A Mixed-Methods Analysis of Western and Hong Kong–based Reporting on the 2019–2020 Protests

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

Abstract

We apply statistical techniques from natural language processing to Western and Hong Kong–based English language newspaper articles that discuss the 2019–2020 Hong Kong protests of the Anti-Extradition Law Amendment Bill Movement. Topic modeling detects central themes of the reporting and shows the differing agendas toward *one country, two systems*. Embedding-based usage shift (at the word level) and sentiment analysis (at the document level) both support that Hong Kong–based reporting is more negative and more emotionally charged. A two-way test shows that while July 1, 2019 is a turning point for media portrayal, the differences between western- and Hong Kong–based reporting did not magnify when the protests began; rather, they already existed. Taken together, these findings clarify how the portrayal of activism in Hong Kong evolved throughout the Movement.

Hong Kong–based	Western-based
tensions	
us-china, **tension**, **war**, dispute, **uncertainty**, heightened, prolonged, **worsening**, fallout, **turmoil**	culture, state-owned, protections, tourists, market, base, rise, travel, closer, **argued**

Table 1: 10 nearest neighbors of *tensions*; words indicating negative emotion in LIWC (Pennebaker et al., 2015) are bolded. The Western sources' use of *tensions* is more descriptive and impassive (§5.1).

1 Introduction

On February 12, 2019, the Hong Kong Government revealed its plan to amend the Fugitive Offenders Ordinance and the Mutual Legal Assistance in Criminal Matters Ordinance (Li, 2019). Hong Kong Chief Executive Lam stated that the bill had been inspired by Poon Hiu-wing's murder: a judicial loophole was preventing the Hong Kong courts from trying her killer for murder, and it needed to be closed (Victor and May, 2019). Chief Executive Lam proposed changing the law so that Hong Kong could extradite fugitives, on a case-by-case basis, to jurisdictions it did not have an agreement with, including Mainland China, Macau, and Taiwan.

Opponents of the bill argued that rather than fixing a loophole, Chief Executive Lam's amendment would open the door to extraditions between Hong Kong and Mainland China, something that Beijing sought ever since the 1997 handover. To pass the amendment to the bill before Poon's murderer was released from prison and could (in theory) flee Hong Kong, citizens were only given 20 days to offer their feedback. They saw the bill as threatening their way of life and voiced dissent following the end of the short public consultations. Marches and demonstrations morphed into months of discontent, which were not appeased by Lam's decision to withdraw the bill (Ramzy and Yu, 2019). By then, the protests had evolved into an increasingly violent anti-government, anti-Beijing movement, with demands for greater democracy and police accountability. The Government turned to powers offered by the Emergency Regulations Ordinance (Hong Kong e-Legislation) to end the protests.

From a theoretical perspective, this article can be positioned within what Oliver et al. broadly call "the world of news and politics" (2019, 36) and, especially, how "dissonance across media outlets could be socially and politically consequential" (Tsfati and Walter, 2019, 39). The research behind the article is based on a mixed-methods approach that combines the application of a variety of NLP techniques with experts' insights on Hong Kong politics and society. NLP aids in marking inconsistencies in event characterizations as we analyze news articles related to episodes of civil unrest between January 2019 and June 2020, in both western- and

Proceedings of LaTeCH-CLfL 2021, pages 178–188
Punta Cana, Dominican Republic (Online), November 11, 2021.

Hong Kong–based English-language newspapers. Experts' insights help put into perspective how the volatile context of Hong Kong politics, plus newspapers' tendency to report more dramatic events, may result in reporting bias that either emphasises or undermines the legitimacy of either the protests or the regime against which protests are directed (Snyder and Kelly, 1977; Schrodt et al., 2001; Earl et al., 2004). The powerful mixed-methods interplay guides the analyses and enables this work's rich and nuanced discussion of these three research questions about newspaper depiction of the protests:

(§4) *What* do Hong Kong–based and Western-based newspapers discuss about protests in Hong Kong?

(§5) Do Hong Kong–based and Western-based newspapers differ in *how* they portray the protests?

(§6) Does coverage differ before and after the outbreak of protests in June 2019?

