CASSANDRA CASE STUDY

Using NLP to Understand China's Signaling through News Coverage of the Olympics

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

CENTER FOR LANGUAGE
AND SPEECH PROCESSING

apsa
AMERICAN
POLITICAL
SCIENCE
ASSOCIATION

Special Projects Fund

CASSANDRA

CASSANDRA is an academic initiative that brings together computational science and social science to increase visibility, scholarship, and communication between the two fields, while at the same time addressing the challenges of conducting valid, open, and ethical research at the nexus of political and computational sciences.

CASSANDRA is managed through the Center for Language and Speech Processing of the Johns Hopkins University Whiting School of Engineering, and funded by a grant from the APSA Special Project Fund.

Abstract

Content analysis can reveal intangible phenomena through clues that they leave in text. In this case study we employ content analysis of newspaper articles to understand China's efforts to build soft power by hosting the Olympic Games in 2008 and 2022. We construct a manually curated corpus of 4871 articles from seven western and Chinese-based newspapers spanning from 2000 to 2022. We use a mix of qualitative and computational techniques which collectively reveal that although China accrued soft power through hosting the Olympics, it failed to convert it into international primacy. Coverage across news sources is tonally and topically dichotomized within a limited set of frames, suggestive of opposing agendas confirmed qualitatively. Finally, we find evidence for wolf-warrior diplomacy in the increasingly assertive manner of Chinese reporting over time.

1. Introduction

The Olympic Games are inherently newsworthy, and international media coverage of the Olympics is one of the avenues through which host countries are able to engage with the world at large to promote their image, national prestige, and interests. In fact, since the very inception in 1896, the modern Olympic Games "have provided a stage for the symbolic contestation of modernity as well as between different models of development" (Brownell et al., 2011).

This case study contrasts China- and western-based newspapers portrayals of the 2008 Summer and 2022 Winter Beijing Olympic Games to assess the PRC's success in using the Olympics as a conduit for soft power (Nye, 1990). The growth of accessible digital archives affords the opportunity to conduct computational analyses of language to guide, support, and challenge hypotheses about host countries' efforts, their reception and presentation in mainstream printed media sources. Content analysis (Berelson, 1952), in general, is a set of non-invasive techniques for studying communication artifacts such as documents, photographs, and recordings. Computational methods have supercharged content analysis by complementing subject matter expertise with the potential for massive scale (Field et al., 2018; Demszky et al., 046 2019; Lucy et al., 2020). Computational content analysis approaches can be traced at least to the dictionary-based approaches of the General Inquirer (Stone and Hunt, 1963) and of Laffal (1969) at the third-ever COLING. In this work, we perform content analysis to discover how newspapers' coverage captures intangible concepts, with particular focus on soft power. Our research revealed the following about how China- and western-based newspapers portray the Beijing Olympics:

1. **China failed to convert the soft power accrued 058 through hosting the Olympics into international political primacy.**

2. **China- and western-based newspapers portray the Olympics differently in terms of volume of coverage, tone and topic, which is symptomatic of opposing agendas vis-a-vis China' emergence as global power.**

3. **Collocational shifts show that the Chinese press becomes more assertive by using keywords 067 in increasingly politicized contexts over the extended timeline of our research.**

2. The Olympic Games as Soft Power

Soft power is the ability of a country to affect others through its power of attraction and persuasion. Whether a country is attractive is inherently subjective and depends on how the country is perceived by the international community. As such, building and projecting a positive national image internationally, based on the attractiveness of culture, appeal of state's values, and legitimacy of foreign policy (Nye, 1990), is key to soft power.

The Olympic Games as a soft power strategy has a long history. The most notable example is the 1936 Berlin Olympic Games, when Hitler took advantage of the first televised Games in history to show the world a powerful and well-organized Germany (Boykoff, 2016). The 1964 Tokyo Games and the 1988 Seoul Games demonstrated the political reliability of their hosts and showcased high-technology national industries (Dennis et al., 1988). Further, Australia used the 2000 Sydney Olympics to raise its international profile as "being more than merely a good source for raw material" (Palmer and Pomianek, 2007, 9).

In 2008 and 2022, China leveraged the soft power inbuilt in hosting the Olympics to demonstrate its growing importance worldwide. In 2001, Beijing was awarded the 2008 Summer Olympics, which are often referred to as China's coming out party, and through which China showed off its cultural and economic power. Moreover, China's emergence as a country on equal footing with the world's strongest economies also appealed to domestic audiences and bolstered national pride. The 2022 Winter Olympics arrived at the tail end of a decade and a half during which the West had increasingly perceived China's growing economic and political stature as threatening. In response to what felt like an unfair portrayal of China's achievements, China's President Xi deployed a new foreign policy strategy which advocated a more assertive role for China on the world stage that became known as Wolf-warrior Diplomacy.[1] This aggressive style of diplomacy strengthened China's media power to gain discourse power in international politics and denounce the widespread criticism of the Chinese government and its policies in the media.[2] In this context, China used the soft power of the 2022 Winter Olympics to assert and safeguard, rather than promote, its interests worldwide.

---

[1] See Appendix A.

[2] During the 18th National Congress of the Communist Party of China in 2014, Xi Jinping remarked that "We should increase China's soft power, give a good Chinese narrative, and better communicate China's messages to the world."

3. The Beijing Olympic Legacy Dataset

The corpus of articles that informs our research comes from a small-N sample of English language newspapers: The New York Times (NYT), The Wall Street Journal (WSJ), The Financial Times (FT), and Barron's, and three China-based English language newspapers: the China Daily, the Shanghai Daily, and the South China Morning Post (SCMP). We used the East Coast edition of the NYT and WSJ, the UK edition of the FT, and the Beijing edition of 128 the China Daily.[3]

Current limitations of natural language processing (NLP) comparison across languages led us to include only English-language newspapers (Earl et al., 2004; Baden et al., 2021). While multilingual NLP approaches are being proposed, they are typically proven only for languages related to English (Mimno et al., 2009; Reber, 2019; Chan et al., 2020). In fact, Bender (2009) discusses the limitations generally, and Baden et al. (2021) focus on specific issues in multilingual NLP for the social sciences. Moreover, a recent WMT shared task (Barrault et al., 2019) shows why translation as a preprocessing step is sub-optimal (cf. Field et al., 2018), as current translation systems struggle to maintain coherence across long document contexts.

