Characterizing News Portrayal of Civil Unrest in Hong Kong, 1998–2020

James A. Scharf and **Arya D. McCarthy** and **Giovanna Maria Dora Dore**
Johns Hopkins University
jscharf8@jhu.edu

Abstract

We apply statistical techniques from natural language processing to a collection of Western and Hong Kong–based English-language newspaper articles spanning the years 1998–2020, studying the difference and evolution of its portrayal. We observe that both content and attitudes differ between Western and Hong Kong–based sources. ANOVA on keyword frequencies reveals that Hong Kong–based papers discuss protests and democracy less often. Topic modeling detects salient aspects of protests and shows that Hong Kong–based papers made fewer references to police violence during the Anti-Extradition Law Amendment Bill Movement. Diachronic shifts in word embedding neighborhoods reveal a shift in the characterization of salient keywords once the Movement emerged. Together, these raise questions about the existence of anodyne reporting from Hong Kong–based media. Likewise, they illustrate the importance of sample selection for protest event analysis.

1 Introduction

In this era where movements against entrenched power structures are both widespread and well documented, we can conduct computational analyses of language to guide, support, and challenge hypotheses about unrest and its discussion in mainstream written media sources. We direct these tools to analyze portrayals of protest and unrest in Hong Kong over a period of 22 years.

Public protests in Hong Kong date back to British colonial rule and have evolved from the bloody riots of the 1960s to the protests of 2019–2020, when up to two million people took to the streets over an extradition bill. They feared it would make the Hong Kong inhabitants subject to China's legal system in violation of the Basic Law[1], which guarantees that Hong Kong's capitalist system, judicial independence, and existing civil and political liberties would remain unchanged until 2047. Hong Kong protests captured the world's attention with defiant crowds commemorating the 1989 Tiananmen Square incidents, the July 1, 1997 transfer of sovereignty from the UK to China, and students blockading roads in the Admiralty district while doing their homework during the pro-democracy Umbrella Movement in 2014 (Weiss and Aspinall, 2012). Over time, the instability created by the protests has become a threat to the credibility of Hong Kong as a financial hub and the possibility of applying the principles of *one country, two systems* beyond Hong Kong and Macau (Overholt, 2021).

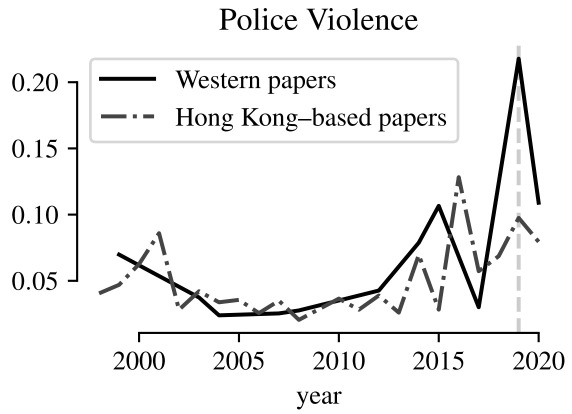

Figure 1: Words related to the topic of police violence in Hong Kong sharply rose in prominence in 2019, but only in Western news sources. The corresponding increase in Hong Kong–based news was muted.

We apply a host of techniques from natural language processing to mark inconsistencies in event characterizations, analyzing news articles related to episodes of civil unrest between 1998 and 2020, in both western- and Hong Kong–based English-language newspapers. In the volatile context of

[1] https://www.basiclaw.gov.hk/en/basiclaw

Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) , pages 43–52
August 5–6, 2021, ©2021 Association for Computational Linguistics

Hong Kong politics, newspapers' tendency to report more dramatic than ordinary events may encourage reporting bias that either emphasizes or undermines the legitimacy of the protests or the legitimacy of the regime against which the protests are directed (Snyder and Kelly, 1977; Earl et al., 2004; Schrodt et al., 2001).

Our contributions are manifold. Foremost, our work is novel amongst work on protests and natural language due to the expanse of our time horizon. Second, we characterize crucial differences in Western- and Hong Kong–based portrayals of protest: statistically significant differences in protest-related lexical choice (§5.1), reinforced by differences in treatment of democracy and police violence (§5.2), though with no major differences in sentiment (§5.4). Third, we find several key points where coverage differs (§5.3), including a major shift in the notion of "confrontation".

