Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2891–2900
Marseille, 20-25 June 2022

Hong Kong: Longitudinal and Synchronic Characterisations of Protest News between 1998 and 2020

Arya D. McCarthy♦ and **Giovanna Maria Dora Dore**♠
♦Center for Language and Speech Processing, Johns Hopkins University
♠Krieger School of Arts and Sciences, Johns Hopkins University

Abstract

This paper showcases the utility and timeliness of the *Hong Kong Protest News Dataset*, a highly curated collection of news articles from diverse news sources, to investigate longitudinal and synchronic news characterisations of protests in Hong Kong between 1998 and 2020. The properties of the dataset enable us to apply natural language processing to its 4522 articles and thereby study patterns of journalistic practice across newspapers. This paper sheds light on whether depth and/or manner of reporting changed over time, and if so, in what ways, or in response to what. In its focus and methodology, this paper helps bridge the gap between "validity-focused methodological debates" and the use of computational methods of analysis in the social sciences.

Keywords: dataset, Hong Kong, news, protests, longitudinal

1. Introduction

Protests constitute an important means in contemporary societies by which citizens voice their concerns. However, protests' ability to communicate their messages and achieve their outcomes depends significantly on whether and how mass media, and especially newspapers, portray them (Agnone, 2007; King, 2014). Newspapers, in fact, can either amplify and legitimise the protesters' voices (Gamson and Wolfsfeld, 1993) or marginalise and delegitimise protests by portraying them as dangerous or irrelevant (Boykoff, 2006; Small, 1994). To investigate the dynamic relationship between protests and newspapers, we have constructed the *Hong Kong Protest News Dataset*, an original collection of news articles covering the occurrence and evolution of protests in Hong Kong between 1998 and 2020.

Protest participation has long been an undercurrent in Hong Kong's political culture (Rawnsley and Rawnsley, 2002) dating back to British colonial rule, and has evolved from the bloody riots of the 1960s to the protests of 2019–2020, when up to two million people took to the streets against the proposed amendment to the Fugitive Offenders and Mutual Legal Assistance in Criminal Matters Legislation Bill (ELAB). Hong Kong protests captured the world's attention with defiant crowds commemorating the 1989 Tiananmen Square incidents, nostalgically marking the transfer of sovereignty from the UK back to China every July 1 since 1997, and students blockading roads for 79 days in the Admiralty district during the pro-democracy Occupy Central protests in 2014 (Weiss and Aspinall, 2012). The news value of these actions grew, as early demonstrations to voice dissent morphed into an increasingly violent anti-government, anti-Beijing movement with demands for greater democracy.

To showcase the relevance, utility, and timeliness of the *Hong Kong Protest News Dataset* in investigating longitudinal and synchronic characterisations of news protests in Hong Kong, we apply natural language processing (NLP) to study patterns of journalistic practice across newspapers, shed light on whether depth and/or manner of reporting changed over time, and if so, in what ways, or in response to what. As language is at the heart of our research, NLP emerges as especially important for its ability "to analyse signals ranging from simple lexical clues to word clusters to choices of syntactic structure" (Boydstun et al., 2014, 2) as well as its speed, scale, reliability, and granularity when analysing text.

This paper builds on Scharf et al. (2021) who consider a subset of our techniques with a preliminary form of the dataset, and McCarthy et al. (2021) who focus solely on the recent anti-ELAB protests. It also complements a small collection of articles, currently under review and/or being drafted, that aims at bridging the gap between "validity-focused methodological debates" (Baden et al., 2021, 13) and the use of computational methods of analysis in the social sciences.

2. The Dataset

As newspapers' coverage remains one of the most useful records of protest events (Earl et al., 2004), we took steps to include a diverse array of news sources, even though Chinese, including Hong Kong SAR, North American and British newspapers sit at opposite ends of the spectrum in terms of ownership and state control. The corpus of articles we construct comes from six western-based, English language newspapers: *The New York Times* (NYT), *The Wall Street Journal* (WSJ), *The Washington Post* (WaPo), *The Financial Times* (FT), *The Guardian*, and *The Times*; and two Hong Kong–based, English language newspapers: *China Daily* and *South China Morning Post* (SCMP).

Current NLP limitations of comparison across languages (Pires et al., 2019; Baden et al., 2021, *inter alia*) make the challenge of including news sources in

Cantonese and Mandarin formidable, and ultimately resulted in the decision to include only Hong Kong– and western-based English language newspapers (Earl et al., 2004; Lee, 2014; Du et al., 2018; Baden et al., 2021).

We use news articles focusing on eight non-randomly selected episodes of civic unrest in Hong Kong to compare their news value and newsworthiness in the volatile social and political setting of post-handover Hong Kong (Chan and Lee, 1984; Lee, 2014; Tsfati and Walter, 2019). We focus on (i) the 1998–2002 July anniversary marches; (ii) the 2003 protests against national security reform; (iii) the 2004–2019 July 1 protests; (iv) the 2006–07 save the Star Ferry Pier protests; (v) the 2012 Protests against Moral and National Education; (vi) the 2014 Occupy Central protests; (vii) the 2016 Riots; and (viii) the 2019–2020 anti-extradition protests. Taken together, these protests represent a sustained and organised citizens' effort asking Hong Kong and Chinese authorities for a clear and faster path to democratisation for Hong Kong (Chan, 2015; Wong, 2021) and as such have received significant coverage in both Hong Kong– and western-based newspapers.