2 Related work

In answering our three questions, we benefit from the methods of *content analysis* (Berelson, 1952) by complementing subject matter expertise with the potential for massive scale. In this vein, Lucy et al. (2020) use word embedding similarity, topic models, and dependency parsing to generate clues toward differing portrayals of race and gender in U.S. history textbooks. Field et al. (2018) relate the content of Russian state-run news articles to the nation's economic performance, finding an agenda of distraction. Other content analysis and stylometry considers authorship (Mosteller and Wallace, 1984; Bergsma et al., 2012), native language identification (Koppel et al., 2005; Bergsma et al., 2012), and deceptive product reviews (Ott et al., 2013).

Considering civil unrest generally, Wueest et al. (2013) apply topic models and named entity recognition to protest event analysis. Hürriyetoğlu et al. (2019) show that not all protests are alike, computationally speaking: event extraction models for protests perform much worse on countries outside the training set. Inverting this, we call into question different views on protest in the same location.

Within the specific focus of Hong Kong, the closest work to ours is the contemporaneous Scharf et al. (2021), who perform a longitudinal analysis of Hong Kong news media since Hong Kong's transfer of sovereignty. Our work "zooms in" on the 2019–2020 protests—the largest in Hong Kong's history.

3 Data

We rely on a corpus we have collected of 3398 news articles 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: *China Daily* and *South China Morning Post*, over a period of 18 months, from January 2019 to June 2020. Although we took steps to include a diverse array of newspapers from within given regions, the newspapers used in this sample were purposefully selected because they were English-language newspapers. This still allows for important insights into differences across cultures and regions. Future research ought to extend news coverage of protest beyond English-only news sources (Earl et al., 2004; Lee, 2014; Du et al., 2018).

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, and be about the protests.

A one-by-one, manual screening process eliminated irrelevant items such as eventual duplicates within each publication, readers' letters, and articles that included any of the chosen keywords but whose content was not about the protest incidents. Following the manual screening, we retained 3398 articles. The mean length was 783 tokens.

4 *What* do Hong Kong–based and Western-based newspapers discuss about protests in Hong Kong?

The foremost question we answer is about the lay of the land: what is discussed? In this section, we first look at the frequency of articles about protests in Hong Kong: over time, how much are the protests

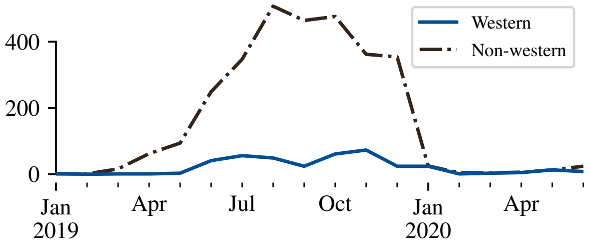

Figure 1: Hong Kong–based sources published far more throughout the protests and tapered off later.

discussed at all? Next, we discover topics that the newspapers tie into the protests, like legislation, air travel, and Mainland China. Bolstered by this, statistical analysis of word counts uncovers a discrepancy between western- and Hong Kong–based sources' use of 15 of 19 protest-related keywords like "protest" and "freedom".

4.1 Publication frequency

Counting articles about a theme is a simple, crude way to gauge its interest or relevance. However, newfound censorship in Hong Kong[1] adds a dimension to these numbers. Here, we report the trends in the number of collected articles over time, and we discuss the causes of the trends.

South China Morning Post and *The New York Times* published the largest number of articles about protests in Hong Kong. *South China Morning Post* published the most articles on Hong Kong protests amongst all newspapers, and *The New York Times* published the most among western-based newspapers. Both Hong Kong– and western-based newspapers published the lion's share of their articles in the second half of 2019 (see Figure 1). Out of a total of 3268 articles published in 2019, 94 % of them were published between June and December 2019.

There are a few reasons for this symmetry. First, *if it bleeds, it leads* (Altheide, 2018). Citizens' disquiet over the proposed amendment to the extradition bill boiled over into direct and violent action during the summer and fall months and, perhaps unsurprisingly, resulted in more articles published in both Hong Kong– and western-based newspapers. Then, the anniversary of the July 1, 1997 handover may have played a role. Historically, Hong Kong–based newspapers have used the handover anniversary as an opportunity to highlight aspects of the *one country, two systems* principle that are proven beneficial for the economic and political wellbeing of both Hong Kong and Mainland China. Western-based newspapers, on the other hand, tend to focus on the slow erosion of the very independence that *one country, two systems* was intended to preserve and guarantee.