2 Related Work

Content analysis (Berelson, 1952), in general, is a set of non-invasive techniques for studying communication artifacts such as documents, photographs, and recordings. Computational methods have supercharged content analysis by complementing subject matter expertise with the potential for massive scale. Lucy et al. (2020) consider the content of United States history textbooks in Texas, using word embedding similarity, topic models, and dependency parsing to generate clues toward differing portrayals of race and gender. Field et al. (2018) relate the content of Russian state-run news articles to the nation's economic performance, finding an agenda of distraction through the framework of Granger causality (Granger, 1988). Other attempts at content analysis and stylometry consider authorship (Mosteller and Wallace, 1984; Bergsma et al., 2012), native language identification (Koppel et al., 2005; Bergsma et al., 2012), and deceptive communication in reviews (Ott et al., 2013).

With the advent of fast-paced 'social' media, recent work (De Silva and Riloff, 2014; Alsaedi et al., 2017; Sech et al., 2020) has aimed to characterize unrest through Tweets, short communiques on the platform Twitter.

Within the specific focus of protests, the closest work to ours in longitudinal scope is Papanikolaou and Papageorgiou (2020), whose 541 thousand news articles (albeit not all about protest) reflect Greece from 1996 to 2014; other similarly broad-scale work is rare. Wueest et al. (2013) apply topic models and named entity recognition to protest event analysis. The CLEF 2019 Protest-News shared task asked participants to perform event extraction, even in news articles about a country outside of the training set. The organizers report consistent drops in performance after this shift. Inverting this, our work calls into question different views on protest in the same location.

3 Data

We collected a corpus of news articles collected from six Western-based English language newspapers: *The New York Times*, *The Wall Street Journal*, *The Washington Post*, *The Financial Times*, *The Guardian*, and *The Times*; and two Hong Kong–based English language newspapers: *The China Daily* and *The South China Morning Post*, covering multiple incidents of protests that took place between January 1998 and June 2020. The newspapers were purposefully selected because they are English-language newspapers; the selection ensures newspaper diversity within western- and Hong Kong–based newspapers to allow for insights into differences across cultures.

The articles were collected through keyword-based searches in ProQuest Newspapers for the western English language newspapers, and Newsbank Access World News Research Collection for the English language Hong Kong newspapers. We searched for the keywords "Hong Kong" + "protests", "Hong Kong" + "rallies", "Hong Kong" + "marches", and "Hong Kong" + "riots". We used the East Coast editions for *The New York Times* and *The Wall Street Journal*; the UK editions for *Financial Times*, *The Guardian*, and *The Times*, and the overseas edition for *China Daily* (which is run and printed in Hong Kong). To be eligible for collection, articles had to be at least 300 words long.

We manually screened the collected articles to eliminate irrelevant items such as duplicates within each publication, readers' letters, and articles that included any of the research chosen keywords but whose content was not about the protest incidents.

Following the manual screening, we retained 4175 articles, with a mean length of 782 tokens. The *South China Morning Post* and *The New York Times* published the largest number of articles about protests in Hong Kong between 1998 and 2020. The *South China Morning Post* published the most articles on Hong Kong protests among all newspapers, and *The New York Times* published the most among western-based newspapers.

4 Method

We aim to contrast the treatment of civil unrest in Hong Kong, both across news sources and over time. Here we outline four techniques to suit this purpose: analysis of word choice with ANOVA, analysis of word clusters with latent Dirichlet allocation, analysis of word usage with embedded neighborhood shifts, and analysis above the word level with sentiment analysis.

4.1 Comparing lexical frequency

Word frequency exposes obvious discrepancies in word choice and word usage. A lack of event-related keywords in contemporaneous articles from different newspapers may signal the omission of events in some of them.

Each source will have some degree of variation in keyword counts. An author's voice accounts for some mismatch in frequency, but not all. It is therefore challenging to determine whether the distribution of keyword counts is due to pure chance or something more meaningful. Analysis of variance (ANOVA) is a sampling theory–based method for comparing the means of a quantitative response variable, when the explanatory variable is categorical (Agresti, 2017). A statistically significant p-value supports that the means of both populations are different. According to Agresti (2017), ANOVA is analogous to regression with a continuous response variable and a categorical explanatory variable.

