Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

China in the eyes of news media: a case study under COVID-19 epidemic

Hong HUANG[1,2,3,4], Zhexue CHEN[1,2,3,4], Xuanhua SHI[‡1,2,3,4], Chenxu WANG[1,2,3,4], Zepeng HE[1,2,3,4], Hai JIN[1,2,3,4], Mingxin ZHANG[5], Zongya LI[5]

[1]*National Engineering Research Center for Big Data Technology and System, Huazhong University of Science and Technology, Wuhan 430074, China*

[2]*Services Computing Technology and System Lab, Huazhong University of Science and Technology, Wuhan 430074, China*

[3]*Cluster and Grid Computing Lab, Huazhong University of Science and Technology, Wuhan 430074, China*

[4]*School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*

[5]*School of Journalism and Information Communication, Huazhong University of Science and Technology, Wuhan 430074, China*

E-mail: {honghuang, chenzhexue, xhshi, wangchenxu, hezepeng, hjin, mingxinzhang}@hust.edu.cn; lzy901014@sina.com

Received Dec. 10, 2020; Revision accepted Feb. 8, 2021; Crosschecked Mar. 31, 2021; Published online May 26, 2021

Abstract: As one of the early COVID-19 epidemic outbreak areas, China attracted the global news media's attention at the beginning of 2020. During the epidemic period, Chinese people united and actively fought against the epidemic. However, in the eyes of the international public, the situation reported about China is not optimistic. To better understand how the international public portrays China, especially during the epidemic, we present a case study with big data technology. We aim to answer three questions: (1) What has the international media focused on during the COVID-19 epidemic period? (2) What is the media's tone when they report China? (3) What is the media's attitude when talking about China? In detail, we crawled more than 280 000 pieces of news from 57 mainstream media agencies in 22 countries and made some interesting observations. For example, international media paid more attention to Chinese livelihood during the COVID-19 epidemic period. In March and April, "progress of Chinese vaccines," "specific drugs and treatments," and "virus outbreak in U.S." became the media's most common topics. In terms of news attitude, Cuba, Malaysia, and Venezuela had a positive attitude toward China, while France, Canada, and the United Kingdom had a negative attitude. Our study can help understand China's image in the eyes of the international media and provide a sound basis for image analysis.

Key words: Country image; COVID-19 epidemic; Topic mining; Entity; Tone of news; Emotion
https://doi.org/10.1631/FITEE.2000689 **CLC number:** TP311.13

1 Introduction

The news is influenced by various social factors. It is difficult to truly reflect the whole picture of an event due to the journalists' and editors' participation at all levels. Every country, every media, and even every person would have its own tone, emotion, and focus toward China, which affects China's image in various ways. The traditional analysis of national image often uses surface analysis based on corpus and media content (Manheim and Albritton, 1984). According to the traditional analysis, China's image has been portrayed as a peace-loving country (Wang, 2003; Zhang L and Wu, 2017), a developing country, and an anti-hegemonic nation (Wang, 2003) by Chinese media. China's image portrayed by overseas media has usually been

‡ Corresponding author
ORCID: Hong HUANG, https://orcid.org/0000-0002-5282-551X; Xuanhua SHI, https://orcid.org/000-0001-8451-8656

mixed and conflicted (Zhang L and Wu, 2017), and portrayed as a socialist country, a significant power, an authoritarian state, and a militant obstructive force (Wang, 2003; Zhang L and Wu, 2017). Traditional methods lack scalability, and the analysis of granularity is relatively simple. Chen et al. (2021) investigated China's image during the COVID-19 epidemic period from a sentiment analysis perspective, and analyzed only aspects of sentiment on Twitter data from the open public. A more intuitive question is, what is China's image at multiple levels and portrayed by overseas mainstream media in a specific period? How do they report China? How do other factors influence their reporting over time? To our knowledge, none of these issues has been thoroughly examined.

In this study, we look at China during the COVID-19 epidemic period as an example to see how overseas media portray China. As one of the early COVID-9 epidemic outbreak areas, China attracted the global media's attention early in 2020. At the beginning of the epidemic, China and the Chinese people suffered from much criticism. What was China's image during this epidemic period? How do international news media portray China over time? Specifically, we aim to answer the following questions:

Q1: What has the international media focused on during the COVID-19 epidemic period?