Besides these NLP limitations, other factors validate our methodological choice. The SCMP and China Daily are the printed English-language newspapers with the largest readership in mainland China and Hong Kong. Also, as our research focuses on whether China's hosting the Olympic Games helped further its interests and promote its image, investigating the perception and reception of these efforts in English language newspapers is key to assessing China's soft power success. Finally, our methodological approach is consistent with the work of Bhatia (2015), Yu (2015); Wong and Liu (2018); Du et al. (2018), Lee (2014), McCarthy and Dore (2022), and Scharf et al. (2021), who rely on English language news sources to investigate news portrayals of protests in Hong Kong.

The articles were collected through keyword-based searches in ProQuest Newspapers for western English language newspapers, and Newsbank Access World News Research Collection for China-based English language newspapers. We searched for "Beijing" + "Olympics"; "Beijing" + "Olympic Games"; "Beijing" + "2008" + "Olympics"; "Beijing" + "2022" + Olympics"; "Beijing" + "Olympic" + "legacy"; "Beijing" + "Olympics" + "infrastructure"; "Beijing" + "investments" + "Olympics". To be eligible for collection, articles had to be published between January 1, 2000 and June 30, 2022, be at least 300 words long, and focus on the Beijing Olympic Games. A one-by-one, manual screening process eliminated

---

[3] See Appendix B.

eventual duplicates within each publication, readers' letters, and articles that included any of the chosen keywords but whose content was not pertinent to the scope of our research. Following the manual screening, we retained 3229 articles from western-based newspapers and 1642 from the China-based ones. A total of 4871 articles makes up the novel and highly curated Beijing Olympic Legacy Dataset (BOLD).

4 Methods

In this research, we conceive of interdisciplinarity as the space between computer science and social science. An inquiry must go farther than matching computational techniques to a social science research question (O'Connor et al., 2011); it involves the design of a synergistic methodology that connects the norms and standards for empirical evidence from these two strange bedfellows.

This means partnering computer science's preference for the structured, generalizable, and objective with the unstructured, critical, and contextual that the social sciences champion. This level of interdisciplinarity allows moving beyond individual findings to explanations of their broader importance and contextual understanding. Skepticism can remain toward findings not drawn from the standard practices of one's own field (Armstrong, 1967). To assuage doubts, we leverage *predictive validity* —i.e., expected correspondence between a measure and exogenous events uninvolved in the measurement process -- and *convergent validity* — i.e., correlation with other measures of the same construct" (Quinn et al., 2010; Grimmer and Stewart, 2013).

To meet the challenges inherent in operationalizing our interdisciplinary research, we use a mixed-method approach. As language is at the heart of our research, computational analysis emerges as important for its ability "to analyze signals ranging from simple lexical clues to word clusters to choices of syntactic structure" (Boydstun et al., 2014) as well as its speed, scale, and granularity. In using statistical techniques to analyze text, the case study builds on research of Field et al. (2018) on the use of word embedding similarity, topic models, and dependency parsing to generate clues toward differing portrayals of race and gender in US history textbooks; Field et al. (2018), who relate the content of Russian state-run news articles to the nation's economic performance to push an agenda of distraction; Mosteller and Wallace (1984) and Bergsma et al. (2012) on content analysis and stylometry in consideration of authorship; Jatowt and Duh (2014) and Kulkarni et al. (2015) on the detection of shifts in word meaning, whether between groups or over time, and gradual or at specific change points; Wijaya and Yeniterzi (2011) on whether relevant historical events lead to differences in word usage; and Blei and Lafferty (2007) on tracking shifts in entire topics, rather than in the semantics of individual words.

In terms of qualitative analysis, we employ qualitative content analysis (Fulcher, 2010; Howitt and Cramer, 2007; Braun and Clarke, 2006) and descriptive interference—in the sense of King et al. (1995)—to corroborate the results emerging from the application of computational techniques to BOLD. Our choice to use qualitative analysis to complement quantitative techniques builds on Achen and Snidal (1989)'s recommendation to use historical case studies as a useful complement to statistical research; their plea was strengthened by Verba's work in the early 1990s (Verba et al., 1993, 1995; Verba, 1996), and Tarrow (1995), which openly called for bridging qualitative and quantitative modes of research in social science. In the last decade, many authors have made a quantum leap Levy (2007) in social science methodology by providing a highly structured approach to qualitative analysis (Coppedge, 1999; Gerring, 2004; Lieberman, 2010).[4]

4.1 Topic Modeling

Cohen states that "newspapers may not be successful in telling people what to think, but they are stunningly successful in telling readers what to think about" (Cohen, 1963, 13). Newspapers do so through agenda-setting (McCombs and Shaw, 1972), which steers perceptions of issue importance, and framing (Entman, 2006), which involves selection and guides how issues should be interpreted. The interplay of agenda setting, and framing provides a unique vantage point for readers learn about a given issue. We tested for which frames China- and western-based newspapers used to characterize the Beijing Olympics to investigate how sports mega events can be used as a conduit for soft power.

We build on established scholarship and use topic models to find the frames used in the articles across news sources and over time (Jacobi et al., 2016; Dehler-Holland et al., 2021; Ylä-Anttila et al., 2022), then manually validate these in terms of their semantic, predictive, and convergent validity (Quinn et al., 2010; Grimmer and Stewart, 2013). For our topic models, we use latent Dirichlet allocation (LDA; Blei et al., 2003), a hierarchical admixture model of text, which allows us to capture and convey the prevalence of various topics, so that we can contrast these across news sources, and over time. We perform topic modeling with MALLET (McCallum, 2002), and to pre-process the articles, we lemmatize 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, we subsequently evaluated the coherence score (Mimno et al., 2011) of the resulting topics and manually spot-checked them. We estimate a topic's prevalence in a news source or year by averaging the topic's weight across the articles from that

---

[4] See Appendix C.

source or year. For the period 2000–2022, we operationalized issue framing by creating models, setting the number of topics from 8 to 25, and evaluating the coherence of the resultant topics. We found that using 10 topics produced the highest coherence score, and we identified each of these topics with an identifying label.