We apply ANOVA to our corpus to determine important differences in frequencies. We first select 19 keywords of interest related to Hong Kong protests.[2] Then, for one keyword at a time, we perform the following steps.

1. Split the corpus in two by some categorical attribute,
2. Obtain the keyword's frequency in each article of both corpora, and then
3. Apply ANOVA to establish whether our categorical variable is associated with a variation in frequency.

In this work, we use the location of the article's publisher as the categorical variable.

This statistical analysis cannot, however, reveal the *motive* for a difference in lexical choice. It merely raises the question to subject matter experts. It then befalls those experts to determine whether the difference arises due to intentional omission, niceties of a newspaper's style guide, or some other feature.

ANOVA uses the F-test to check equality of the word frequencies in each group. We set a significance level of $\alpha = 0.05$ and employ the Bonferroni correction (Dunn, 1961).

We also attempted to identify discrepancies between the words used by different subsets of articles using a weighted log-odds ratio (Monroe et al., 2017) with an informative Dirichlet prior (following Jurafsky et al., 2014; Field et al., 2018; Lucy et al., 2020), to mixed results. We omit this from later discussion.

4.2 Topic modeling

Topic modeling characterizes documents by the topics they contain, automatically identifying the topics from corpora. We use latent Dirichlet allocation (LDA; Blei et al., 2003) for our topic models. It is a probabilistic generative model that maintains distributions over the words within each topic and the topics with each article, representing each article in the traditional vector space model (Salton et al., 1975). With LDA, we capture and convey the prevalence of various topics, so that we can contrast these across news sources and over time.

We perform topic modeling with MALLET (McCallum, 2002). To preprocess the articles, we lemmatize all tokens with WordNet's `morphy` feature (Miller, 1995). We also extract common bigrams. The resulting unigrams and bigrams were converted to term–document matrices and provided as inputs to MALLET. We created models, setting the number of topics from $k = 10$ to 60, and evaluated the coherence of the resultant topics according to Mimno et al. (2011). We found that using 13 topics produced the highest coherence score. We then identified each of these topics with an identifying label (see Table 2).

Our topic model represents each article as a mixture of topics. More prevalent topics have higher mixture weight, and the weights sum to 1 for each article. (In LDA, these can be interpreted as sam-

[2] *confront*, *confrontation*, *crackdown*, *democracy*, *freedom*, *freedom_of_speech*, *independence*, *occupation*, *protest*, *protests*, *resistance*, *rights*, *riot*, *rule_of_law*, *severe*, *tension*, *terrorism*, *terrorist*, *unrest*.

ples from a k-dimensional Dirichlet distribution.) We can estimate a topic's prevalence in a news source or year by averaging the topic's weight across the articles from that source or year.

4.3 Comparing lexical usage

Complementary to the previous methods which consider *which* words are used, we would like to investigate the evolution of *how* words are used differently, both in the Western/non-Western split and over time.

Diachronic shifts in word usage are often identified with changes in words' neighborhoods in an embedding space (Hamilton et al., 2016; Gonen et al., 2020). For instance, Hamilton et al. (2016) used these to find a shift in the word "broadcast" from agricultural to television contexts between the 1850s and 1900s. A word embedding model seeks to assign similar vectors (measured by dot product) to words in similar contexts, and different vectors to words in different contexts. If the usage of a word changes, then this should be reflected in changes to the word's context and consequent changes in the word's embedding.

We re-implement and extend the difference-in-usage model of Gonen et al. (2020), which measures how the contexts of words differ.

1. Partition the corpus $\mathcal{C}$ into $\mathcal{C}_a$ and $\mathcal{C}_{\overline{a}}$ based on the attribute of interest a.

2. Fit separate word embedding models for each partition: $\mathcal{M}_a$ and $\mathcal{M}_{\overline{a}}$.

3. Select a keyword w of interest.

4. Obtain the set of nearest neighbors $\mathrm{NN}_a(w)$ and $\mathrm{NN}_{\overline{a}}(w)$ of w according to each of $\mathcal{M}_a$ and $\mathcal{M}_{\overline{a}}$.[3]

5. Score the usage-change of w as the size of the intersection, $|\,\mathrm{NN}_a(w) \cap \mathrm{NN}_{\overline{a}}(w)|$.