Hong Kong has a long tradition of a free and opinionated media, with broad readership in both Cantonese and English. As the main English-language outlet in Hong Kong, *SCMP* shapes the international understanding of events in Hong Kong, whereas the (Hong Kong edition) *China Daily* in its efforts to connect Hong Kong readers to Mainland China finds itself consistently highlighting the *one country* perspective of the *once country, two systems* principles. These differences punctuated these newspapers' coverage of the 2019–2020 protests. *SCMP* published (roughly) ten times more articles than *China Daily* over the course of the 18-month timeline of our research. The articles' style was also markedly different, with the SCMP favoring the use of juxtapositions between positive and negative views towards the government, and the use of polls as indicators of public opinion's sentiment, whereas *China Daily* used academics as apolitical authorities on public affairs, and often adopted rhetorical strategies for the articles to read neutral and objective.

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 the 18 months of our study.

Method 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 $k = 7$ to 60, and evaluated the coherence of the resultant topics according

[1] https://www.nytimes.com/interactive/2020/09/04/world/asia/hong-kong-speech.html

Topic	Top 10 words
Finance	cent, per_cent, hk, market, property, billion, million, price, sale, financial
Social Media	n't, movement, young, have_been, political, do_n't, life, ha_been, hongkongers, social
Legislation	extradition, taiwan, lawmaker, council, election, chan, law, party, mainland, camp
Air Travel	business, mainland, airport, service, staff, august, day, company, tourist, cathay
International	china, chinese, state, world, trade, united, beijing, international, trump, mainland
Force and Order	officer, force, public, court, yesterday, case, arrested, law, post, source
Mainland	beijing, law, china, system, national, country, chinese, central, affair, two_system
Clashes	station, officer, gas, tear, violence, road, sunday, march, rally, tear_gas
Students	student, university, school, group, campus, ho, education, support, event, more_than
Chief Executive	lam, chief, executive, chief_executive, public, carrie, cheng, carrie_lam, extradition, political

Table 2: The 10 topics found in news coverage of Hong Kong's 2019–2020 protests.

to Mimno et al. (2011). We found that using 10 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 samples 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.[2]

Findings Figure 2 shows how the prominence and timing of the ten topics from Table 2 differ. We may expect these differences because of a media organization's desire to appeal to their own readership, preference for big picture issues, and the fact that the local nature of Hong Kong–based media might encourage a focus on the details of domestic issues. The treatment of the topics of MAINLAND, LEGISLATION, and SOCIAL MEDIA help showcase these differences.

MAINLAND Western-based newspapers' coverage of MAINLAND peaks twice. First at the time when the proposal to amend the extradition bill is announced, the short public consultations opened, and citizens become vocal about the amended extradition bill likely to weaken Hong Kong's independent judiciary. The second time, following Beijing's decision on Hong Kong's new security law, whose drafting and approval bypassed Hong Kong's legislature and citizenry. These trends suggest that Western-based newspapers may have used the coverage of the extradition bill to show the failure of the *one country, two systems* concept, and strengthen the perception that the terms of the 1984 Sino-British Declarations were being violated with no recourse for the people of Hong Kong. A toned down coverage of the topic MAINLAND underscores the domestic nature of the issue for the Hong Kong–based newspapers. Coverage, however, picks up at the tail end of 2019, after the amendment proposal was formally withdrawn from LegCo[3] legislative agenda, and even more over the

[2] As a complement to topic models, we also used Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015; Tausczik and Pennebaker, 2010) to analyze the language used in the articles. LIWC characterizes text based on word counts across more than 70 morphosyntactic and psychometric dimensions. There were only discernable trends in two of these categories, so we omit this from further discussion. "Newspaper-ese" has some common stylistic elements across the world, and the subject matter we consider shares one focus: the protests in Hong Kong. We speculate that this leads to the consistency across LIWC categories. The SHEHE and I categories occur with similar frequency because of newspapers' aggregate tendencies to use third- or first-person pronouns with particular frequency. Meanwhile, the INGEST category remains rare due to its irrelevance to the protests. Further, the LIWC categories include finite pre-defined lists of words. We found that the word "fallout", for instance, is not listed in the NEGEMO category, despite its readily apparent negative connotation. Topic modeling adapts to a corpus's extant vocabulary, giving it greater flexibility for our purposes.