4.2 Sentiment Analysis

Sentiment score may provide evidence of stylometric differences between newspapers sources, so 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).[5] The model uses DistilBERT (Sanh et al., 2019) for feature extraction from text. Together with the analysis of lexical usage and topic modeling, sentiment analysis strengthens the understanding of newspapers' portrayal of the Beijing Olympics.

4.3 Lexical Frequency and Collocations

Approaching content analysis through the lens of corpus linguistics, we focused on lexical frequency and collocations. Word frequency exposes discrepancies in word choice and usage. A lack of event-related keyword in contemporaneous articles from different newspapers may signal omission of events in some of them. Our research builds on 26 keywords[6], selected by the consensus of three subject-matter experts who independently created keyword lists and discussed which candidates to retain. Collocations further the understanding of how words are used differently across news sources. A word embedding model is a smoothed model of collocation that employs a vectorial, rather than symbolic, representation of words. It seeks to assign similar vectors 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 thus changes in the word's embedding (Kulkarni et al., 2015).

---

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

[6] In this case study the term keyword is used in the information retrieval rather than the corpus linguistic sense, the latter of which signifies a term that is statistically characteristic in a text. See Biber and Reppen (2015, 90–150).

To assess these semantic and collocational shifts, we both replicate and extend the difference-in-usage model of Gonen et al. (2020).

1. Partition the corpus C into C_a and C_a based on the attribute of interest a.
2. Fit separate word embedding models for each partition: M_a and M_a.
3. Select a keyword w of interest.
4. Obtain the set of nearest neighbors $NN_a(w)$ and $NN_a(w)$ of w according to M_a and M_a.[7]
5. Score the usage-change of w as the size of the intersection, $| NN_a(w) \cap NN_a(w)|$.

After this process, if a word w is used differently based on the presence or absence of the attribute, we expect its score to be low. Words whose usage does not depend on a will have similar neighborhoods in each split. To extend Gonen et al. (2020), we contextualize the similarity score of a given word with the percentile in which the score falls. This distributional measure is more interpretable than the raw similarity score.

---

[7] Following Wendlandt et al. (2018) and Gonen et al. (2020), we use 1000 nearest neighbors.

5 Discussion

5.1 Who Covers the Beijing Olympics?

How China- and western-based newspapers covered the 2008 and 2022 Beijing Olympics is telling of whether and how China leveraged the Olympics as soft power to proactively tell the story of its economic and political rise. As such, the volume and timing of newspapers' coverage of the Games is indicative of the attention they attracted and importance they commanded at home and abroad.

Between 2000 and 2022, China- and western-based newspapers published consistently about the Beijing Olympics, even though western-based newspapers published more than their China-based counterparts. China-based newspapers' coverage suffers peaks and dips over time, whereas Western-based newspapers show a more sustained and voluminous coverage over time. China-based newspapers, however, published more than their western-based counterparts between 2001 and 2003, when the 2008 Summer Games were awarded to Beijing and repeated pledges of improving the country's record on human rights and the rule of law were made; in 2006, when a Winter-Summer Olympics Partnership was set up between Turin, host of the 2006 Winter Games and Beijing; and 2008 when the Summer Olympics were held. These events were unprecedented for a country with China's global standing at the time. We interpret this as China capitalizing on that fact to showcase the rising profile it grew through the Games.

With 2274 and 654 articles, the FT and the China Daily emerge as the newspapers that published the most, whereas the WSJ, and Barrons published the least (i.e., 172; and 14). Overall, the western-based newspapers published 2.4 times more articles than their China-based counterparts (i.e., 3059 vs. 1252). This created the opportunity for western-based newspapers to set the agenda on how to frame the 2008 and 2022 Olympics, and weaken China's efforts to leverage the soft power of the Olympics, especially in light of China's authoritarian system. Fifty-five percent (55%) of all China-based newspapers' articles were published between 2000 and 2011, whereas 63% of western-based newspapers' coverage was published over the last decade. Our findings further show that while coverage of "all things Olympics" decreases both in China- and western-based coverage between 2009 and 2011, the decline is significantly sharper in the China-based newspapers than in the western-based ones. In the case of the 2022 Beijing Olympics, however, in the months following the end of the Games (i.e., March-April 2022), western-based newspapers' coverage declines of about 52% (i.e., 259 vs. 135), whereas that of China-based newspapers increases 2.6 times (i.e., 147 vs 390). Descriptive inference points to the trend-reversal as evidence of wolf-warrior diplomacy in action. Chinese news sources bolstered positive Olympics-related coverage to

11

counter western-based newspapers' characterizations of China as struggling to maintain international legitimacy as exemplified by mass diplomatic boycott of the 2022 Olympics.

5.2 The Tenor of the Coverage of the Beijing Olympics

The western-based newspapers write not only more, but also more negatively about the Beijing Olympics.[8] Between 2000 and 2022, China-based newspapers' average positivity is significantly higher than that of western-based newspapers (62.16% vs. 43.8%). Western-based newspapers adopted a steady negative tone vis-a-vis China being awarded and hosting the 2008 and 2022 Olympics. The WSJ has the most negative tone of all western-based newspapers, followed by the NYT, and the FT. Among the China-based newspapers, the SCMP emerges as the least positive, whereas the China 437 Daily and the SD present the most positive coverage of the 2008 and 2022 Beijing Olympics. Nevertheless, the SCMP remains more positive than any of the western-based newspapers. Over time, the consistent high positivity of the China Daily coverage supported President Xi's call for "a good Chinese narrative."

Western-based newspapers exhibit imperceptible changes in the negative tone of their coverage of the 2008 and 2022 Olympics. While China is no stranger to critical reporting from the West, it has changed its diplomatic style over the past two decades, from keeping a low-profile diplomacy to a wolf-warrior diplomacy to counter what it is perceived as Western-based newspapers' biased reporting style on China's image and interests worldwide. China-based newspapers' more positive coverage of the 2022 Winter Games (than that of the 2008 Summer Olympics) is seen as both an exercise in self-assertion as well as stating to the world that China's aspiration to global primacy would be beneficial for the world at large (Heath et al., 2021).