Q2: What is the media's tone when they report China?

Q3: What is the media's emotional tone when they are talking about China?

To answer these questions, we first designed several crawlers to crawl news articles from overseas news media. We found more than 280 000 pieces of news from 57 mainstream media agencies in 22 countries. After data cleaning and annotation, we generated a highly qualified dataset for news analysis and natural language processing (NLP) related tasks. Then we explored the multi-level media focus problem from three levels: entity level, coarse-grained topic level, and fine-grained topic level. An entity is a real-world object, such as a person or an organization. At the entity level, we concentrated mainly on the entities on which the media focused in their news reports. At the coarse-grained topic level, all news articles were classified into topic categories like society or politics to determine the types of news reported in different media. In contrast, at the fine-grained topic level, each news article was further examined to determine a more concrete topic, such as "Wuhan is under lockdown" and "the progress of the vaccine." As for the tone of the news against China, we studied the tone of the news in different countries, on different topics, and at different periods. For the third question, we designed two methods for determining news emotions toward China. One is to use sentiment intensity to quantitatively measure the media's influence toward China; the other is to use emotional labels to examine the emotional situation qualitatively.

We also made some interesting observations. For example, international media paid more attention to Chinese livelihood during the COVID-19 epidemic period, and most media presented a negative tone against China, such as American and French media. Our contributions are as follows:

1. As far as we know, we are the first to study the image of China as a country in the eyes of overseas news media with a large-scale, multi-level study, especially during the COVID-19 epidemic period.

2. We built a high-quality dataset for country image study and NLP tasks, including crawling a large amount of news media data and annotating parts of the data with crowdsourcing techniques.

3. We made some interesting discoveries. For example, from the analysis of topic distribution over time, we saw that in February and March COVID-19 was the most serious in China. In March and April, "progress of Chinese vaccines," "specific drugs and treatments," and "virus outbreak in U.S." became the media's most reported topics. In terms of news emotion toward China, Cuba, Malaysia, and Venezuela had a positive attitude, while France, Canada, and the United Kingdom had a negative attitude.

2 Dataset

The dataset was crawled from 57 media outlets in 22 countries between December 1, 2019 and June 30, 2020. For case study purposes, we focused mainly on news related to COVID-19. Information, including news titles, authors, and content, was collected. Below we will describe the details of data collection, cleaning, translation, and annotation.

1. Data collection

We studied the mainstream news media (official

news media or media with the most massive audience) in 22 countries including some powers and countries within the Belt and Road Initiative (BRI). We selected 57 official and influential news media sites from the Chinese Ministry of Foreign Affairs and other authoritative websites (www.fmprc.gov.cn/web/gjhdq_676201/gj_676203/yz_676205/, www.fec.mofcom.gov.cn/article/gbdqzn/index.shtml) as our data sources (see Table A1 in the appendix). We designed crawlers for each news medium and collected news that contained keywords related to COVID-19 (e.g., 2019-nCOV, COVID-19, coronavirus, pneumonia) in the context or in the title between December 1, 2019 and June 30, 2020.

2. Data cleaning

For collected data, we deleted duplicated items that were crawled twice or more. In addition, we calculated each news text's similarity score and deleted news that was almost the same. Furthermore, we used regular expressions to fix and replace faulty fields in the dataset, e.g., replacing some of the header fields automatically generated by the website or some label fields in the body.

3. Translation

To ensure the authenticity and integrity of the news data, we collected the original news from the site, which means that the data is available in multiple languages. To facilitate the model for processing the data, we used an online translation application programming interface (www.fanyi-api.baidu.com) to translate all the data into English. We did some processing in the translation process to preserve the integrity of the context information and the sentences and paragraphs after the translation of long text.

4. Annotation

We used crowdsourcing technology to label some randomly selected data from our dataset for supervised learning of the analysis model. We developed a system for multi-person collaborative labeling (https://203.195.140.107:8088). Then we trained 100 labeling experts to annotate the data. Each news item was randomly assigned to at least five experts, who randomly read each news item and marked the news with emotion tags, tone tags, news object tags, topic tags, and news genre tags. The detailed annotated labels are shown in Table 1. After the first annotation, the data with disputed results were annotated a second time by more experts. We ended up annotating about 4115 pieces of high-quality news data.