[3] The Legislative Council, Hong Kong's legislature, comprises 70 members, 35 of whom are directly elected through five geographical constituencies under a proportional representation. The other 35 are indirectly elected through interest-group-based functional constituencies with limited electorates. Since its establishment in 1843 as an advisory council to the

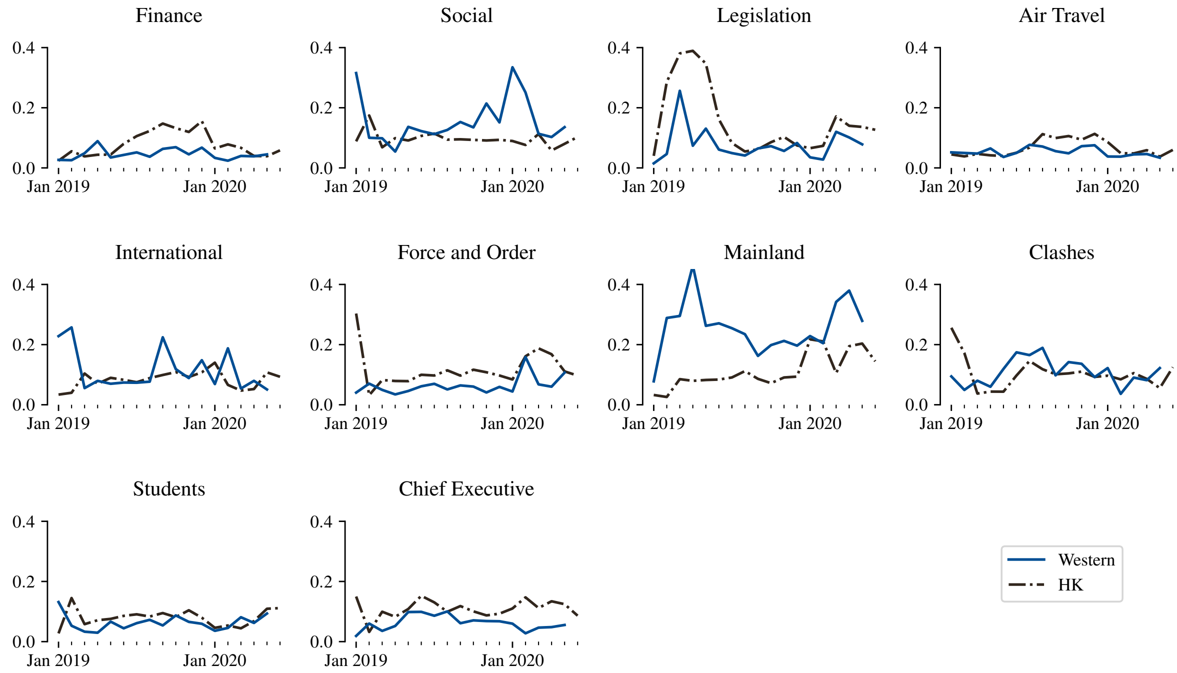

Figure 2: The 10 topics found and used in our topic modeling analysis.

first six months of 2020, (possibly) in an effort to address allegations that Beijing used the amendment to the extradition bill to usher in limitations of Hong Kong's existing civil and political liberties.

LEGISLATION Hong Kong– and western-based newspapers' trends for the coverage of the topic of LEGISLATION mimic each other; however, the volume of coverage is different. Hong Kong–based newspapers published consistently more on LEGISLATION than western-based ones. This is especially the case in the first half of 2019, when the proposal to amend the existing extradition bill was introduced to LegCo, the short public consultations held, and citizens began voicing their criticism; following this, Chief Executive Lam decided to indefinitely postpone the LegCo's vote on the amendment. The local nature of Hong Kong–based media also help explain the larger volume of coverage of an issue that is essentially domestic. In Western-based newspapers coverage is both less voluminous and less consistent. Two, distinct peaks underscore the announcement of the amendment proposal, and Chief Executive Lam's decision to indefinitely postpone the LegCo's vote on the amendment.