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 edition for the NYT and WSJ, the UK edition for the FT, The Guardian, and The Times, and Hong Kong edition for China Daily. To be eligible for collection, articles had to be at least 300 words long and to focus on the protests. A one-by-one, manual screening process eliminated irrelevant items such as eventual duplicates within each publication, readers' letters, and (crucially) articles that included any of the chosen keywords but whose content was not relate to the Hong Kong protest incidents. Following the manual screening, we retained a total of 4522 articles; 793 articles come from western-based newspapers and 3729 from Hong Kong–based newspapers, with a mean length of 783 tokens.

The *Hong Kong Protest News Dataset* has a 22-year time horizon, spanning from January 1, 1998, to June 30, 2020, to capture changes in how Hong Kong– and western-based news sources cover protests in Hong Kong, and test the relevance and robustness of changes in how newspapers treat protests over time. The extended time horizon together with the size of our sample represent a significant departure from other datasets on Hong Kong protests, which tend to include a much smaller samples of articles, focus on a particular episode of protest, or attempt comparisons between no more than two incidents of protests at different points in time. For instance, Bhatia (2015) uses approximately 100 articles the SCMP published over the last two months of the 2014 Occupy Central protests to understand the SCMP's characterisation of those protests. Yu (2015) uses 249 news stories to examine the frames that the SCMP, the NYT, and *The Guardian* use in their coverage of the 2014 Occupy Central protests. Wong and Liu (2018) examine newspapers' representations of the aggressive behaviour of social actors in the 2014 Occupy Central protests based on 875 articles from the *China Daily* and the SCMP. Du et al. (2018) rely on 191 articles from the FT, the NYT, *Ming Pao*, *People's Daily*, and the *United Daily News* to show how differently these newspapers frame news stories about the 2014 Occupy Central protests. Lee (2014) uses 1,767 articles from the *Apple Daily*, the *Oriental*, and *Ming Pao* to investigate whether news organisations exercise any social control function in their discussion of protests that took place in Hong Kong between 2001 and 2012.

Author	Protest event	No. of articles	News source
Bhatia (2015)	Occupy Central	100	SCMP
Yu (2015)	Occupy Central	249	SCMP, NYT, The Times, The Guardian
Wong and Liu (2018)	Occupy Central	875	China Daily, SCMP
Du et al. (2018)	Occupy Central	191	FT, NYT, Ming Pao, People's Daily, United Daily News
Lee (2014)	HK Protests, 2001–2012	1,767	Apple Daily, The Oriental, Ming Pao

Table 1: Existing Hong Kong protest news datasets.

3. Related Work

Broad-scale research on news coverage like the one presented in this paper remains limited to date. 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. Federico et al. (2000) report on the development and evaluation of an Italian broadcast news corpus at ITC-irst, consisting of 30 hours of recordings transcribed and annotated with conventions like those adopted by the LDC for the DARPA HUB-4 corpora, to reproduce verbal and non-verbal sounds of speech recording, and associate certain signal, speaker, and content conditions with speech and its transcription. Thanks to a dataset of 5000 English news headlines (but not the entire articles) annotated via crowdsourcing, Bostan et al. (2020) further research on emotion analysis by addressing emotions as a phenomenon to be tackled with structured learning. Field et al. (2018) analyse 118,532 articles over a 13-year timeline of the Russian newspaper *Izvestia* to identify government strategies for subtle media manipulation strategies, at the intersection of agenda-setting and framing.

4. Methods

4.1. Topic modelling

We use topic modelling to contrast the treatment of protests in Hong Kong, both across news sources and over time. We use latent Dirichlet allocation (LDA) (Blei et al., 2003) for our topic models, which 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 modelling with MALLET (McCallum, 2002), and to pre-process the articles, we lemmatise all tokens with WordNet's morphy feature (Miller, 1995), and also extract common bigrams. The resulting unigrams and bigrams were then converted to term–document matrices and provided as inputs to MALLET. We created models exploring varying numbers of automatically discovered topics in ranges we set for each subset of articles, and we subsequently evaluated the coherence of resulting topic according to Mimno et al. (2011).

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 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. For the period 1998–2020, we operationalised issue framing by creating models, setting the number of topics from $k = 5$ to 20, and evaluating the coherence of the resultant topics. We found that using six topics produced the highest coherence score, and we identified each of these topics with an identifying label.

4.2. 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. Analysis of variance (ANOVA) is a class of sampling theory–based methods 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. As our corpus displays a non-parametric distribution, we apply Mann–Whitney U, splitting by newspaper source and using the Holm–Bonferroni correction with significance level of $\alpha = 0.01$, to test whether any of the 19 protests-related keywords has statistically significant differences in usage (i.e., *confront*, *confrontation*, *crackdown*, *democracy*, *freedom*, *freedom of speech*, *independence*, *occupation*, *protest*, *protests*, *resistance*, *rights*, *riot*, *rule of law*, *severe*, *tension*, *terrorism*, *terrorist*, *unrest*).