Table 1 Annotation fields and labels

Field	Label
Emotion	Objective, agreeable, believable, good, hated, sad, worried
Tone against China	Positive, neutral, negative
Subject	About China, not about China
Topic	Politics, society, technology, economy, sports, humanity, entertainment
Type	Fact, interview, essay, remark

5. Data statistics

The processed statistics of our dataset are shown in Table A1 in the appendix. We have made the dataset public and available at http://203.195.140.107/dataset/download. We also did some preliminary studies on the dataset. The distribution of news media source is shown in Fig. 1, with some of the smaller sources combined. We can see that media in different countries paid extra attention to China during the epidemic, and the United States ranks first in the list.

For further research, we separately analyzed the news related to China and COVID-19. The number of news item over time is shown in Fig. 2. There are several peaks in the figure. The first peak appeared on January 23 when Wuhan was locked down, which attracted a large amount of attention worldwide. The second peak was around March 16, when the number of COVID-19 caused deaths outside China surpassed that of China for the first time. We define this time point as the second wave of the COVID-19 epidemic period. The third peak was around May 28, when the World Health Organization (WHO) announced the launch of the "COVID-19 Technology

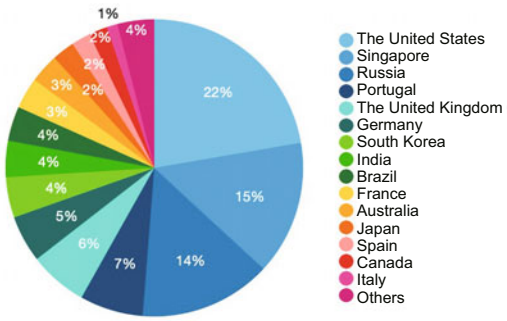

Fig. 1 Media source distribution

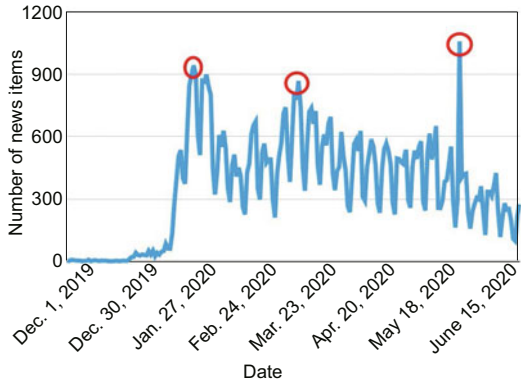

Fig. 2 Number of news items over time

Access Pool" to accelerate the development of vaccines, tests, treatments, and other technologies for COVID-19 through open scientific research.

3 Multi-level media focus (Q1)

In this section, we explore the media focus on China during the COVID-19 epidemic period. Specifically, we conduct our analysis from three levels: entity level, coarse-grained topic level, and fine-grained topic level. At the entity level, we explore the named entities on which these stories are focused. An entity is a real-world object, such as a person or an organization. We extract and analyze the first few entities of most media interest in each category. At the coarse-grained topic level, we analyze the categories to which news items belong thematically, such as social and political ones. In this way, we determine the types of topics related to China to which the media pays more attention. At the fine-grained topic level, we further analyze specific topics of media interest, such as major events or topical trends.

3.1 Entity-level media focus

Entities in the news corpus represent essential elements, including people, organizations, places, and things. We identify entities from these news corpora using the named entity recognition (NER) method. To better understand these news corpora, we extract entities from the news using an NLP tool named spaCy (Honnibal and Montani, 2017). We focus on 10 types of entities, which are listed in Table 2. After obtaining entities using spaCy, we align them with Wikidata (www.en.wikipedia.org/wiki/Wikidata).