SOCIAL MEDIA Social media characterized the protests that engulfed Hong Kong starting in 2019, nurtured a sense of community based on collective, horizontal, and participatory decision-making. As the battleground in Hong Kong's protests, social media emerges as a topic whose coverage in Hong Kong– and western-based media notably diverges. In Western-based media, the discussion of social media peaks and dips as citizens leverage the catalytic features of social media in civic participation to voice their concerns about amending the extradition bill (in the first half of 2019), and civil disobedience, when concerns for the proposed amendment morphed into discontent over the lack of democratic political reforms, starting in the summer of 2019 and continuing through the spring of 2020. Hong Kong–based media's coverage of social media is consistently low, possibly reflecting both the fact that protests' social media use created digital space for activism, and Beijing's unhappiness with such a difficult-to-counter accomplishment.

Finally, we contrast our findings with those of Scharf et al. (2021). Their longitudinal analysis of unrest in Hong Kong found that the topic of POLICE VIOLENCE became especially prominent in western sources in 2019. Their topic shares many words with our CLASHES and FORCE AND ORDER topics; still, we see no marked discrepancy between these during coverage of the protests.

Governor, its powers and functions have expanded.

4.3 Comparing lexical frequency

Having categorized salient themes in the articles, we now turn to the use of specific, protest-relevant keywords chosen by a subject matter expert. 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.

We first test for whether there are important differences in the frequencies of 19 protest-related keywords[4]. Our procedure is related to Scharf et al. (2021), though we employ the Holm–Bonferroni correction (Holm, 1979) instead of the overly conservative, low-power Bonferroni correction. We set a significance level of $\alpha = 0.01$.

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.

Findings The results of the analysis show that 15 of our 19 selected keywords have statistically significant differences in frequency, with *democracy* ($F = 490.5$); *protest* ($F = 354.6$); *protests* ($F = 300.6$); *freedom* ($F = 137.2$); and *occupation* ($F = 119.4$) having the highest F statistics. Our analysis consistently found less discussion of *protests* in Hong Kong-–based sources, with the high values of F statistic for *protest* and *protests* pointing to a statistically significant disparity in the coverage of protests.

Figures 3 and 4 illustrate these differences for *democracy* and *freedom*. In western-based media, the evolution of citizens' dissent over the amendment is cast as citizens' fight for democracy and the freedoms that come with it, resisting the authoritarian tightening in post-handover Hong Kong. This motivates the frequent and recurrent use of *freedom* and *democracy*. Hong Kong–based newspapers, on the other hand, use *democracy* and *freedom* sparingly, focusing their discourse narrowly on the actual protest events rather than a broader meaning of them. The following section works to further characterize this contrast.