After this process, if w is used differently based on the presence or absence of the attribute, we expect its score to be quite small. Words whose usage does not depend on the attribute will have similar neighborhoods in each split.

To extend the work of Gonen et al. (2020), we contextualize the similarity score of a given word against a reference set. Considering all words that occur at least 100 times, in which percentile does w's similarity score fall? We find this to be more meaningful than the raw similarity score.

We focus on two splits and apply the same methods of analysis to each split. For the first split, we divide the corpus by the location of the source. For the second split, we investigate whether June and July, high points in the 2019–2020 protests, mark any shifts in media coverage. For each, we calculate the scores of words that appear at least 100 times in both sub-corpora. Then, we use those scores to calculate the percentile of a given keyword's score. This makes it clearer to compare these relative scores.

4.4 Sentiment analysis

Sentiment analysis measures the attitude of an author from the tone and connotations of their document. While it may be performed based on handcrafted sentiment (valency) lexica (Mohammad, 2018), we select a technique that is robust to the specific words that are chosen. We select a BERT-based model to classify a given sentence as positive or negative because of its near state-of-the-art sentiment classification abilities.

We treat sentiment as a binary attribute[4] $(+, -)$ and use a probabilistic classifier trained on the Stanford Sentiment Treebank (SST-2; Socher et al., 2013). The model uses DistilBERT (Sanh et al., 2019) for feature extraction from text; DistilBERT has previously been used for sentiment analysis of product reviews (Büyüköz et al., 2020). We split each article into sentences, then classify each sentence. An article's sentiment is taken as the average sentiment over all of its sentences.

While this sentiment score obscures the reason for the author's attitude (*Were they opposed to the protests, or opposed to the police response?*), it still provides coarse-grained evidence of stylometric differences between news sources.

5 Results and Discussion

In this section, we analyze and give historical context for the results of the four techniques we describe in §4.

5.1 Comparing lexical frequency

The ANOVA results in Table 1 show that 15 of our 19 selected keywords have statistically significant differences in frequency. The top five keywords

[3] Following the recommendation of Wendlandt et al. (2018) and Gonen et al. (2020), we use 1000 nearest neighbors.

[4] There is merit to including a third 'it's complicated' class (Kenyon-Dean et al., 2018).

Keyword	p-value	F-statistic
democracy	**7.4e-103**	490.5
protest	**5.3e-76**	354.6
protests	**4.2e-65**	300.6
freedom	**3.2e-31**	137.2
occupation	**1.9e-27**	119.4
crackdown	**5.8e-17**	70.6
confrontation	**1.4e-15**	64.1
tension	**1.5e-15**	64.0
resistance	**3.8e-12**	48.4
confront	**3.4e-08**	30.5
riot	**1.9e-07**	27.2
unrest	**7.3e-06**	20.1
rights	**2.6e-05**	17.6
freedom_of_speech	**6.8e-04**	11.5
independence	**7.2e-04**	11.4
severe	1.3e-02	6.1
rule_of_law	2.3e-01	1.4
terrorist	4.9e-01	0.4
terrorism	5.2e-01	0.4

Table 1: ANOVA of 19 selected keywords' frequency between Western-based and Hong Kong–based articles. Kewords are sorted by F-statistic; significant differences after Bonferroni correction are **bolded**.

with the highest F-statistics, in descending order, are "democracy", "protest", "protests", "freedom", and "occupation".

We find consistently less discussion of protests in Hong Kong–based sources. The high F-statistic of "protest" and "protests" implies a statistically significant disparity in the coverage of protests. Figure 2 shows how the median number of times "protest" is lower in Hong Kong–based media sources than Western-based sources; corresponding plots illustrate the difference for "protests" (Figure 3) and "democracy" (Figure 4).

In conjunction with §5.2's findings of the prevalence of the "democracy movements" topic, the high F-statistics of "democracy" and "freedom" suggest that discourse about democracy is much more common in Western-based sources than in Hong Kong–based sources.