Table 2 Entity type description

Type	Description
Person	People, including fictional
NORP	Nationalities, religious, political groups
FAC	Buildings, airports, highways, bridges
ORG	Companies, agencies, institutions
GPE	Countries, cities, states
LOC	Non-GPE locations, mountain ranges, bodies of water
Product	Objects, vehicles, foods (not services)
Event	Named hurricanes, battles, wars, sports events
Work_of_art	Titles of books, songs
Law	Named documents made into laws

Specifically, each entity has a unique identifier called the QID in Wikidata. For example, the QIDs of "U.S." and "the United States" are both Q30, which means that "U.S." and "the United States" are the same entity. In this way, we align entities without disambiguation.

We further study the extracted entities and find that China, Wuhan, the United States, and WHO appeared with a high frequency. The result is highly associated with the epidemic. We list the top five entities in each category as shown in Fig. 3. We can see that during the second wave of the COVID-19 epidemic period, mainstream media care more about medical scientists, such as Anthony Fauci (an American physician and immunologist) and Zhong Nanshan (a Chinese medical scientist), and events that were closely influenced by this epidemic, for example, postponing the Tokyo Olympics because of the epidemic.

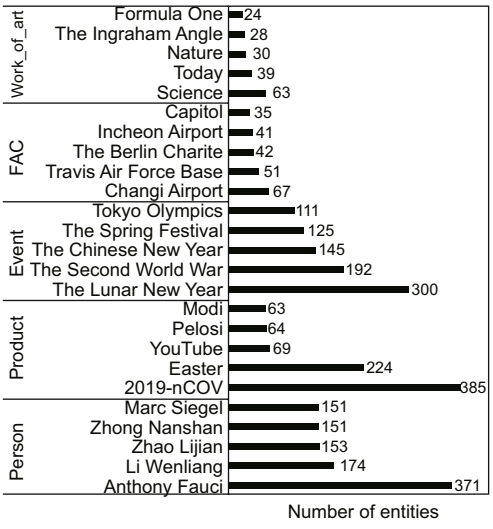

Fig. 3 Top five entities in multiple categories

3.2 Coarse-grained topic level media focus

After examining popular news sites like the BBC and CNN, we set seven topic categories for our study: society, politics, economy, technology, sports, humanity, and entertainment. As introduced in Section 2, we have manually annotated some news articles with these categories' labels. We consider these news articles as training datasets. Then we extract features using term frequency and inverse document frequency (TF-IDF) (Jurafsky and Martin, 2000), build a supervised convolutional neural network (Kim, 2014) for training, and predict the topic labels for the remaining non-annotated news articles.

The coarse-grained topic distribution is shown in Fig. 4. We can see that the media pay special attention to people's livelihood and society issues, followed by politics, economy, and technology topics. The attention to livelihood and society accounts for more than 45%, and entertainment and humanity news together account for less than 1%. During the epidemic, the government's priority is to ensure people's livelihood, accompanied by the promulgation and implementation of a series of political regulations and economic impacts.

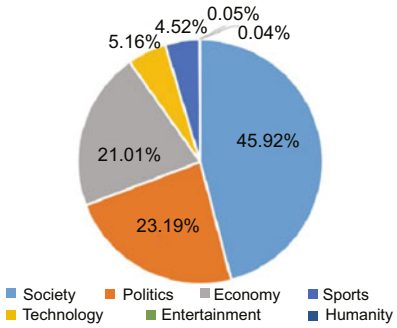

Fig. 4 Topic distribution

Fig. 5 shows the coarse-grained topic distribution in each country, where the horizontal axis represents the country and the vertical axis represents the percentage of news on a particular topic in the country. We can see that different countries show different interest in China. For example, the media in Singapore and Malaysia are more concerned about China's economy. Canada and France are more concerned about Chinese politics. It is worth noting that the proportion of people's livelihood topics reported by the media in Cuba is as high as 62.75%, while that

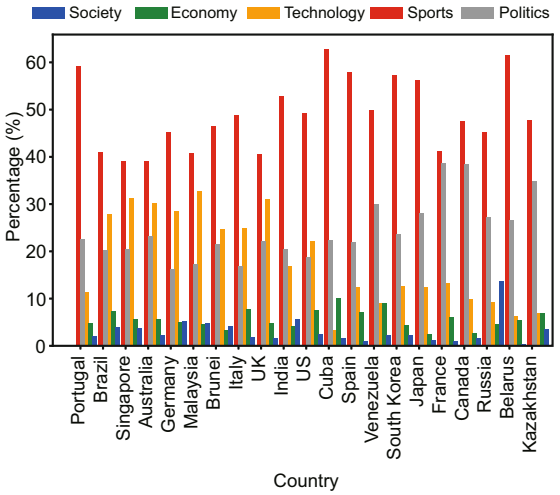

Fig. 5 Topic distribution in each country

in the economic field is only 3.24%; on the contrary, Singapore shows great interest in economic issues, accounting for 31.19%.