5.3 The Opposing Agendas of China and the West on the Beijing Olympics

As shown in Figure 1, ten frames characterize the portrayals of the Beijing Olympics between 2000 and 2022. There is a stark dichotomy between the choice of frames through which newspapers shape their narratives on the Olympics. Only five frames are used by both China- and western-based newspapers. We tested the generalizability of our findings against the Policy Frames Codebook of Boydstun et al.[9] We found consistency with six of the 14 categories listed in the Codebook: *Economic, Capacity and Resources,*

---

[8] See Appendix D.

[9] The Codebook, a reference to aid coding in computational content analysis, "provides a general system for categorizing frames across policy issues (...) and contains 14 categories of frame 'dimensions' (plus an 'other' category) that are intended to be applicable to any policy issue and in any communication context" (2014, 4).

External Regulation and Reputation, and the *Cultural and Identity*, which provides convergent validity for the results of our unsupervised topic modeling. The *Hosting and Winter Sports* frames are used across all news sources, even though China-based newspapers employed them significantly more than their western-based counterparts. Qualitative content analysis, applied to a random sample of 30% of the articles, shows that the SCMP, the SD, and the China Daily used these frames to highlight the economic value of bringing the Summer and Winter Games to Beijing, showcase China's culture and values, and promote China's foreign policy. While skeptical and not necessarily positive, western-based papers' use of these frames peaks when iconic sporting venues and high-tech aspects of the Games are revealed.

Figure 1. Prevalence of frames in western- and China-based news coverage of the 2008 and 2022 Beijing Olympics
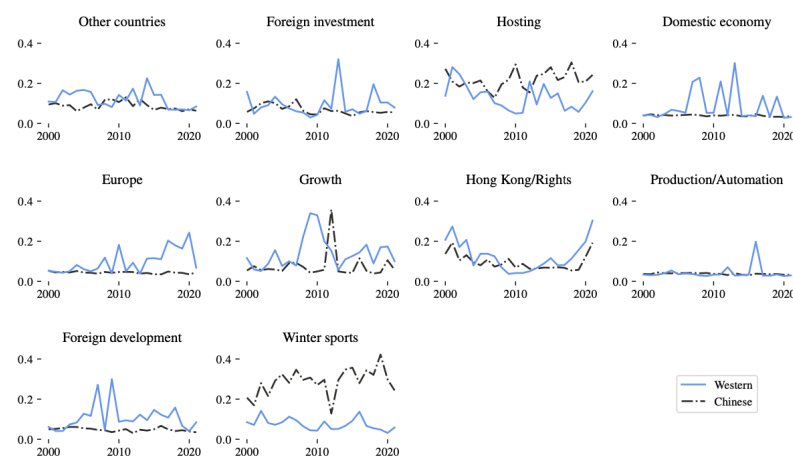
Frame	Top words
Other countries	city, n't, cost, public, brazil, japan, back, ha_been, more_than, business, stadium
Foreign investment	company, market, foreign, technology, business, brand, such_a, group, industry, consumer, service
Hosting	game, olympic, sport, athlete, event, host, bid, committee, winter, team, summer
Domestic economy	asia, growth, asian, market, report, global, cent, region, fund, investor, per_cent
Europe	economic, europe, european, mr, road, eu, trade, london_uk, economy, power, asia
Growth	growth, cent, economy, per_cent, market, price, economic, rate, bank, demand, economist
Hong Kong/Rights	mr., hong, kong, hong_kong, official, state, –, united, president, united_state, foreign
Production/Automation	robot, factory, industrial, manufacturing, job, worker, company, indonesia, japan, make, good
Foreign development	bank, project, development, mr, africa, world_bank, loan, australia, fund, gas, debt
Winter sports	winter, city, ski, daily, percent, news, million, shanghai, skiing, area, winter_sport

The trends for how China- and western-based newspapers employed the Hong Kong frame mimic each other closely. Qualitative content analysis shows that, starting in 2001, China leveraged Hong Kong's worldwide financial relevance to promote China's image and interests in relation to the awarding, preparation, and hosting the 2008 and 2022 Beijing Olympics through a series of carefully choreographed events. Perhaps to be expected, Hong Kong's pro-Beijing Olympics strategy received significant coverage in China-based newspapers. Western-based newspapers' coverage, however, focused on Beijing's authoritarian tightening of Hong Kong's civil and political liberties and touched on the Olympics celebration-related events only in passing. This is iconic of a broader pattern of coverage: using the Olympic Games as the "excuse" to both bring up and negatively discuss "all things China."

Both China- and western-based newspapers used the Foreign Investments and Growth frames, with trends broadly matching, even though western-based newspapers' coverage is more voluminous, which could be seen as a positive development stemming from the soft power of the Olympics. Qualitative

content analysis, however, shows that western-based newspapers do not use these frames to discuss opportunities for economic growth and foreign direct investments that the Beijing Olympics created, rather they use the Olympics simply as the excuse to talk about China's unprecedented economic growth relevance for, and impact on the world economy not in an appreciative manner.

Western-based newspapers almost exclusively use the Domestic Economy, Foreign Development, and Europe frames when publishing on the Beijing Olympics. This finding could be misinterpreted as corroborating China's success in using the Olympics to promote its interests. Qualitative content analysis, however, tells a different story as it reveals that the share of articles that uses these frames increases in coincidence with PRC's announcing economic decisions or launching strategic initiatives (such as the inclusion of provisions relating to property rights in the Constitution in 2005, the launch of the One Belt One Road Initiative in 2013, and the announcement of the Socialism with Chinese Characteristics for the New Era in 2017, but discuss the Games only tangentially. These findings suggest that in hosting the Olympics China failed to reap the gains it hoped for internationally even after the introduction of its overtly assertive wolf-warrior diplomacy.