The Mann–Whitney U, splitting by newspaper source and using the Holm–Bonferroni correction, shows that every word has statistically significant differences in usage except severe. The same test, splitting by before and after June 2019, shows statistically significant differences only for five (out of the 19) keywords: *protests*, *unrest*, *rights*, *rule of law*, and *democracy*. For the Friedman's test with four categories – that is "west before June"; "west after June"; "Hong Kong before June"; and "Hong Kong after June" – no keyword showed statistically significant differences.

4.3. Sentiment analysis

We apply computational sentiment analysis to measures the tone and connotations of articles. While it is common to use hand-crafted sentiment (valency) lexica (Mohammad, 2018), we selected a technique that is robust to the specific words that are chosen. We chose 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 $(+, -)$ 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. Sentiment score may provide evidence of stylometric differences between newspapers sources, which (together with the analysis of lexical usage and topic modelling) strengthens the current understanding of newspaper portrayal of civil unrest in Hong Kong.

4.4. Comparing embedding neighbourhoods

The investigation of word embedding neighbourhoods furthers our understanding of *how* words are used differently between Western- and Hong Kong–based newspapers and *how* the contexts of protest-related keywords differ across news sources. Diachronic shifts in word usage are often identified with changes in words' neighbourhoods in an embedding space (Hamilton et al., 2016; Gonen et al., 2020). For instance, Hamilton et al. (2016) used these shifts to find changes 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 both replicate 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 $\mathcal{M}_{\overline{a}}$.[1]
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 neighbourhoods in each split. To extend the work of Gonen et al. (2020), we contextualise the similarity score of a given word against a reference set. We give the percentile in which word w's similarity score falls. We find this distributional measure to be more meaningful than the raw similarity score.

5. Results and Discussion

5.1. Who covers the protests in Hong Kong?

The number of articles about protests in Hong Kong allows us to gauge the news value of protests as well as the thematic relevance to newspapers in general. Between 1998 and 2020, Hong Kong–based newspapers published more articles about protests in Hong Kong than western-based newspapers, except for 2014 when this trend was reversed.

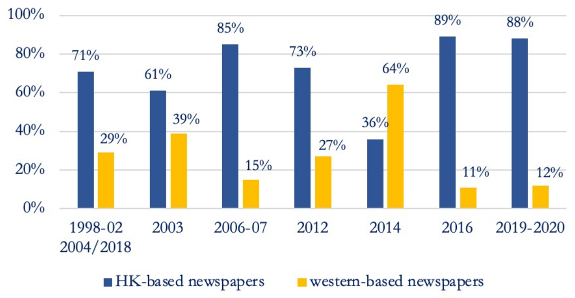


Figure 1: Hong Kong– and western-based newspapers' coverage of selected episodes of civic unrest between 1998 and 2020

Western-based newspapers' coverage of protests in Hong Kong is punctuated by sharps peaks and dips, and declines over time, most significantly between 2014 and 2019–2020. With 425 articles, the NYT published more than half than what all western-based newspapers published on Hong Kong protests over the 22-year timeline of our research. With 3347 articles, the SCMP published almost 8.8 times more articles than the China Daily (i.e., 3347 vs 381), and more than any other newspapers in our sample. Moreover, the SCMP and China Daily together published 4.7 times more articles than the NYT, WSJ, WaPo, FT, the Guardian, the Times of London combined (i.e., 3728 vs 793). As the main English-language outlets in Hong Kong, it is not surprising that the SCMP and China Daily coverage of Hong Kong matters is more frequent and in-depth than that of western-based newspapers, particularly nowadays as newspapers have reduced the space, resources and commitment devoted to a range of topics, and have especially cut back on foreign news.

5.2. What is the tenor of articles about the protests?

We corroborate the findings of more negative tone in Hong Kong–based newspapers between 1998 and 2020. At 36.9% the Hong Kong–based articles' average positivity is slightly lower than the 38.1% of western-based articles. There is, though, wide variation across sources. At 31.3% and 31.5% the SCMP and the Times emerge as the newspapers with the most negative tone overall, even though the SCMP published the most articles, whereas the Times rarely publishes about the protests. Both The Guardian, at 32.9%, and the Financial Times, at 33.9%, also rank low in positivity, which makes UK-based newspapers the more negative about Hong Kong protests among all western-based newspapers. US-based newspapers average a positivity score of 36%, with the NYT articles being almost imperceptibly more positive than the WSJ (i.e., 36.3% vs 36.1%) and WaPo (i.e., 36.3% vs 36%). At 40%, the China Daily has the most positive tone among both Hong-Kong– and western-based newspapers. Finally, Hong Kong–based newspapers articles' average positivity remains lower than that of western-based newspapers articles in 2014 (i.e., 33.2% vs 35.7%), and also in 2019–2019 (i.e., 31.4% vs 32.9%).