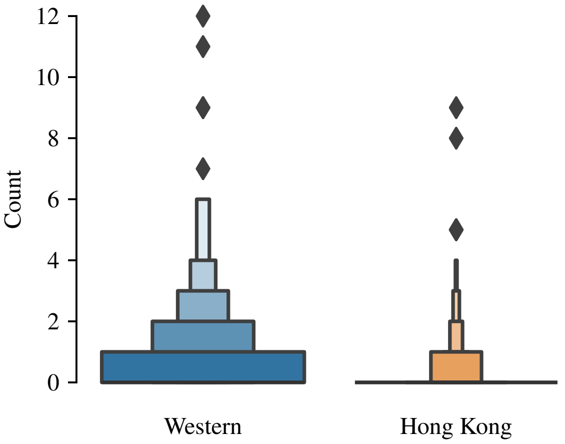

Figure 3: Quantile plot of *democracy* counts per article, by source location

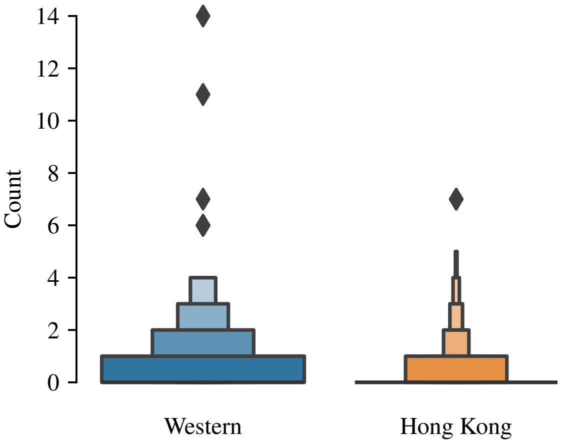

Figure 4: Quantile plot of *freedom* counts per article, by source location

5 Do Hong Kong–based and Western-based newspapers differ in *how* they portray the protests?

Evidence from word embedding neighborhoods shows that Hong Kong–based sources use more negative, emotionally charged imagery; western-based depictions of concepts like *protest* and *tension* are more impassive and detached.

While the methods from §4 consider *which* words are used, we would like to investigate the evolution of *how* words are used differently between Western- and Hong Kong–based sources. In

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

Hong Kong–based	Western-based
confront (17th percentile)	
retreat, intimidated, abused, reminded, understandable, upset, confronting, provoked, regularly, provoke	work, met, acts, spirit, reasons, decisions, trying, voice, intolerable, tsang
protest (66th percentile)	
rally, sit-ins, demonstration, demonstrators, rallies, campaign, strike, movement, march, demonstrations	demonstrations, umbrella, peaceful, movement, began, streets, activists, mass, 1m, referendum
tensions (63rd percentile)	
us-china, tension, war, dispute, uncertainty, heightened, prolonged, worsening, fallout, turmoil	culture, state-owned, protections, tourists, market, base, rise, travel, closer, argued

Table 3: 10 nearest neighbors of three representative keywords in the Hong Kong–based and western-based articles. For *confront*, HK-based reporting is more emotionally charged. For *protest*, language from both sources is clinical, but HK-based language remains more negative. For *tensions*, the HK-based coverage is emotionally charged, higher-stakes, and more negative.

§5.1, we measure how the contexts that our protest-related keywords differ across sources. Sentiment analysis supports the findings that Hong Kong–based sources' writing is more negative. Analysis of articles' headlines shows that Western sources' headlines are more judgmental.

5.1 Comparing embedding neighborhoods

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. The same procedure can identify differences between words' usage when separated by something other than time. 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 $\mathcal{M}_{\overline{a}}$.[5]
5. Score the usage-change of w as the size of the intersection, $\left|\mathrm{NN}_a(w) \cap \mathrm{NN}_{\overline{a}}(w)\right|$.

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 word w's similarity score fall? We find this to be more meaningful than the raw similarity score.

The analysis of lexical usage reveals semantic divergence in certain keywords between Western-based and Hong Kong–based news sources, as shown in Table 3.

A visual inspection of the term's nearest neighbors for the Western-based model suggests the prevalence of neutral or descriptive lexicon (e.g., *work*; *reason*; *rise*). In contrast, the nearest neighbors in the Hong Kong– based model relate to adversarial behavior as in the case of *confront* (e.g.,

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

retreat; *intimidated*; *provoke*); *protests* (e.g., *rally*; *strike*; *march*); *tension* (e.g., *war*; *dispute*; *uncertainty*; *turmoil*); and *rule of law* (e.g., *safeguard*; *undermined*). These trends are evidence of the dichotomous framing of anti-government demonstrators as well as their actions only having a negative connotation for Hong Kong.

5.2 Sentiment analysis

We corroborate the findings of more negative tone in Hong Kong–based publications with computational sentiment analysis. Sentiment analysis measures the attitude of an author from the tone and connotations of their document. While it is common and interpretable to use hand-crafted 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[6] $(+, -)$ 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.

The Hong Kong–based articles' average positivity (31.4 %) is slightly lower than the western-based articles (32.9 %). Across sources, there is wide variation, though; in fact, *The Times* is the most negative overall (28.3 % positive) but writes about the protests rarely. *China Daily* has the most positive tone. Coupled with the word-level negativity we noted in §5.1, this document-level negativity suggests that the Hong Kong press coverage of protests is systemically more negative. While the sentiment score sheds little light on authors' attitudes behind articles and commentaries, it still provides evidence of stylometric differences between news sources. This combined with the analysis of lexical usage and topic modeling helps to construct an informative view into news portrayal of civil unrest in Hong Kong.