5.4 Lexicon and Collocations of Newspapers' Coverage of the Beijing Olympics

The following 26 expert-selected keywords guided the investigations of changes in lexical frequency: Olympic games, Olympic, games, winter sports, venues, snow, tourism, Beijing, China, legacy, technology, reforms, foreign policy, power, politics, partnership, green, sustainability, infrastructure, innovation, investments, economic growth, economics, GDP, development, and human rights. 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 26 Beijing Olympics-related keywords has statistically significant differences in usage. Our analysis reveals that out of 26 keywords, the following 3 do not show statistically significant differences in usage: *human rights*, *sustainability*, and *legacy*.

Also of interest what news coverage does not discuss. Several candidate keywords, in fact, had to be excluded because either the China-based or western-based articles, as a group, never use them. We interpret this as the Chinese government's use of strategic ambiguity (Cheung, 2003); that is, warning media about topics whose coverage may not be welcome, without clearly defining which particular aspects of a topic should not be discussed. As such, strategic ambiguity widens the space between what is and is not allowed journalistically while at the 570 same time increasing the space for self-censorship.

The Mann–Whitney U test, splitting by before and after 2011, to check differences in keyword use in the characterization of the 2008 and 2022 Beijing Olympics, shows no statistically significant differences for 10 of the 26 keywords: games, human rights, legacy, green, investments, politics, economic growth, development, GDP, venues, power. The majority of the keywords that lose statistical significance after 2011 broadly relate to finance and 580 the economy. On the one hand, this shift, which also coincides with China surpassing Japan as the world's second largest economy, can be interpreted as China's success in using the soft power of the 2008 Olympics to further its economic interests. On the other, it can be revealing of the wolf-warrior diplomacy in action, with news coverage shifting its focus to the rising political profile of China at a time when the salience of China's economic achievements has been established.[10]

Table 1. Embedding 10 nearest neighbors of Innovation and Power before and after 2012.

2000–2011	2012–2022
Innovation (21st percentile)	
first-rate, commitments, responsible, prestige, communication, communications, conflicts, supervision, oversight, security	cutting-edge, industries, snow-making, upheaval, technologies, tended, materials, technology, shenzhen-based, dji
Power (27th percentile)	
plants, ports, roads, railways, dams, transport, materials, bridges, coal-fired, nuclear	nuclear, neocolonial, waned, military, favours, gambit, plants, coal-fired, influence, embrace

Lexical frequency and usage aid in the discussion of the evolution of how keywords are used differently, both in the China/western split and over time. Innovation is used 474 times across all news sources, with the CD and the FT using it more than any other newspapers overall (i.e., CD =37; FT =350), and also more in relation to the 2022 than the 2008 Olympic Games (i.e., 2000–2011: CD =17 vs FT=89; 2012–2022: CD=30 vs FT=261). This suggests that China used the soft power of the Olympics to showcase its "ability to innovate" (Dychtwald, 2021) both in 2008 and 2022. The lexical usage of this keyword reveals semantic divergence between China- and western-based newspapers over time. As shown in Table 1, the term's nearest neighbors show the prevalence of vocabulary relating to communication or security between 2000 and 2011, but a more clear focus on winter Olympics matters between 2012 and 2022. We interpret these findings as reflective of the Chinese economy shift from replication to innovation (Dychtwald, 2021) that took place in the 2000s, and China's determination to use the 2022 Olympics to showcase its very own take on technological innovation. Qualitative content analysis, however, shows how the West remained doubtful about China's ability to innovate and its consistent negative sentiment about it (Abrami et al., 2014).

---

[10] See Appendix E.

Power is used 2845 between 2000 and 2022. While western-based newspapers use this keyword 9.3 times more than China-based ones, both news sources use power more in relation to the 2022 Winter Olympics than the 2008 Summer Games (i.e., 61.7% vs 37.9%). Between 2000 and 2011, articles show a prevalence of vocabulary relating to 624 infrastructure and energy, whereas, between 2012 and 2022, the context in which the keyword power is most frequently used is political (i.e., neocolonial; waned; military; influence). Qualitative content analysis corroborates these findings in that they are consistent with China's current efforts to shift away from accumulating material power to strengthening non-material power, more specifically its structural (i.e., hard) and discursive (i.e., soft) power capabilities, also embodied in Xi's emphasis on China's wolf-warrior diplomacy (Gustafsson, 2014).

6 Conclusion

In this case study, we combine sentiment analysis, topic modeling and embedding neighborhood shifts with qualitative content analysis to characterize the effectiveness of China hosting the 2008 and 2022 Olympic Games to build soft power. We apply this mixed-method research approach to a corpus of 4871 English language articles from the Chinese and western press over a 22-year timeline.

Topic modeling and sentiment analysis show that presses of China and the West portray the Olympics differently in terms of tone and focus of their articles. This is symptomatic of opposing agendas vis-a-vis China using the Olympics to build soft power, with Chinese coverage focusing on technological advances and details of the Games while the western press coverage uses the Games as an inroad for criticizing China writ large.

Collocations reveal that China fails to attract western praise for its economic and political growth; instead, it is greeted with skepticism and hostility by a world that remains profoundly divided on what to think about China. This resulted in the deployment of a more assertive communication strategy to "tell China's story right", in the words of China's President Xi. The new strategy moves the conversation to the political realm but still fails to sway the west's opinion and garner praise for its accomplishments. Nevertheless, China's assertion of soft power is reflected in "strategic rituals" of journalistic practice obeyed by the Chinese and western press alike. In this way, China's search for soft power is an opportunity lost for the West to talk about key issues of relevance that China must address to become an internationally respected player on the world stage.

References

Regina M Abrami, William C Kirby, and F Warren McFarlan. 2014. Why China can't innovate. Harvard Business Review, 92(3):107–111.

Christopher H. Achen and Duncan Snidal. 1989. Rational deterrence theory and comparative case studies. World Politics, 41(2):143–169.

J. Scott Armstrong. 1967. Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. The American Statistician, 21(5):17– 21.

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

Loïc Barrault, Ondˇrej Bojar, Marta R. Costa-jussà, Chris- tian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 1–61, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? pages 26– 32, Athens, Greece. Association for Computational Linguistics.