5.3. What do headlines hint about the protests?

News headlines are bait. They are meant to catch readers' attention by using narrative mechanisms and sensational or provoking words (Blom and Hansen, 2015), and help the reader get the most out of the news with minimum effort (Dor, 2003). We tested for the presence of long, short, and judgemental headlines vis-à-vis protests in Hong Kong. Sixty-three percent (63%) of articles in the corpus have long headlines (i.e., include six or more words), whereas the remaining 36% have headlines with less than six words, with these trends not varying significant over the extended timeline of the research. With headlines like "The Worst of Times" or "Hong Kong: A City Divided" the NYT emerges as the newspaper with the highest likelihood of having telegraphic headlines (i.e., 6.3 times more likely), whereas with headlines like "Hong Kong Extradition Bill: Business Groups Breathe Collective Sigh of Relief Over Government Decision to Delay Legislation" the SCMP is the least likely of the newspapers to have short headlines (i.e., 11% less likely).[2] The NYT emerges as

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

[2] The full regression model containing all predictors was statistically significant, X2 (8; N = 4522) = 723.787, p ¡ 0.001. The model correctly classifies 75% of cases, and explains between 16.9% (Cox and Snell R2) and 21.5% (Nagelkerke

the one newspaper whose headlines offer a clear and dramatic view of what the article is about it to stimulate its readership's curiosity. SCMP long headlines showcase key information from the articles, and in consistently doing this, the SCMP emerges as the newspapers that more efficiently succeeds at both story summarisation, immediacy satisfaction, and attention direction among all newspapers.

Headlines can be structurally classified as either verbal or nonverbal. Some 75% of the headlines in the NYT articles were nonverbal, while only 25% of them were verbal; most of the nonverbal headlines were modified—i.e., they may include a term that adds descriptive information to the headline (Quirk et al., 2010, 65). Furthermore, on average, about 53% of the headlines of western-based newspapers other than the NYT were also of the nonverbal kind. We also found that, in their headlines, western-based newspapers used "presupposition" (van Dijk, 1995, 273) (Bonyadi and Samuel, 2013, 5) 45% more than Hong Kong–based newspapers to posit a negative attribute for what articles identify as *others* (e.g., Hong Kong Chief Executive; Hong Kong government; Beijing; China; police) and positive ones for *us* (e.g., protesters; citizens; rights; freedoms).

5.4. How do newspapers frame the protests?

Our unsupervised topic modelling reveals that, both over time and in the case of specific protest events (i.e., the anti-ELAB protests) Hong Kong– and western-based newspapers use the same topics. The prominence and timing of how the same topics are used is, however, different. As such, what emerges from the topic modelling analysis should be understood as journalistic frames, unique and specific to western- and Hong Kong–based newspapers' coverage of the Hong Kong protests between 1998 and 2020.[3] The treatment of the police violence topic/frame helps showcase the role and relevance that factors such as norms and practices of the news industry, newspapers' desire to appeal to their own readership, preference for big picture issues, and/or focus on the details of domestic issues play in shaping the narrative of protests coverage.

Police violence Between 1998 and 2018, the trends for how the police violence frame has been used in western- and Hong Kong–based newspapers hardly mimic each other or move in opposite directions, as between 2003 and 2007, when the use of this frame peaks in western-based newspapers and dips for Hong Kong–based ones, or 2008 and 2012, when the opposite happens. The trends, though, mimic each other in Hong Kong– and western-based newspapers between 2019–2020. This points to the police violence frame having equal significance for both sets of newspapers when discussing the protest-related cycle of violence that the Hong Kong government could not break.

Topic	Top 10 words
Chief executive	bill, lam, extradition, public, court, executive, legal, case, cheng
Finance	cent, per_cent, hk, business, company, market, million, property, trade, billion
Mainland	beijing, chinese, country, system, state, mainland, national, law, foreign, central
Police violence	officer, station, violence, force, arrested, yesterday, students, road, university, day
Legislative council (LegCo)	election, party, leung, council, candidate, lawmaker, vote, executive, camp, legislative
Activism and democracy	student, n't, movement, street, Chinese, leader, mr, day, beijing, democracy

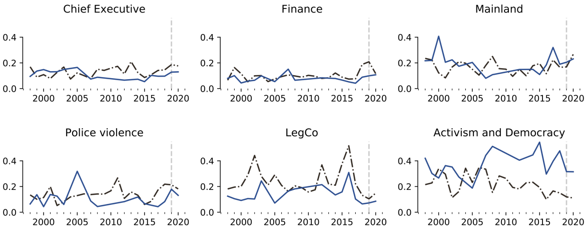


Figure 2: Principal topics/frames used in protest news construction in Hong Kong– and western-based newspapers, 1998–2020. Solid blue: Western. Dashed black: HK.