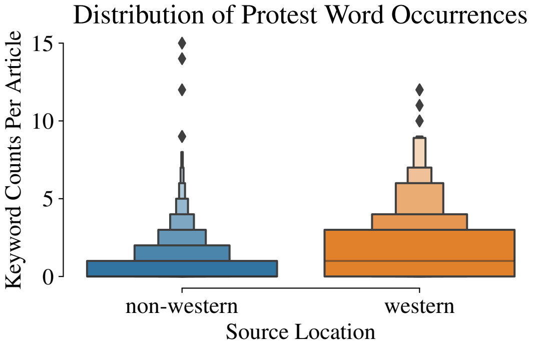

Figure 2: Quantile plot of "Protest" Counts Per Source Location

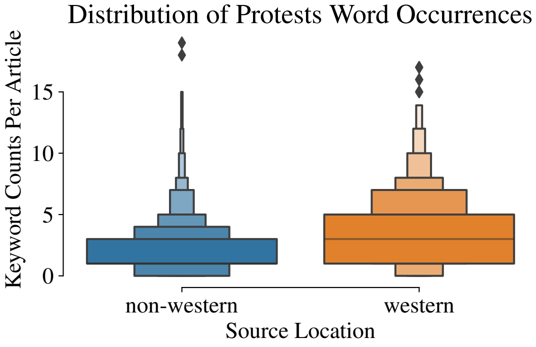

Figure 3: Quantile plot of "Protests" Counts Per Source Location

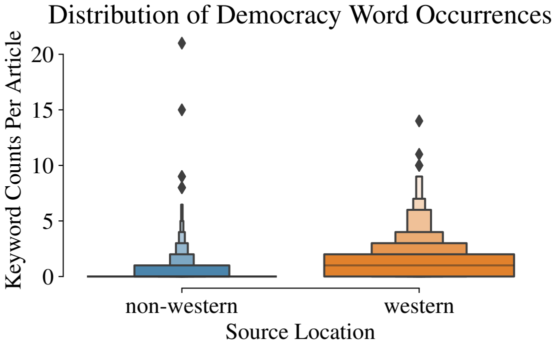

Figure 4: Quantile plot of "Democracy" Counts Per Source Location

5.2 Topic modeling

Table 2 shows the most prominent words for the 13 topics we identified in §4.2.

Figure 5 shows the evolution of topics over time, revealing that at several key points in Hong Kong's history, Western-based and Hong Kong–based sources wrote about different topics. This is not entirely unexpected for a number of reasons, including a media organization's possible desire to appeal to their own readership and therefore maintain loyal readers.

Furthermore, the local nature of Hong Kong–based media might encourage them to include more domestic events and details. This might be shown

Topic	Top 10 words
Students/schools	student, university, school, young, education, campus, class, group, family, child
Airports	station, mtr, airport, staff, service, yesterday, sha, cathay, day, airline
Legal	court, law, case, chan, legal, yesterday, wong, justice, mask, charge
Democracy movements	mr, democracy, leader, chinese, movement, pro, party, street, occupy, pro_democracy
Bills	bill, pro, council, extradition, lawmaker, election, party, legislative, mainland, camp
Foreign states	state, foreign, chinese, united, president, united_state, country, trump, international
Finance	cent, per_cent, hk, market, property, billion, company, million, price, sale
Chief executive	lam, bill, executive, chief_executive, carrie, carrie_lam, extradition, cheng, extradition_bill, demand
News	chinese, medium, mainland, taiwan, news, social, state, post, company, social_medium
Marches/rallies	march, rally, group, july, civil, june, yesterday, front, day, organiser
Police violence	officer, force, violence, gas, tear, tear_gas, attack, riot, police_officer, arrested
Mainland	mainland, world, event, number, day, tourist, local, ha_been, place, unrest
One country, two systems	law, system, national, country, security, central, one_country, rule_of_law, two_system, tung

Table 2: The 13 topics found and used in our topic modeling analysis.

by the pervasiveness of the Marches/Rallies topic and the Bill topic in Hong Kong–based media when compared to the presence of the same topic in Western-based media. Hong Kong–based newspapers may have reported any marches or rallies that took place between 1998 and 2020, whereas Western-based newspapers may have focused only on landmark ones such as those organized around the anniversaries of the July 1 Handover or the June 4 Tiananmen Square incidents. As for the Bill topic, Hong Kong–based media coverage peaks in 2010, when the Legislature debated a number of legal initiatives, whereas the western-based coverage of the same topic remain relatively stable and much lower overtime.