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.

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

Douglas Biber and Randi Reppen, editors. 2015. The Cambridge Handbook of English Corpus Linguistics. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of Science. The Annals of Applied Statistics, 1(1):17 – 35.

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

Amber E Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A Smith. 2014. Tracking the development of media frames within and across policy issues. Unpublished.

Jules Boykoff. 2016. Power games: A political history of 726 the Olympics. Verso Books. 727

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative research in Psychology, 3(2):77–101.

Susan Brownell et al. 2011. The Beijing Olympics as a turning point? China's first Olympics in East Asian perspective. The Olympics in East Asia: Nationalism, regionalism, and globalism on the center stage of world sports, page 185.

Chung-Hong Chan, Jing Zeng, Hartmut Wessler, Marcungblut, Kasper Welbers, Joseph W Bajjalieh, Wouter van Atteveldt, and Scott L. Althaus. 2020. Reproducible extraction of cross-lingual topics (rectr). Communication Methods and Measures, 14(4):285–305.

Anne SY Cheung. 2003. Hong Kong press coverage of China–Taiwan cross-straits tension. In Hong Kong in 742 Transition, pages 219–234. Routledge.

Bernard C. Cohen. 1963. Press and Foreign Policy. Princeton University Press.

Michael Coppedge. 1999. Thickening thin concepts and theories: Combining large n and small in comparative politics. Comparative Politics, 31(4):465–476.

Joris Dehler-Holland, Kira Schumacher, and Wolf Fichtner. 2021. Topic modeling uncovers shifts in media framing of the German renewable energy act. Patterns (N Y), 2(1):100169.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

M. Dennis, A. AC01719122], M. Dennis, and B. Szajkowski. 1988. German Democratic Republic: Politics, Economics, and Society. Marxist regimes series. Pinter Publishers.

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).

Sarah A. Soule. 2004. The use of newspaper data in the study of collective action. Annual Review of Sociology, 30(1):65–80.

Robert M. Entman. 2006. Framing: Toward Clarification of a Fractured Paradigm. Journal of Communication, 778 43(4):51–58.

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.

R. Fulcher. 2010. Critical discourse analysis. Longman.

John Gerring. 2004. What is a case study and what is it good for? American Political Science Review, 98(2):341–354.

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.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21(3):267–297.

Karl Gustafsson. 2014. Is China's discursive power increasing? the "power of the past" in Sino-Japanese relations. Asian Perspective, 38(3):411–433.

Timothy R. Heath, Derek Grossman, and Asha Clark. 2021. China's Quest for Global Primacy: An Analysis of Chinese International and Defense Strategies to Outcompete the United States. RAND Corporation, Santa Monica, CA.

Dennis Howitt and Duncan Cramer. 2007. Introduction to research methods in psychology. Pearson Education.

Carina Jacobi, Wouter van Atteveldt, and Kasper Welbers. 2016. Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism, 4(1):89–106.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In IEEE/ACM Joint Conference on Digital Libraries, pages 229–238.

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.

Gary King, Robert O. Keohane, and Sidney Verba. 1995. The importance of research design in political science. American Political Science Review, 89:454–481. A response to five authors in the symposium "The Qualitative-Quantitative Disputation: Gary King, Robert O. Keohane, and Sidney Verba's Designing Social Inquiry: Scientific Inference in Qualitative Research".

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection 840 of linguistic change. In Proceedings of the 24th International Conference on World Wide Web, WWW '15, page 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Julius Laffal. 1969. Total or selected content analysis. In International Conference on Computational Linguistics COLING 1969: Preprint No. 24, Sånga Säby, Sweden.

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

Jack S Levy. 2007. Case studies: Types, designs, and logics of inference. Conflict management and peace science, 25(1):1–18.

Evan S. Lieberman. 2010. Bridging the qualitative-quantitative divide: Best practices in the development of historically oriented replication databases. Annual Review of Political Science, 13(1):37–59.

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 865 machine learning for language toolkit. http://mallet.cs.umass.edu.

Arya D. McCarthy and Giovanna Maria Dora Dore. 2022. 868 Hong Kong: Longitudinal and synchronic characterizations of protest news between 1998 and 2020. In Language Resources and Evaluation.

Maxwell E. McCombs and Donald L. Shaw. 1972. The agenda-setting function of mass media. The Public Opinion Quarterly, 36(2):176–187.

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 882 Linguistics.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In Proceedings of the 2009

Conference on Empirical Methods in Natural Language Processing, pages 880–889, Singapore. 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.

Joseph S. Nye. 1990. Soft power. Foreign Policy, (80):153–171.

Brendan O'Connor, David Bamman, and Noah A Smith. 2011. Computational text analysis for social science: Model complexity and assumptions. In Proc. of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds.

Craig T. Palmer and Christina Nicole Pomianek. 2007. Applying signaling theory to traditional cultural rituals. Human Nature, 18(4):295–312.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. American Journal of Political Science, 54(1):209–228.

Ueli Reber. 2019. Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. Communication Methods and Measures, 13(2):102–125.

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

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.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: Studies using the general inquirer system. In Proceedings of the May 21-23, 1963, Spring Joint Computer Conference, AFIPS '63 (Spring), page 241–256, New York, NY, USA. Association for Computing Machinery.

Sidney Tarrow. 1995. Bridging the quantitative-qualitative divide in political science. American Political Science Review, 89(2):471–474.

Sidney Verba. 1996. The citizen as respondent: Sample surveys and American democracy. American Political Science Review, 90(1):1–7. Presidential Address, American Political Science Association, 1995.

Sidney Verba, Kay Lehman Schlozman, Henry Brady, and Norman H. Nie. 1993. Citizen activity: Who participates? what do they say? The American Political Science Review, 87(2):303–318.

Sidney Verba, Kay Lehman Schlozman, and Henry E. Brady. 1995. Voice and equality: Civic voluntarism in American politics. Harvard University Press.

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

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural DiversiTy on the Social Web, DETECT '11, page 35–40, New York, NY, USA. Association for Computing Machinery.