Finally, as a complement to topic modelling identifying the most relevant news frames, we also used Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015; Tausczik and Pennebaker, 2010) to analyse the language used in the articles. LIWC characterises text based on word counts across more than 70 morphosyntactic and psychometric dimensions. We found only discernible trends in two of these categories: assent and affiliation. Newspaper-ese has some common stylistic elements across the world, and the subject matter we consider shares one focus: the protests in Hong Kong, which is likely to be the reason for the consistency across LIWC categories, and our related findings. The "I" and "SHEHE" 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 negative connotation, and also that the western-based newspapers' use of the phrase *freedom_of_speech* relates more to the principles and values that the expression embodies, whereas the use in Hong Kong's newspapers is more descriptive.

R2) of the cases. The strongest predictor for short headlines is the variable for the NYT, with an Exp(β) of 6.3, whereas the weakest predictor for short headlines is the variable for the SCMP, with an Exp(β) of – 0.89.

[3] TADA 2021, 11th Annual Conf. on New Directions in Analyzing Text as Data. Panel *Longitudinal Studies of Language*; discussant Philip Resnik. `https://tada2021.org`

Hong Kong–based	Western-based
freedom_of_speech	
far, society, deprived, worry, evolving, true, kind, expect, struggle, different	expression, freedom, protecting, exercising, freedoms, protected, legitimate, upholding, liberty, erosion

Table 2: 10 nearest neighbours of *freedom_of_speech*.

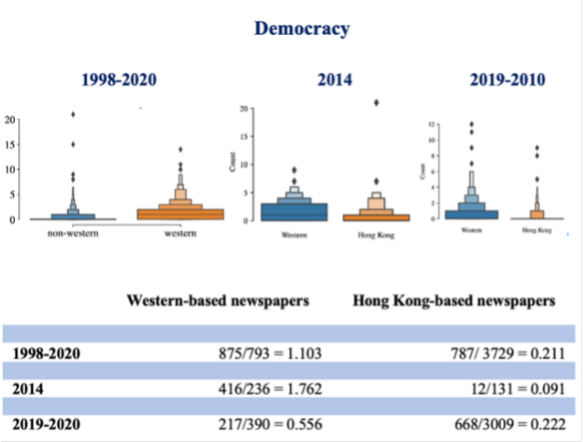
5.5. How often do newspapers use protest-relevant terms?

The investigation of the evolution of how words are used differently, both in the Western/Hong Kong split and over time, reveals that, with the exception of 2019–2020, western-based newspapers have used the terms democracy and freedom more often than Hong Kong–based newspapers.

Between 1998 and 2020, *freedom* appears 492 in western-based newspapers and 70 times in Hong Kong–based newspapers. In the same time period, *democracy* appears 242 times in western-based newspapers and 107 in Hong Kong–based newspapers. In 2014, *freedom* appears 127 times in western-based and only 18 times in Hong Kong–based newspapers, whereas *democracy* appears 34 times more in western-based than Hong Kong–based newspapers (i.e., 416: 12). These trends are reversed between 2019 and 2020, when *freedom* appears 755 times in Hong Kong–based newspapers and 365 in western-based ones. *Democracy* appears 668 times in Hong Kong–based newspapers and 217 times in western-based ones.

However, as shown in Figure 3, the frequencies of use of *democracy* and *freedom* are, overall, lower in Hong Kong than in western-based newspapers. Moreover, western-based newspapers use *democracy* and *freedom* predominantly as a noun, whereas Hong Kong–based newspapers tend to use both terms more often as qualifiers rather than as nouns.

The difference in frequencies may be partially rooted in the type of articles that the newspapers publish. Western-based newspapers tended to cast citizens' civic assertiveness as their fight for democracy and the freedoms that come with it, or resistance against authoritarian tightening that Hong Kong has been experiencing following the 1997 handover. This narrative may require a more frequent use/discussion of democracy and freedom as concepts and values. On the other hand, Hong Kong–based newspapers tend to focus their discourse narrowly on the details of the protests rather than on their meaning. With such specific narrative, it is, perhaps, not surprising that *democracy* and *freedom* are used sparingly and as qualifiers across the large number of articles published.



	Western-based newspapers	Hong Kong-based newspapers
1998-2020	875/793 = 1.103	787/ 3729 = 0.211
2014	416/236 = 1.762	12/131 = 0.091
2019-2020	217/390 = 0.556	668/3009 = 0.222

Figure 3: Lexical frequencies for the term *democracy*

5.6. How are terms used differently between the West and Hong Kong?

The analysis of lexical usage reveals semantic divergence in certain keywords between Western- and Hong Kong–based newspapers between 1998 and 2020, and also with regards to particular episodes of protests. As shown in Table 3, between 1998 and 2020, the most significant semantic divergence (in terms of the immediate neighbourhood) is found in the lexical usage of the words *riot* (98th percentile), *protest* (88th percentile), *occupation* (80th percentile), *confrontation* (70th percentile), *tensions* (59th percentile), and *crackdown* (51st percentile). Moreover, a visual inspection of the term's nearest neighbours for the Western-based model suggests the prevalence of neutral or descriptive lexicon as in the case of *scene*, *clearance*, *dispersal*, *crowds* for the word *riot*; *sit-ins*, *rally*, *campaign* for the word *protest*; or *dispute*, *turmoil*, *uncertainty* for the word *tensions*.