5.3 Analysis of headlines

Headlines are meant to catch readers' attention while deliberately seeking to influence through the use of narrative mechanisms and sensational or provoking words (Blom and Hansen, 2015). We believe that both the length and the tone of the headline are informative about attempts to influence.

63 % of articles in the corpus have long headlines (i.e., include six or more words), whereas the remaining 37 % have headlines with fewer than six words. The *New York Times* emerges as the newspaper with the highest likelihood to have telegraphic headlines (i.e., 9.7 times more likely), whereas *South China Morning Post* is the least likely to have short headlines (i.e., 11 % less likely).[7]

In terms of sentiment, 52 % of all headlines are neutral, 32 % judgmental, and 15 % are mild, with the Hong Kong–based newspapers being the least likely to have judgmental headlines (i.e., 52 % less likely), and—perhaps unsurprisingly—the *Financial Times* being the most likely to have judgmental headlines (i.e., 2.5 times more likely).[8]

6 Does coverage differ before and after the outbreak of protests in June 2019?

Given that we have found differences between western- and Hong Kong–based reporting on the 2019–2020 protests, we might ask whether there are differences *in these differences* over time. Responding to our observation in §4.3, we use a two-way test and discover no significant interaction between source and the start of the protests on word frequency. We therefore restrict our focus to the aggregate changes in word usage over time, finding that the semantic context of the protest keywords becomes more polarizing and intense.

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

[7] The full regression model containing all predictors was statistically significant, $\chi^2(8; N = 3398) = 726.788, p < 0.001$. The model correctly classified 74 % of cases, and explained between 14.9 % (Cox and Snell R^2) and 20.5 % (Nagelkerke R^2). The strongest predictor for short headlines is the variable for *The New York Times* with an Exp($\hat{\beta}$) of 9.7, whereas the weakest predictor for short headlines with the *SCMP* with an Exp($\hat{\beta}$) of .89.

[8] The full regression model containing all predictors was statistically significant, $\chi^2(8; N = 3398) = 190.459, p < 0.001$. The model correctly classified 68.7 % of cases, and explained between 4.1 % (Cox and Snell R^2) and 5.8 % (Nagelkerke R^2). The strongest predictor for judgmental headlines is the variable for *Financial Times*, with an Exp($\hat{\beta}$) of 2.5, whereas the weakest predictor for a judgmental headlines is the variable for *SCMP*, with an Exp($\hat{\beta}$) of .48.

6.1 Lexical frequency and interaction terms

A two-way test lets us compare the means of a continuous response variable, modulated by two categorical explanatory variables. This expands on the analysis in §4.3: we can test the significance of not only the two explanatory variables, but also an interaction term between the two. Again, we use the Holm–Bonferroni correction to mitigate false discovery.

In our case, the explanatory variables are the source (western/HK-based) and the date: was the article published before or after July 1, 2019? There were significant differences for five of the nineteen keywords (*unrest*, *democracy*, *rights*, *crackdown*, *protest*). Further, the interaction term is only significant for *occupation* ($p = 2.3 \times 10^{-5}$); for the other 18 keywords, there is no significant interaction between the source and the date. Whatever trends and biases existed were not discernibly altered by the onset of the protests.

6.2 Comparing lexical usage over time

In our analysis, 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, and found that June 2019 emerges as a turning point, after which the meaning of several keywords shifts for at least the remainder of 2019.

We now split the corpus in "pre-June 30th, 2019" and "post-June 30th, 2019" to investigate whether the way in which Hong Kong– and western-based newspapers portrayed episodes of civic unrest differently following the protests and demonstrations that took place over June 6–12th, 2019.

Neighborhood shift analysis (as in §5.1) revealed significant low scores for *resistance*, *severe*, *riots*, *confront*, *confrontation*, and *terrorism*, which suggests that the context and/or semantic meaning for these words changes from early to late 2019, regardless of whether Hong Kong– or western-based news sources are considered. For instance, in the first half of 2019, neighbors for *riots* include terms like *actions*, *open*, *engage*, and *taken*, which that are not charged, and in the context of either a reporting or an opinion piece descriptively inform readers.