Hio Tong Wong and Shih-Diing Liu. 2018. Cultural activism during the Hong Kong umbrella movement. Journal of Creative Communications, 13(2):157–165.

Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. 2022. Topic modeling for frame analysis: A study of media debates on climate change in India and USA. Global Media and Communication, 18(1):91–112.

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

Appendixes

A. China Diplomacy: From Low Profile to Wolf-warrior

Wolf-warrior Diplomacy describes an aggressive style of coercive diplomacy adopted by Chinese diplomats in the 21st century under President Xi Jinping's administration, and is in contrast to prior Chinese diplomatic practices of under Deng Xiaoping, which had emphasized the need to "observe calmly, secure our position, cope with affairs calmly, hide our capacities and bide our time, be good at maintaining a low profile, and never claim leadership." Deng's strategy to avoid controversy and take advantage of using a cooperative rhetoric became known as low profile diplomacy.

The term *Wolf-warrior Diplomacy* originates from the title of a patriotic Chinese action film series. At the end of the first Wolf Warrior movie a cover of the Chinese passport is displayed along with some text, which reads "Citizens of the PRC: When you encounter danger in a foreign land, do not give up! Please remember, at your back stands a strong motherland." In line with the action movie series, Wolf-warrior diplomacy is confrontational and combative, with its proponents loudly denouncing any perceived criticism of the Chinese government and its policies on social media and in interviews.

As an attempt to gain discourse power in international politics, Wolf-warrior diplomacy is an integral component of President Xi's new foreign policy strategy of Major Country Diplomacy, which advocate and legitimizes a more assertive and active role for China on the world stage, including engaging in an open ideological struggle with the West. This new foreign policy strategy become emerging around 2010, when China's GDP overtook Japan's as the world's second largest, and became more confident and assertive, gradually departing from Deng Xiaoping's low profile diplomacy dictum. Wolf-warrior diplomacy is evidenced not only in combative words, but also aggressive actions intended to counter the growing perception of China's economic and political development as a threat to other countries' national interests, and address the highly ideologically biased western media portrayals of China. Wolf-warrior diplomacy is part of the Chinese government's endeavor to "tell the China story."

B. Media Landscape

The South China Morning Post founded in 1903 by an Australia-born Chinese a British journalist with funding from mostly non- 1033 Chinese businesspeople, with the manifest goal of supporting the reform movement in China, where revolutionaries sought to overthrow the imperial Qing dynasty. The

newspaper has long been the broadsheet of the city's elites, and it is arguably the city's most important title internationally, a position gained from a combination of both its size and its ownership.

Through the late 1980s and into the mid-2000s, the paper was owned by the media tycoon Rupert Murdoch and then the Malaysian billionaire Robert Kuok. In 2015, the SMCP was acquired by Alibaba with the "uncertain future for traditional publishing" as a key reason behind the sale. As Hong Kong's leading English-language newspaper, the South China Morning Post reports on issues and topics that are considered sensitive in mainland China, where the websites of several international media are blocked. While Ma is known to be politically well-connected, the shift in ownership is not as drastic as some people have indicated. Previous owners, in fact, were business tycoon with close ties to the Chinese government.

The SCMP is not as well read as the international outlets that it would like to compete with, but because of its unique position—as the main English-language outlet in a strategically important city—its coverage plays an outsize role in shaping international understanding of events not just in Hong Kong but across the border in China, as well. Moreover, its coverage is far more credible than any mainland outlet, and has been courting a global readership hungry for news from China by dropping its paywall. In 2018, it announced a tie-up with Politico signaling the newspaper's "growing credibility and authority." (Gary Liu, SCMP's CEO, internal email). Thanks to those factors, as well as drastically increased interest in China, where, of course, the coronavirus pandemic began, the SCMP has seen a sharp rise in readership. Though its daily print circulation is relatively limited, at just over 100,000, it averages more than 50 million monthly active users—a tenfold increase over the past three years—and nearly 200 million pageviews a month.

The China Daily is part of the China 1078 Daily Group, which runs 16 print publications in China and abroad, and according to its mission statement, 'is an authoritative provider of information, analysis, comment and entertainment for global readers with a special focus on China.' The China Daily is the largest English-language daily in China, and although it is state-owned, it is not officially a mouthpiece of the ruling Communist Party and is considered more liberal than the other Chinese state newspapers that circulate in Hong Kong. The China Daily Group also publishes the Hong Kong, US, European, African, Asian and Latin American editions of China Daily, with, according to its own statistics, a total circulation of 900,000 copies. China Daily in mainland China is published on a daily basis; there are two different forms of appearance outside the mainland. There is a China Daily Hong Kong edition and a China Daily USA edition, both with daily frequency, and there are weekly editions for Asia, Europe, North America, Africa and Latin America. The newspaper also publishes China Watch, which is circulated as a supplement with the Washington Post, Los Angeles Times and London's Daily Telegraph.

The North American and British Media Landscapes have been characterized by the centrality of large-scale cultural industries since the development of the penny press in the 1830s. For several decades in the mid-twentieth century, an equilibrium existed in the media system, with stable markets that made the dominant media companies highly profitable and very influential as social institutions. Newspapers invested heavily in newsrooms, and the journalist profession grew in autonomy and influence. Journalism was characterized by a low level of political parallelism, with the norm of objectivity dominating journalistic ethics, and most news organizations avoiding identification with political parties. Over time, economic, technological, and political change began disrupting those very elements that guaranteed the stability and fairness of the press system in the late twentieth century. Stable boundaries that once separated markets have been disrupted by digital convergence and deregulation. Changing business models have impacted newsrooms' sizes, and the Internet has accelerated the fragmentation of broadcast industries. Political parallelism has increased, and newspapers have become more inclined to adopt partisan identities, with little concern about the reporting pitfalls of adopting such clear-cut political orientation, including declining public trust.

C Qualitative Analysis

Qualitative Content Analysis follows four stages: (i) decontextualization; (ii) recontextualization; (iii) categorization; and (iv) compilation.

