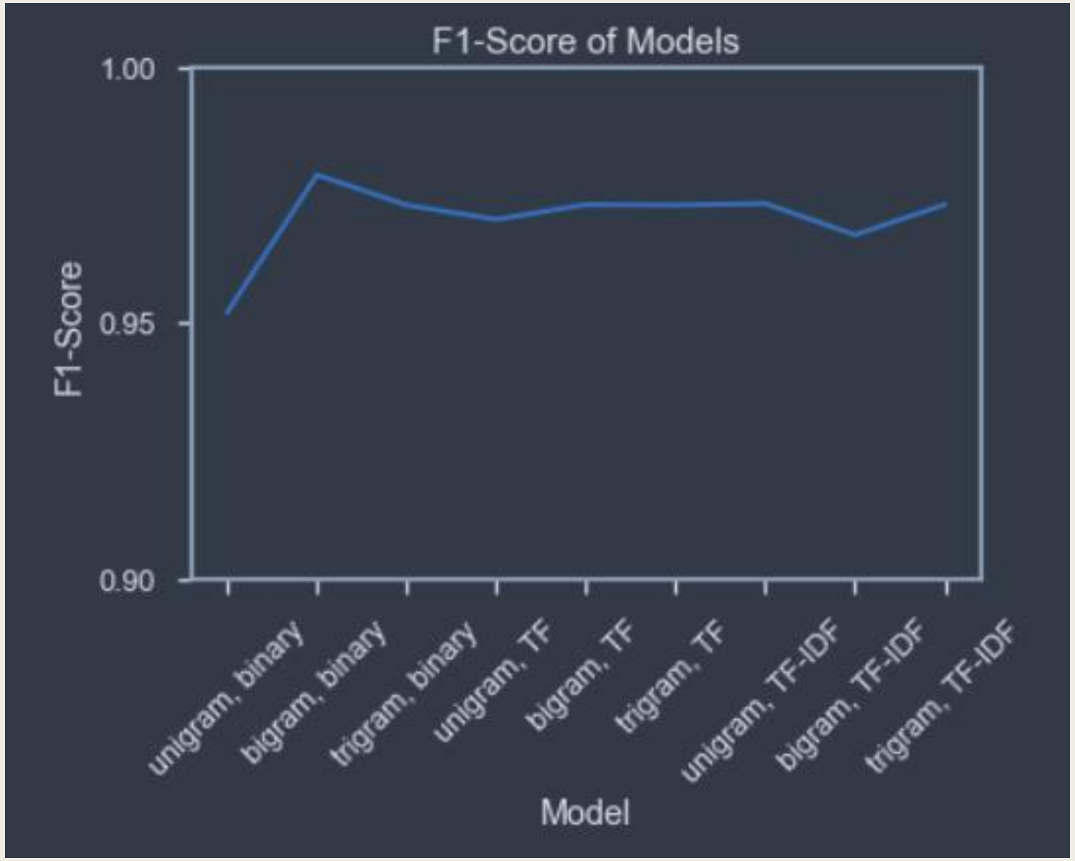
- More importantly,
  - *Chinese and U.S. newspapers contain a substantively lower cosine similarity of 0.55, indicating that they are quite different.*

Modeling

- We ran a total of 27 models, in which we varied three parameters: the minimum sparsity threshold, the term weights, and the n-grams

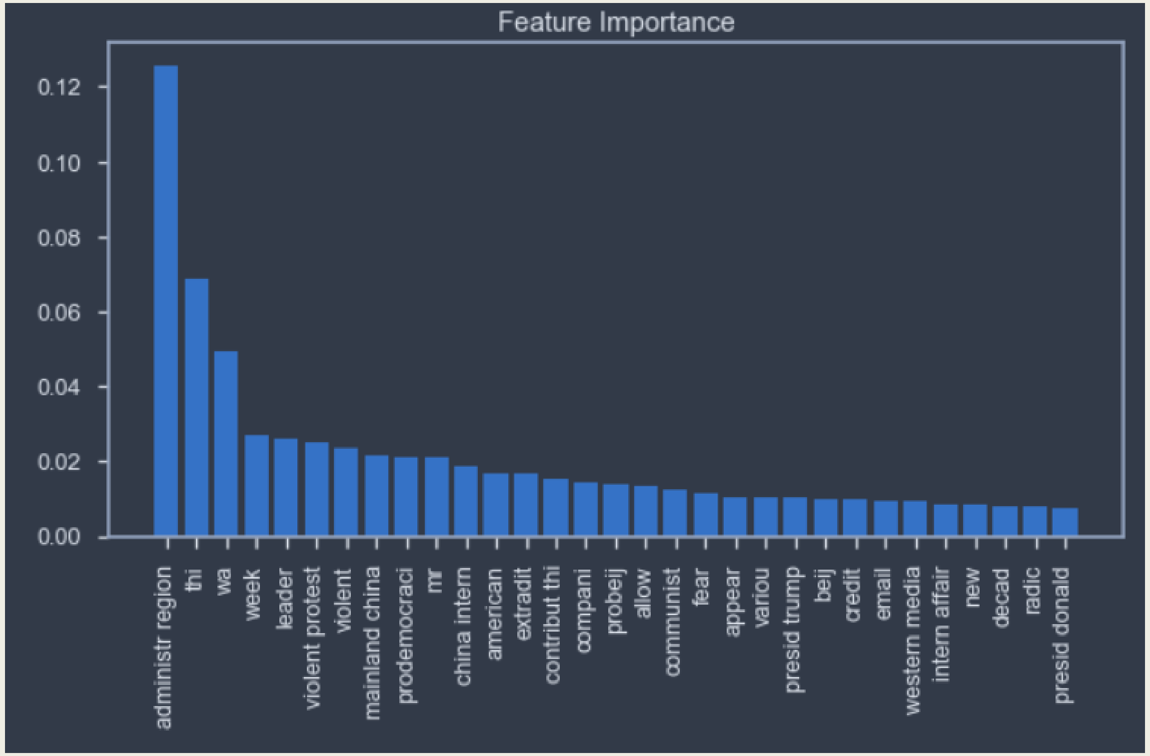
Parameters	Variations		
Minimum sparsity	1% threshold	5% threshold	10% threshold
Term weights	Binary occurrence	TF	TF-IDF
N-grams	Unigram	Bigram	Trigram

Modeling



- We evaluated our models based on their F1-scores.
- A 1% minimum sparsity threshold proved better than a 5% or 10%, but only by a little.
- With a 1% minimum sparsity threshold—all containing similar F1-scores ranging from 0.955 to 0.979.

Modeling



- Our best model configuration included an F1-score of 0.979
- On our test set, comprise of 331 news articles (30% of our corpus)
  - *7 misclassifications*
  - *3 false positives*
  - *4 false negatives.*
- Notable words in the model include: administrative region, violent protest, violent, China internal, internal affair, and radical

Conclusion

- Chinese newspapers
  - *pay more attention to HK government and violence*
  - *tend to talk about narrow specific events and did not link the protest story to wider global politics*
- US newspapers
  - *pay more attention to wider issues and democracy*
  - *tend to use the head of country to represent each country*
  - *contained more variety because they linked the protest story to wider US-China relations and global politics.*
- The similarity between Chinese newspapers is high which likely improves our ability to distinguish it from U.S. newspapers
- As a result:
  - *The model configurations did not make much difference*

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

THANK YOU