In contrast, the nearest neighbours in the Hong Kong–based model relate to adversarial or hostile behaviour as in the case of *fired*, *barricades*, *pepper*, *teargas* for the word *riot*; *break*, *standoff*, *storm*, *chaotic*, *dislodge* for the word *confrontation*. These trends are evidence of Hong Kong–based newspapers' choice of the protest paradigm when publishing about civic unrest in Hong Kong, and also that the SCMP and China Daily reporting about protests has remained the same, although Hong Kong protests have evolved over time. Moreover, the fact that the most significant semantic divergence is found in in the lexical usage of nouns used in their singular form, suggests that Hong Kong–based newspapers' consistent use of the protest paradigm, to frame the discussion of protests in Hong Kong, could be a strategy used to criticise the values embodied in those nouns while reporting about them more distantly as objects or actions.

Tables 3 and 4 show that also in the case of 2014 the Occupy Central protests and the 2019–2020 anti-ELAB protests the analysis of lexical usage reveals semantic divergence in certain keywords between Western- and

Hong Kong–based	Western-based
riot (98th percentile)	
fired, spray, barricades, officers, pepper, station, rubber, cocktails, firing, teargas	mob, rampage, mobs, siege, scene, clearance, dispersal, crowds, clash, radicals
protest (88th percentile)	
activists, peaceful, rally, mass, organizers, demonstrations, streets, occupied, admiralty, main	demonstration, sit-ins, rally, demonstrations, campaign, rallies, march, movement, sit-in, protests
occupation (80th percentile)	
denounce, supporters, peacefully, confrontation, join, occupying, radical, momentum, peaceful, anti-government	mayhem, rallies, demonstrations, demonstration, marches, sit-ins, outbursts, bloody, 79-day, scenes
confrontation (70th percentile)	
peacefully, break, dramatically, preparing, standoff, storm, driving, demonstrate, chaotic, dislodge	turning, stand-offs, mayhem, resorting, confrontations, extreme, resorted, quickly, disruptive, chaotic

Table 3: Neighbours in 1998–2020 for select protest-related keywords

Hong Kong–based newspapers that are consistent with what found for the 1998–2020 period.

In the case of the 2014 Occupy Central protests, the most significant semantic divergence is found in the lexical usage of *occupation* (75th percentile), *protest* (61st percentile), *confrontation* (53rd percentile), whereas the least significant semantic divergence is found for some of the very words that are used most differently over time (i.e., *tensions* (24th percentile), *riots* (20th percentile), and *crackdown* (18th percentile).

Hong Kong–based	Western-based
occupation (75th percentile)	
join, started, even, protesting, threat, thought, umbrella, planning, probably, revolution	a, court, support, work, admiralty, go, even, legal, protest, they
protest (61st percentile)	
movement, main, pro-democracy, student, peace, group, district, site, admiralty, love	court, admiralty, even, a, pan-democrats, occupation, social, three, ?, way
confrontation (53rd percentile)	
scene, losing, showing, despite, grew, avoid, businesses, families, workers, demonstration	line, lai, came, much, wong, democratic, number, democracy, sit-in, still
tensions (24th percentile)	
became, questions, laws, prevent, little, half, winning, helped, closely, internal	participants, views, rights, meant, yesterday, month, also, city, protesters, students

Table 4: Neighbours in 2014 for select protest-related keywords

As for the 2019–2020 anti-ELAB protests, the most significant semantic divergence is found in the lexical usage of *protests* (66th percentile) and *tensions* (63rd percentile) (Table 5). On the one hand, the consistency of the prevalence of neutral or descriptive lexicon for the Western-based models, and the recurrence of adversarial or hostile behaviour lexicon for the Hong Kong–based model, and the fluctuations in the magnitude of the semantic divergence in certain keywords when comparing across protests, are likely to be linked to the characteristics specific of the various episodes of protests. On the other, they may be explained by Hong Kong media "norms of political correctness" (Lau and To, 2002, 74) vis-à-vis Beijing, or the "strategic rituals" (Lee, 2000, 317) Hong Kong newspapers have established to counter Beijing's "strategic ambiguity" (Cheung, 2003) and ensuing self-censorship, or by cultural co-orientation, resulting from Hong Kong journalists' views shifting closer to China's official views.

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 5: Neighbours in 2019–2020 for select protest-related keywords

5.7. Does coverage differ before and after the onset of protests in June 2019?

We investigated whether there are differences in these differences over time, in the 2019–2020 anti-ELAB protests. We found that, over time, the semantic context of the protest keywords becomes more polarizing and intense. Statistical analysis lets us compare the means of a continuous response variable, modulated by two categorical explanatory variables. We use the Holm–Bonferroni correction to mitigate false discovery. In our case, the explanatory variables are the source (western/HK-based newspapers) and the date: was the article published before or after July 1, 2019?