The topics reflect known events in Hong Kong's history; spikes in the students/schools topic track the Scholarist movement and its resurgence in 2014 in the Umbrella Revolution. Several spikes emerge around discussions of the election process for Hong Kong's chief executive.

However, at key points in Hong Kong's history of social unrest Western-based media and Hong Kong–based media the topics diverge completely. For example, in July 2019 Western-based newspapers reported police violence to a far greater extent than Hong Kong–based media.

5.3 Comparing lexical usage

The methods from §4.3 reveal semantic divergence in certain keywords between Western-based and non-Western-based news sources. We also find that June–July is a turning point, after which the meaning of several keywords shifts for at least the remainder of 2019.

Western-based vs. Hong Kong–based sources We divide the data by the location of each article's publisher. Corpus $\mathcal{C}_{\text{West}}$ is composed of all 711 articles published by Western-based sources. Corpus $\mathcal{C}_{\overline{\text{West}}}$ is composed of all 3464 articles published by Hong Kong–based sources.

We then trained word2vec models on both corpora. Despite the relatively small size of corpus W, a visual inspection of the resulting Word2Vec model shows sound performance. We then scored each keyword from Table 1 and compared each model's nearest neighbors.

We observe noticeable semantic differentiation between the two models for several keywords. For example, "resistance" has an unexpectedly low score. In comparison to the scores of all words that appear more than 100 times in both corpora, the score of "resistance" is only in the 17$^{\text{th}}$ percentile.

A visual inspection of the term's nearest neighbors in Table 3 for the Western-based model suggests an association with the feelings of protesters (ex. "frustration", "anxiety"). In contrast, the nearest neighbors of "resistance" in the Hong Kong–based model relate to adversarial behavior. This is evidence of the dichotomous framing of anti-government demonstrators.

Authors commonly employ the the words "tension" and "severe" to describe protest events and

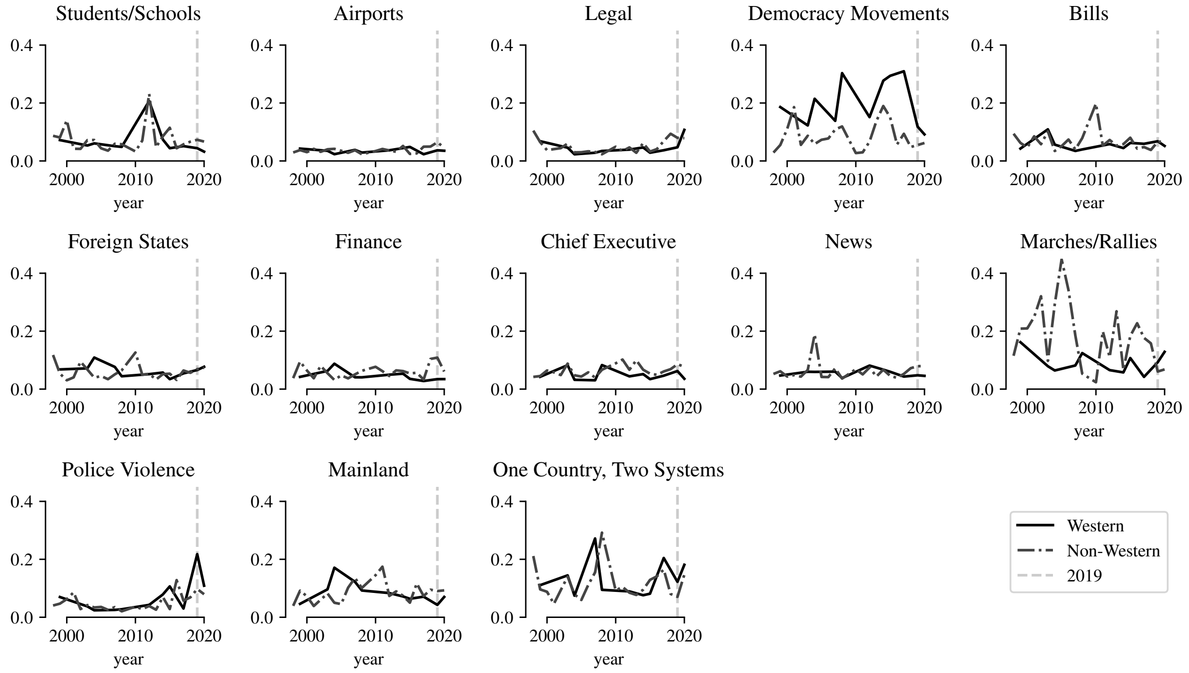

Figure 5: Mean topic representation (out of 1) over time, for Western and Hong Kong–based sources. 2019 is highlighted in dashed grey.