We show the topic distribution over time in Fig. 6. The horizontal axis represents the month and the vertical axis represents the percentage of all news on a given topic in a given month. We focus on topic distribution in different periods. In the social news about China during the epidemic, the number of news articles increases and then declines with time. The proportion in February and March reaches a peak, and then the proportion gradually decreases in the next few months. In February and March, COVID-19 was the most serious in China. Much news followed the social topics, focusing on the impact of the epidemic on people's lives and societies. After April, as the epidemic began to be effectively controlled in China, the proportion of news reports continued to decline.

In terms of technology, the proportion of news reports about China has generally risen over time. The proportion in the first three months increased

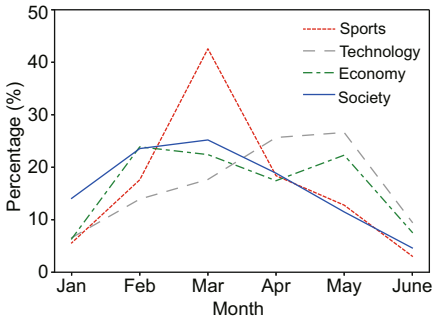

Fig. 6 Topic distribution over time

gradually and reached its peak in April and May. Then the proportion dropped sharply in June. During this period, with the outbreak and continuation of COVID-19, the number of news articles about vaccine development and experimental work by Chinese research institutes continued to grow. Most of the reports in technology news were related to Chinese vaccine research.

In terms of sports, the proportion of reports showed a general trend with time. The number of articles in January and February increased gradually. The proportion of sports news reached its peak in March, and then declined sequentially from April to June. During this period, on March 22, the International Olympic Committee officially announced the postponement of the 2020 Tokyo Olympics to 2021, and much news reported related events in March.

3.3 Fine-grained topic level media focus

The second layer of topic discovery is fine-grained topic detection, which automatically identifies topics from news streams. Different from coarse-grained topic classification, it does not have any preset topics. Its primary purpose is to learn from news articles and find topics about which the most news is concerned. We model this problem as an unsupervised learning problem rather than a supervised classification problem. On the other hand, due to the continuous emerging news, we have to simultaneously deal with vast amounts of data. Thus, we design an efficient and effective supervised topic detection method for clustering news by topics and detecting topics. Specifically, inspired by Keygraph (Sayyadi and Raschid, 2013), we construct an entity graph that contains only entities and keywords (including proper nouns, adjectives, and ordinal words) to make the algorithm more efficient. In addition, we use the SIFRank (Sun et al., 2020) algorithm based on the ELMo (Embeddings from Language Models) pre-training model to extract keywords, which improved performance. The learned topics are as follows:

1. In the first 20 days of January, the topics reported by various media focused mainly on the "novel coronavirus found in China," "virus outbreak in China," and "Wuhan is under lockdown." In late January, with the virus spread to Japan, Italy, Iran, and other countries, the media shifted its focus from China to other countries.

2. In February, the media reported mainly on topics such as "Aggregated events were postponed or canceled," "stock market volatility," and "the progress of the vaccine." It seems that the media began to pay more attention to the impact of the epidemic on human activities and the economy.

3. In March and April, "progress of Chinese vaccines," "specific drugs and treatments," and "virus outbreak in U.S." became the media's most concerned topics. The focus shifted away from China because the epidemic in China was well controlled. In March, topics such as "Tokyo Olympics" and "events postponed" also occupied a lot of forums. This corresponds to the sports news percentage peak in March in Fig. 6.