Decontextualization requires familiarization with the data, read through the text to get a sense of the content, and to learn "what is going on" before the content can be broken down into smaller meaning units to try answering the question at the heart of the inquiry (Catanzaro, 1988; Graneheim & Lundman, 2004). Each identified meaning unit is labeled with a code, understood in relation to the context. Codes can be generated inductively or deductively, depending on the study design.

Recontextualization. After the meaning units have been identified, it is important to check whether all aspects of the content have been covered in relation to question at the heart of inquiry (Burnard, 1991). The original text is re-read alongside the final list of meaning units.

Categorization. Before beginning to create categories, extended meaning units must be condensed. This entails that the number of words is reduced without losing content of the unit (Graneheim & Lundman, 2004). The depth of the meaning units determines the level at which the analysis can be performed. To extract the sense of the data, the coded material can be divided into domains: broad groups based on different attentions of the study (Catanzaro, 1988; Patton, 2002), generally called themes

and/or categories, and in sub-categories and sub-themes. There are no informal ways to describe specific strategies when categorization is performed (Burnard, 1991).

Compilation. Once the categories are established, the analysis and writing up process begins. As a final check, the researcher must consider how the new findings correspond to the literature and whether or not the result is reasonable and logical.

In the context of qualitative content analysis, validity means that the results truthfully reflect the phenomena studied, and reliability requires that the same results would be obtained if the study were replicated (Graneheim & Lundman, 2004; Morse & Richards, 2002; Burnard, 1991).

Qualitative Discourse Analysis. Discourse analysis is intended to uncover "assumptions that are hidden in the worlds of written text or oral speech" (McGregor, 2010:4). It aims to explore often opaque relationships between discursive practices, texts, and events and wider social and cultural structures, relations, and processes. Discourse analysis tries to unite and determine the relationship between three levels of analysis: (i) the actual text; (ii) discursive practices; and (iii) the larger social context that bears upon the text and the discursive practices. According to Van Dijk (2006:259), the theory and practice of discourse analysis look to establish connections between properties of texts, futures of discourse practice (text production, consumption and distribution), and wider sociocultural practice (Fairclough 1995:87). The purpose is to analyze "opaque" as well as transparent structural relationships of dominance, discrimination, power and control as manifested in language (Blommaert & Bulcaen 2000: 448). Since discourse analysis is an interpretive and deconstructing reading, there are no specific guidelines to follow. However, one can make use of the theories of Jacques Derrida (deconstruction), Michael Foucault (genealogy and social criticism and analysis), Julia Kristeva (feminist interpretations of current social practices), or Fredric Jameson (post modernism). McGregor (2010:5) argues that discourse analysis does not have a unitary theoretical framework or methodology because it is best viewed as a shared perspective encompassing a range of approaches framing the details into a coherent whole. McGregor (2010:5) mentions the following techniques of discourse analysis that can be used to facilitate sentences, phrases and words:

- *Topicalisation*, that is, in choosing what to put in the topic position, the writer creates a perspective or slant that influences the reader's perception.
- Sentences can also convey information about power relations.
- *Nominalisation*, that is, converting a verb into a noun.
- *Presupposition* can also occur at the sentence level in the form of persuasive rhetoric that can be used to convey the impression that what an agent of power says carries more weight.

- *Insinuations*, that is, when the facts, or the way the facts are presented, are challenged, the originator of the discourse can readily deny any culpability.
- *Connotations*, that is, even one word can convey strong meaning.
- *Modality*, that is, the tone of the text is set with the use of specific words to convey the degree of certainty and authority.
- *Phony register*, that is, writers can deceive readers by affecting a phony register, one that induces mistrust and skepticism.

D Sentiment Analysis

Table 2. Average positive sentiment by both year and newspaper, with marginals.

source year	Barrons	CD	FT	NYT	SCMP	SD	WSJ	All
2000	–	77.8	52.3	45.5	52.4	–	59.4	56.5
2001	–	66.2	57.4	48.5	56.4	–	32.0	53.2
2002	–	73.4	53.2	52.2	56.9	–	44.0	59.0
2003	–	75.0	48.5	51.3	51.0	–	42.4	53.7
2004	–	78.1	53.8	56.7	60.6	–	51.7	56.6
2005	–	72.3	54.0	48.8	34.5	–	40.3	55.5
2006	–	67.1	42.8	47.9	49.1	61.6	50.4	53.7
2007	–	75.1	52.0	44.1	65.4	60.5	57.6	53.6
2008	48.6	65.8	44.5	51.0	59.3	60.6	48.8	56.2
2009	–	62.8	45.2	46.3	33.0	40.7	57.5	46.1
2010	–	78.5	43.3	39.5	55.3	64.9	38.7	45.8
2011	56.3	–	56.7	41.7	–	47.7	50.6	55.7
2012	49.4	39.9	51.4	45.1	44.0	58.0	46.9	50.8
2013	–	68.3	53.1	45.1	47.1	69.1	48.4	55.1
2014	–	70.8	46.8	37.3	59.2	64.0	48.9	55.3
2015	34.1	81.4	47.3	44.8	51.7	61.6	49.1	50.1
2016	–	63.7	44.1	48.3	62.1	73.1	45.4	46.9
2017	33.0	78.1	48.9	46.9	65.6	73.6	39.7	52.8
2018	–	82.1	40.1	50.4	65.9	74.8	42.4	48.7
2019	47.2	84.6	46.7	48.0	58.6	80.5	50.2	53.5
2020	–	69.2	42.8	46.8	53.9	65.5	28.4	47.1
2021	26.0	71.9	41.0	40.9	45.6	65.4	38.3	47.6
All	37.2	70.3	47.0	45.9	53.2	64.5	45.1	51.9

A history is given in Figure 3, prepared and provided by Bert Hoffman; reproduced with permission.

70 years of economic development and policy in the People's Republic of China

Sources: Hofman, Bert (2018). Reflections on 40 years of China's reforms. In R. Garnaut, L. Song, & C. Fang (Eds.), *China's 40 years of reforms and development* (pp. 53-66). ANU Press. Bottelier, P. (2018). *Economic Policy Making in China (1949-2016). The role of Economists.* Oxon, UK and New York, NY.: Routledge. Penn World Tables Version 9.0; World Development Indicators. Authors' calculations.