confrontations. The words "tension" and "severe" both had low similarity scores, with the score of the former in the 1st percentile and the score of the latter in the 7th percentile. This is evidence of high semantic divergence between Western and non-Western news sources in their usage of polarizing framing.

Curiously, "protest" scored only in the 91st percentile. We attribute this finding to be a function of low prevalence of the word in Hong Kong–based protests, which may also betray self-censorship. Additionally, we interpret the finding to mean that the context in which "protest" occurs is not dissimilar in our two corpora.

Before vs. after July 2019 Here, we sought to quantify the degree to which the introduction of the Fugitive Offenders amendment bill acted as a pivotal moment in the style of newspapers' portrayal of the Hong Kong protesters.

We again obtain the scores of words with a frequency higher than 100 in both corpora to contextualize our keywords' scores. We find that "resistance" again has a low score, and therefore high semantic shift. We inspected its nearest neighbors in each model and saw that the term became associated with dissent in the months after July 2019. Examples are shown in Table 4.

We note a similar trend for the words "confront" (9th percentile) and "confrontation" (11th percentile). After July 1, confrontations became associated with words like "provocative", "battles", and "mayhem". These changes may be suggestive of how English-language Hong Kong–based newspapers intended to shape the international understanding of what was happening in Hong Kong, favoring the inclusion of strong and negative terms to portray the 2019 street protests.

5.4 Sentiment analysis

Here we report the mean article sentiments across news sources, following §4.4. We find a consistent pessimism across news sources: all display positive sentiment in only 30 % to 40 % of their content. While no clear-cut relationship can be established between whether an article is from a western source from its sentiment, Hong Kong–based sources are more negative.

There is, however, internal variation. The *China Daily* with a share of 37.9 % positive articles is the second most positive in sentiment, following the *Wall Street Journal*, whereas the *South China Morning Post*, with a share of 30.5 %, displayed the least positive sentiment across articles.

“resistance”		“tension”		“protest”	
Western-based	Hong Kong–based	Western-based	Hong Kong–based	Western-based	Hong Kong–based
conflict	dissent.	demonstrating	worsening	streets	rallies
beyond	approaches	careful	disputes	violent	rally
frustration	pragmatism	saw	tensions	umbrella	campaign
cited	insurrectionists	cars	continues	movement	non-cooperation
helps	odds	treated	controversies	thousands	demonstrators
anxiety	adversaries	eyes	risks	demonstrations	demonstrations
meant	outpouring	walked	crises	march	movement
word	nerve	watched	turmoil	clashes	citywide
stark	inflict	bus	conflict	hundreds	demonstration
uprisings	craft	deleted	divisions	marched	strike

Table 3: 10 nearest neighbors of three words in Western-based source model vs. Hong Kong–based source model

“resistance”		“confront”		“confrontation”	
Pre-July 2019	Post-July 2019	Pre-July 2019	Post-July 2019	Pre-July 2019	Post-July 2019
canada	uprising	yan	alike	chance	scenes
global	eroding	station	innocent	break	confrontations
influence	humane	glass	resorting	procedural	mayhem
governments	dissent	chat	motivated	letting	battles
deep	sow	minutes	letting	leave	chaotic
reverse	advocating	wore	provoke	circumstances	tense
woman	define	yuen	endangering	refusing	respite
initial	anthems	lines	deny	agreed	frequent
relationship	authoritarianism	walls	treat	rational	clashes
growing	labelling	throwing	insulting	based	stand-offs

Table 4: 10 nearest neighbors of three words in Western-based source model vs. Hong Kong–based source model

6 Conclusion

We show that techniques from natural language processing can guide, answer, and suggest questions in social science. While past work focuses on single movements or eras, we characterize the portrayal of civil unrest in Hong Kong over a period of 22 years. Using a curated and manually filtered corpus of 4175 articles from Western-based and Hong Kong–based newspapers, we identified clear differences in framing both across time and between Western-based and Hong Kong–based newspapers.