However, in the second half of 2019, the nature of neighboring terms for *riots* changes to include more polarizing terms such as *violent*, *escalated*, *destructive*, *triggered*, *anti-government*, and *sparked*. Similarly, pre-July 2019, neighbors for *terrorism* include, among others, terms like: *covered*, *lawyers*, and *negative*. Post-June 2019, neighboring words become politically charged, and include *criminals*, *destructive*, *extreme*, *lawless*, *punishing*, and *barbaric*. As "Hong Kong Summer of uprising" unfolded (Lee et al., 2019), feelings of social danger prevailed in newspapers' accounts of the events. Citizens' willingness to assert their political and civic agency via marches and demonstartions was described more and more harshly over time. Articles pre-July 2019 focused on general descriptions of protestors' tactics, whereas post-June 2019 the writing focused on detailed description of violent actions that took place during the protests as well as mentions of the negative social impacts that such actions may cause, which are implied to be detrimental to the well being of Hong Kong.

7 Conclusion

The Hong Kong protests captured domestic and international media attention, and their news value increased as their objectives shifted from protesting against a proposed amendment to the existing extradition bill to protecting democracy.

Our research shows that techniques from natural language processing help strengthen our understanding of how the portrayal of activism in Hong Kong profoundly evolved over the 18-month timeline of our study. While Hong Kong– and western-based publishers cover similar topics in their articles on the 2019 protests, the framing of those topics is different, which in turn carries implications for the public's perception, understanding, and support of those events. We also show that, in general, the use of charged words is more prominent in Hong Kong–based sources than western-based ones. This has implications for the extraction of protest-related events from corpora with politically opposed polarities such as ours. Finally, July 1, 2019, emerges as a turning point both for Hong Kong– and western-based news sources: afterwards, lexical choice related to protest descriptions becomes more negative and hostile in articles from both sources.

In exposing these findings, we combined expert domain knowledge and techniques from natural language processing. The interplay allows faster, more efficient analysis with larger sample sizes, which can be contextualized and interpreted to go beyond the descriptive and uncover the "So what?"

References

David L Altheide. 2018. *Creating fear: News and the construction of crisis*. Routledge.

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.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100.

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).

Y. Du, Lingzi Zhu, and Fan Yang. 2018. A movement of varying faces: How "occupy central" was framed in the news in Hong Kong, Taiwan, mainland China, the UK, and the U.S. *International Journal of Communication*, 12(0).

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.

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.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hong Kong e-Legislation. Cap. 241 emergency regulations ordinance.

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, Osman Mutlu, and Arda Akdemir. 2019. Overview of CLEF 2019 Lab ProtestNews: Extracting protests from news in a cross-context setting. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 425–432. Springer.

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.

Francis L. F. Lee. 2014. Triggering the protest paradigm: Examining factors affecting news coverage of protests. *International Journal of Communication*, 8(0).

Francis L. F. Lee, Samson Yuen, Gary Tang, and Edmund W. Cheng. 2019. Hong Kong's summer of uprising: From anti-extradition to anti-authoritarian protests. *China Review*, 19(4):1–32.

Jeff Li. 2019. Hong Kong-China extradition plans explained. *BBC News*.

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.

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

Mary Beth Oliver, Arthur A Raney, and Jennings Bryant. 2019. *Media effects: Advances in theory and research*. Routledge.

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.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Austin Ramzy and Elaine Yu. 2019. Hong Kong's leader, Carrie Lam, to withdraw extradition bill that ignited protests. *The New York Times*.

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.

James Scharf, Arya D. McCarthy, and Giovanna Maria Dora Dore. 2021. Characterizing news portrayal of civil unrest in Hong Kong, 1998–2020. In *Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 43–52, Online. Association for Computational Linguistics.

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.

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.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Yariv Tsfati and Nathan Walter. 2019. The world of news and politics. *Media Effects: Advances in Theory and Research*.

Daniel Victor and Tiffany May. 2019. The murder case that lit the fuse in Hong Kong. *The New York Times*.

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.