In the case of *unrest*, *democracy*, *rights*, *crackdown*, and *protest* our analysis found significant differences in the way these terms were used in newspapers before and after July 1, 2019. For the Friedman's test with four categories (western-based newspapers, Hong Kong–based newspapers) x (before, after) no keyword showed significant differences. These results suggest that any already existing biases were not discernibly altered by the onset of the anti-ELAB protests.

Moreover, building on the richness of our dataset, we also sought to quantify the degree to which the introduction of ELAB acted as a pivotal moment in how newspapers portray the Hong Kong protests, 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 split the corpus into "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 the month of June 2019. Neighbourhood shift analysis 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, neighbours 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 neighbouring terms for *riots* changes to include more polarising terms such as *violent*, *escalated*, *destructive*, *triggered*, *anti-government*, and *sparked*. Similarly, pre-July 2019, neighbours for *terrorism* include, among others, terms like *covered*, *lawyers*, and *negative*. Post-June 2019, neighbouring words become politically charged, and include *criminals*, *destructive*, *extreme*, *lawless*, *punishing*, and *barbaric*.

These findings reflect well the extent to which June 2019 was a pivotal moment in the context of the 2019–2020 Hong Kong protests. As protests escalated exponentially during the month of June, feelings of social danger prevailed in newspapers' accounts of the events. Citizens' civic assertiveness was described more and more harshly over time. Our dataset allowed to see that articles pre-July 2019 focused on general descriptions of protesters' tactics, whereas post-June 2019 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.

1998–2020	2014	2019–2020
98th percentile	20th percentile	99th percentile
fired, spray, barricades, officers, pepper, station, rubber, cocktails, firing, teargas	batons, fired, shield, canisters, umbrellas, rubber, disperse, bullets, officers	violent, escalated, destructive, triggered, anti-government, sparked

Table 6: Neighbouring terms for the word *riot* in Hong Kong–based newspapers in 1998–2020, 2014, and 2019–2020

However, 2019 was not the first time that the semantic context of protest keywords had become more polarising and intense. As Table 6 shows, between 1998 and 2020 as well as in 2014, the nature of neighbouring terms for the word riot was the same as that found in Hong Kong–based newspapers in the second half of June 2019. The inefficacy of the Hong Kong government's response to the 2019–2020 protests, and the Hong Kong's shrinking freedom of speech that may help explain why, in 2019, highly polarised protest keywords impacted Hong Kong in such consequential way, whereas their impact was barely noticed in 2014, or between 1998 and 2020.

6. Conclusions

We show how powerful the curated dataset of 4522 articles, spanning over a 22-year time horizon of the *Hong Kong Protest News Dataset* can be in revealing longitudinal and synchronic changes in how Hong Kong– and western-based news sources cover protests in Hong Kong.

The sheer volume of the articles in our dataset validates the news value of Hong Kong protests for both Hong Kong– and western-based newspapers, and shows that coverage of Hong Kong–based newspapers remains consistent and sustained over time, whereas that of western-based newspapers is punctuated by sharp peaks and dips, declining between 2015 and 2020. We speculate that these differences create an opportunity for the Hong Kong–based press to set the agenda for how protests are framed and reported. Within these results, 2014 emerges as an outlier, with western-based newspapers publishing twice as many articles as Hong Kong–based ones on Occupy Central protests.

We prove diachronic consistency between the topics/frames that western- and Hong Kong–based newspapers use to cover the protests in Hong Kong, which points to the generalisability of our findings. We also found that newspapers rely on a limited set of frames, which portray protests as deviant actions characterised by violence and vandalism and detrimental for society.

The analysis of stylometric differences across newspapers sources shows evidence of a more negative tone in Hong Kong–based newspapers when reporting about the protests in Hong Kong, and that western-based newspapers use the terms democracy and freedom more and predominantly as nouns, whereas Hong Kong–based newspapers use them less frequently and, generally, as qualifiers rather than as nouns.

Our investigation of word embedding neighbourhoods broadened the current understanding of how words are used differently between Western- and Hong Kong–based newspapers. We confirmed semantic divergence in certain keywords both between Western- and Hong Kong–based newspapers over time and with vis-à-vis particular episodes of protests. We confirmed that, over time, the semantic context of the protest keywords became more polarising, and that June 2019 is as a pivotal moment, after which the meaning of several keywords shifts for at least the remainder of 2019.

Finally, the extended time horizon of the *Hong Kong Protest News Dataset* also allowed us to capture how the semantics of protest keywords became more polarising and intense during episodes of protest beyond the 2019–2020 anti-ELAB ones. We show that between 1998 and 2020 as well as in 2014, the neighbouring terms for the word *riot* were remarkably similar to those Hong Kong-based newspapers used in 2019. We hypothesise that the differing impact of similarly polarised protest keywords over time can be explained by "shifts in journalistic paradigms" (Chan and Lee, 1984, 97) which altered the boundaries of press freedom in Hong Kong.

7. Acknowledgments

This publication was made possible in part by a grant from the American Political Science Association to A.D.M. and G.M.D.D. The statements made and views expressed are solely the responsibility of the authors.

8. Bibliographical References

Agnone, J. (2007). Amplifying Public Opinion: The Policy Impact of the U.S. Environmental Movement. *Social Forces*, 85(4):1593–1620, 06.