Our approaches shed light on the ways in which Western and Hong Kong–based portrayals have evolved over time. For instance, while both discussed the Scholarist movement’s rise to prominence in 2012 in roughly equal proportions, the discussion of police violence was much more prominent in Western sources than in Hong Kong–based sources. Similarly, Western-based sources are far more likely to discuss protests than Hong Kong–based sources. This has implications for the extraction of protest-related events from corpora with politically opposed sources such as ours. Further, July 1, 2019 marked a turning point across Western and non-Western sources in the characteristics of usage for confrontation-related vocabulary.

The efficacy of event extraction models presupposes that the event in question is discussed in the considered collection of documents. In characterizing significant differences in portrayal across news sources, we implore that a critical eye be applied to the data *selection* process. We are working to quantify the degree to which event extraction systems are stymied by content and framing differences.

Finally, we have binned all articles from each year for much of our analysis. This blends news coverage leading up to unrest and portrayals of it afterward. Does language in news media cause (or at least, Granger-cause) protest sizes? Future work will more precisely measure differences in news content and framing around points of civil unrest.

References

Alan Agresti. 2017. *Statistical methods for the social sciences*. Pearson.

Nasser Alsaedi, Pete Burnap, and Omer Rana. 2017. Can we predict a riot? Disruptive event detection using Twitter. *ACM Trans. Internet Technol.*, 17(2).

Bernard Berelson. 1952. *Content analysis in communication research.* Free press.

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *HLT-NAACL*, pages 327–337.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. 2020. Analyzing ELMo and DistilBERT on sociopolitical news classification. In *Proceedings of the Workshop on Automated Extraction of Sociopolitical Events from News 2020*, pages 9–18, Marseille, France. European Language Resources Association (ELRA).

Lalindra De Silva and Ellen Riloff. 2014. User type classification of tweets with implications for event recognition. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 98–108, Baltimore, Maryland. Association for Computational Linguistics.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Jennifer Earl, Andrew Martin, John D. McCarthy, and Sarah A. Soule. 2004. The use of newspaper data in the study of collective action. *Annual Review of Sociology*, 30(1):65–80.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

C.W.J. Granger. 1988. Some recent development in a concept of causality. *Journal of Econometrics*, 39(1):199–211.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 624–628, New York, NY, USA. Association for Computing Machinery.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas U.S. history textbooks. *AERA Open*, 6(3):2332858420940312.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Frederick Mosteller and David L Wallace. 1984. *Applied Bayesian and classical inference: the case of the federalist papers*. Springer.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia. Association for Computational Linguistics.

William H. Overholt. 2021. Hong Kong: The rise and fall of "one country, two systems". Harvard Kennedy School.

Konstantina Papanikolaou and Haris Papageorgiou. 2020. Protest event analysis: A longitudinal analysis for Greece. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 57–62, Marseille, France. European Language Resources Association (ELRA).

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Philip A Schrodt, Erin M Simpson, and Deborah J Gerner. 2001. Monitoring conflict using automated coding of newswire reports: a comparison of five geographical regions. In *Conference 'Identifying Wars: Systematic Conflict Research and it's Utility in Conflict Resolution and Prevention', Uppsala*, pages 8–9. Citeseer.

Justin Sech, Alexandra DeLucia, Anna L. Buczak, and Mark Dredze. 2020. Civil unrest on Twitter (CUT): A dataset of tweets to support research on civil unrest. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 215–221, Online. Association for Computational Linguistics.

David Snyder and William R. Kelly. 1977. Conflict intensity, media sensitivity and the validity of newspaper data. *American Sociological Review*, 42(1):105–123.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Meredith Leigh Weiss and Edward Aspinall. 2012. *Student activism in Asia: Between protest and powerlessness*. U of Minnesota Press.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

Bruno Wueest, Klaus Rothenhäusler, and Swen Hutter. 2013. Using computational linguistics to enhance protest event analysis. In *ENCoRe Workshop 'Tools and Techniques for Conflict Event Data Collection'*, Konstanz.