4. In May, financial topics such as "the stock market," "crude oil," and "exchange rate" received continuous attention. In the first half of this month, topics on "compensation from China for COVID-19", "virus origin," "vaccine competition," "China's second wave of epidemics," and "China-Australia relations" were hot. "NPC and CPPCC China" became the focus at the end of May.

5. In early June, the topics of media concern were scattered. Topics such as "People's Bank of China buys bank loans," "Stocks surged," "China will strengthen global cooperation in vaccine trials," "Trump administration says it will block Chinese airlines from flying into the U.S.," "China urges citizens to avoid Australia," "Harvard research," and "New virus cases raise fears in Beijing" were reported.

4 Tone of news against China (Q2)

Media usually bare three tones against China: support China (positive), oppose China (negative), and neutral. As stated in Section 2, we annotate our training dataset with tone labels (news media's tones against China), so we model this problem as a supervised learning problem. We first learn all word embeddings using word embedding techniques like BERT (Devlin et al., 2019), and then feed them into a supervised classifier.

After training, we are able to predict the news tones toward China. We find that most of the news has a neutral tone against China in our dataset of all media, accounting for 62%, as shown in Fig. 7. This result is in line with principles of news reporting.

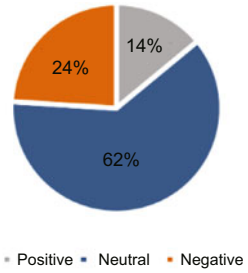

Fig. 7 Distribution of tone of news against China

4.1 Tone of news against China in different countries

We further analyze the tone of news at the country level. The tone of news against China in different countries is shown in Fig. 8. The horizontal axis represents the country and the vertical axis represents the percentage of news in a country that has a certain tone. We can see that France, the United Kingdom, and the United States hold a relatively negative tone toward China, while Russia, Singapore, Cuba, and Brunei bear a positive tone.

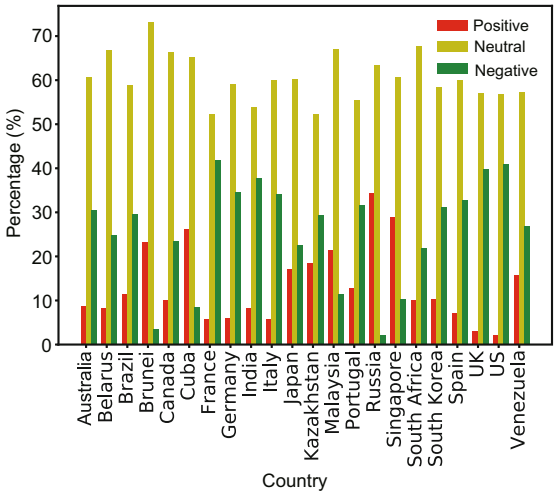

Fig. 8 Tone of news against China in different countries

We also calculate the similarity scores of each country against China. We collect all the news in a country and create statistics of their tones toward China. Hence, the proportions of positive, neutral, and negative tones can be viewed as a vector. By calculating this vector's similarity score, we can find a similar country that bears a similar tone against China. We find that France and the United Kingdom have a similar result of 0.913. The similarity score between France and the United States is 0.824, and the similarity between the United States and Germany is 0.817. This shows that Western powers' tones toward China are consistent, and their news shows more of their negative tones against China. On the contrary, Russia and Cuba have a similarity of 0.849, Russia and Brunei have a similarity of 0.797, and Brunei and Malaysia have a similarity of 0.735, showing a positive tone toward China.

4.2 Tone of news against China on different topics

The tone of news against China on different topics is shown in Fig. 9. The horizontal axis represents the topic and the vertical axis represents the percentage of a given topic that reports a particular tone. We find that international news media hold the most negative tone toward China when reporting political news and the second most negative tone when reporting economic news.