Agresti, A. (2017). *Statistical methods for the social sciences*. Pearson.

Baden, C., Pipal, C., Schoonvelde, M., and van der Velden, M. A. C. G. (2021). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 0(0):1–18.

Bhatia, A. (2015). Construction of discursive illusions in the 'Umbrella Movement'. *Discourse & Society*, 26(4):407–427.

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

Bonyadi, A. and Samuel, M. (2013). Headlines in newspaper editorials: A contrastive study. *SAGE Open*, 3(2):2158244013494863.

Bostan, L. A. M., Kim, E., and Klinger, R. (2020). GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France, May. European Language Resources Association.

Boydstun, A. E., Card, D., Gross, J., Resnick, P., and Smith, N. A. (2014). Tracking the development of media frames within and across policy issues. Unpublished.

Boykoff, J. (2006). Framing dissent: Mass-media coverage of the global justice movement. *New Political Science*, 28(2):201–228.

Büyüköz, B., Hürriyetoğlu, A., and Özgür, A. (2020). Analyzing ELMo and DistilBERT on socio-political news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18, Marseille, France, May. European Language Resources Association (ELRA).

Chan, J. M. and Lee, C.-C. (1984). The journalistic paradigm on civil protests: A case study of Hong Kong. *The news media in national and international conflict*, pages 183–202.

Chan, C. K. (2015). Contested news values and media performance during the umbrella movement. *Chinese Journal of Communication*, 8(4):420–428.

Cheung, A. S. (2003). Hong Kong press coverage of China–Taiwan cross-straits tension. In *Hong Kong in transition*, pages 219–234. Routledge.

Du, Y., Zhu, L., and Yang, F. (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).

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

Federico, M., Giordani, D., and Coletti, P. (2000). Development and evaluation of an Italian broadcast news corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May. European Language Resources Association (ELRA).

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y. (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, October-November. Association for Computational Linguistics.

Gamson, W. A. and Wolfsfeld, G. (1993). Movements and media as interacting systems. *The ANNALS of the American Academy of Political and Social Science*, 528(1):114–125.

Gonen, H., Jawahar, G., Seddah, D., and Goldberg, Y. (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, July. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (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, August. Association for Computational Linguistics.

King, B. G. (2014). The Tactical Disruptiveness of Social Movements: Sources of Market and Mediated Disruption in Corporate Boycotts. *Social Problems*, 58(4):491–517, 07.

Lau, T.-y. and To, Y.-m. (2002). Walking a tight rope: Hong Kong's media facing political and economic challenges since sovereignty transfer. In Ming K. Chan et al., editors, *Crisis and Transformation in China's Hong Kong*, page 322. Hong Kong University Press.

Lee, C.-c. (2000). The paradox of political economy: Media structure, press freedom, and regime change in Hong Kong. *Power, money, and media*, pages 288–336.

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

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

McCarthy, A. D., Scharf, J., and Dore, G. M. D. (2021). A mixed-methods analysis of western and Hong Kong–based reporting on the 2019–2020 protests. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 178–188, Punta Cana, Dominican Republic (online), November. Association for Computational Linguistics.

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

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (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., July. Association for Computational Linguistics.

Mohammad, S. (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, July. Association for Computational Linguistics.

Papanikolaou, K. and Papageorgiou, H. (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, May. European Language Resources Association (ELRA).

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Technical report, University of Texas at Austin.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Quirk, R., Greenbaum, S., Leech, G., and Svatvik, J. (2010). *A comprehensive grammar of the English language*. Longman, London.

Rawnsley, G. D. and Rawnsley, M.-Y. T. (2002). *Political communications in greater China*. Curzon.

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

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Scharf, J., McCarthy, A. D., and Dore, G. M. D. (2021). Characterizing news portrayal of civil unrest in Hong Kong, 1998–2020. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 43–52, Online, August. Association for Computational Linguistics.

Small, M. (1994). Covering dissent: The media and the anti-Vietnam War movement. 70.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (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, October. Association for Computational Linguistics.

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

Tsfati, Y. and Walter, N. (2019). The world of news and politics. *Media Effects: Advances in Theory and Research*.

van Dijk, T. A. (1995). Discourse semantics and ideology. *Discourse & Society*, 6(2):243–289.

Weiss, M. L. and Aspinall, E. (2012). *Student activism in Asia: Between protest and powerlessness*. U of Minnesota Press.

Wendlandt, L., Kummerfeld, J. K., and Mihalcea, R. (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, June. Association for Computational Linguistics.

Wong, H. T. and Liu, S.-D. (2018). Cultural activism during the Hong Kong umbrella movement. *Journal of Creative Communications*, 13(2):157–165.

Wong, M. Y. (2021). Democratization as institutional change: Hong Kong 1992–2015. *Asian Journal of Comparative Politics*, 6(1):92–106.

Yu, M. (2015). Framing Occupy Central: A content analysis of Hong Kong, American and British newspaper coverage. Master's thesis, University of South Florida.