4.3 Tone of news over time

We report the tone of news over time in Fig. 10. The horizontal axis represents the month and the vertical axis represents the percentage of a given topic that reports a particular tone. We can see that the tone of the news media changed over time. In January, the negative proportion of most countries toward China was relatively low. With the continuous aggravation of the epidemic situation, the media of many countries (such as Singapore, Spain, and Germany) gradually increased their negative tone toward China in February. After March, the domestic epidemic situation gradually eased. The negative tone showed a significant downward trend, or it was

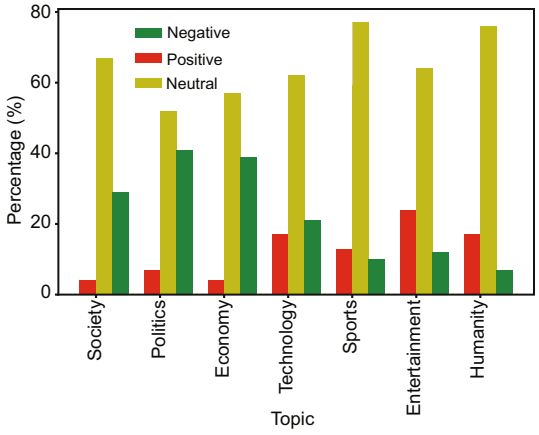

Fig. 9 Tone of news against China on different topics

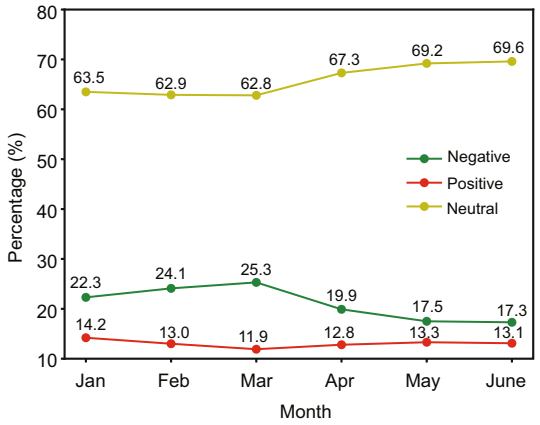

Fig. 10 Tone of news against China over time

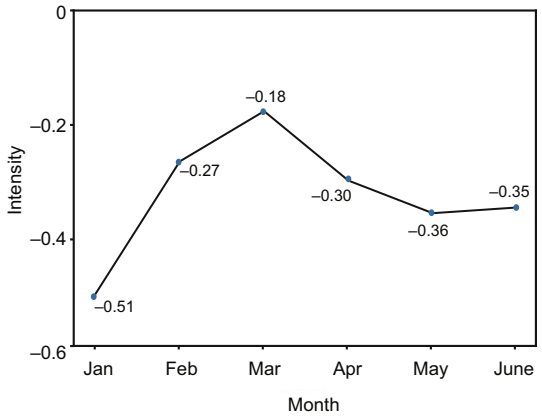

Fig. 11 Sentiment intensity over time

related to the worsening of the epidemic situation in other countries. The epidemic situation in China has been well controlled. The overall negative position shows an increasing trend with time first and then a decreasing one.

5 News emotion tones toward China (Q3)

We use two methods to determine news emotion tones toward China. One is to use sentiment intensity to measure the media's influence toward China quantitatively; the other is to use emotional labels to examine the emotional situation qualitatively.

5.1 Sentiment intensity index based news emotions

For news sentiment intensity, we use Vader (Hutto and Gilbert, 2014) to calculate the sentiment intensity of a news article. The original paper's sentiment intensity ranges from -1 to 1. -1 represents the most negative emotional value, while 1 represents the most positive emotional value. To distinguish the emotional intensity of news reports more clearly, we uniformly extend the range to $[-5, 5]$.

As shown in Fig. 11, news all over the world has reflected negative sentiment obviously, which means the intensity score is not equal to zero. Sentiment intensity has fluctuated for half a year. In January, when the epidemic began, the score was the lowest. In February, medical teams all over China galloped to Wuhan. In March, the epidemic was effectively controlled. Therefore, the score gradually returned to zero from January to March. From April to May, the outbreak of coronavirus had already taken place in various parts of the world. The world was full of different opinions of China for coronavirus. Remarkably, the United States passed the buck one after another, and the score gradually dropped. The score in June slightly increased compared with that in May.

The sentiment intensity in different countries is shown in Fig. 12. We can see that Malaysia has the highest sentiment intensity, while Canada has the lowest intensity. A country with high positive sentiment intensity means that it has a positive attitude toward China, and vice versa. This result corresponds to our discussion in Section 4.

5.2 Label-based news emotions

A piece of news usually shows or implicitly expresses opinions on an event, a person, or other

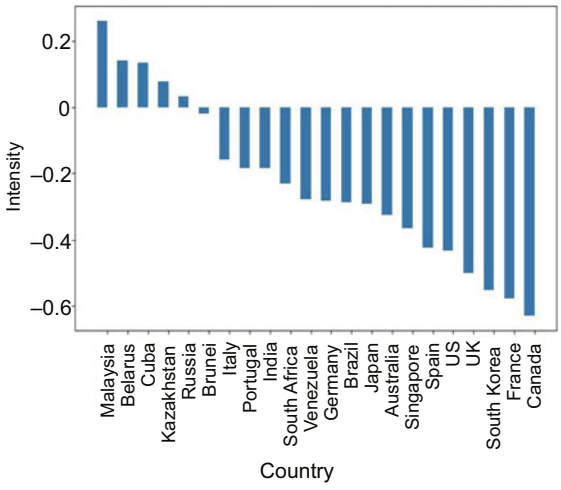

Fig. 12 Sentiment intensity for each country

targets, reflecting the author's emotions. We divide emotions into six categories: agreeable, believable, good, hated, worried, and sad. If a piece of news contains none of these emotions, we think it is objective.

For label-based emotions, considering that a piece of news may contain multiple emotions simultaneously, even opposite emotions, we need to design a useful multi-label emotion classification model to sufficiently capture the semantics of news context. We employ BERT (Devlin et al., 2019) as the feature extractor. The input to our model includes news headline and body content. After using the feature extractor, we obtain a sequence of the last hidden states and then retain the first token of the sequence (classification token). This token is fed to a linear layer with a sigmoid activation function, which predicts six probability distributions corresponding to defined emotions. The threshold is set to 0.4. In other words, if the prediction probability of emotion is greater than or equal to 0.4, we consider that the news contains this emotion. Note that if all six probabilities are less than 0.4, we consider the news to be objective. The overall illustration of our multi-label emotion classification model is shown in Fig. 13.

As shown in Fig. 14, considering non-subjective news articles, we find that international news toward China holds more negative emotions than positive emotions, up to 26.0% and 5.5%, respectively. China, one of the countries with an early outbreak of the virus, has suffered from public criticism. The rapidly increasing number of infections makes the emotional tone of overseas media present negative emotions such as "critical" and "anger." Perhaps this explains why positive feelings toward China are only 5.5%.

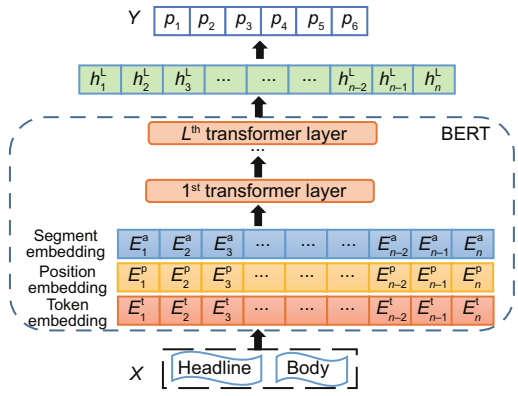

Fig. 13 Illustration of the multi-label emotion classification model

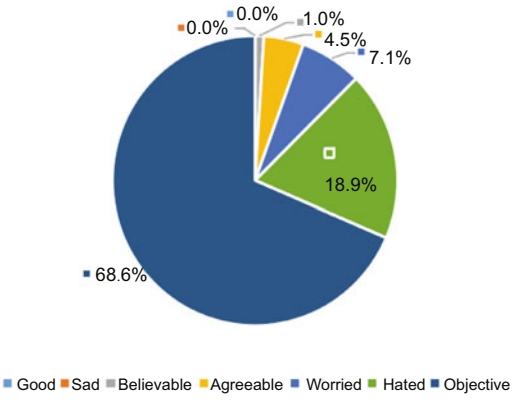

Fig. 14 Emotion distribution

As shown in Fig. 15, we find that France's news reports have an extremely high percentage of "hated" emotion, followed by Canada, the United Kingdom, South Korean, Spain, and the United States.

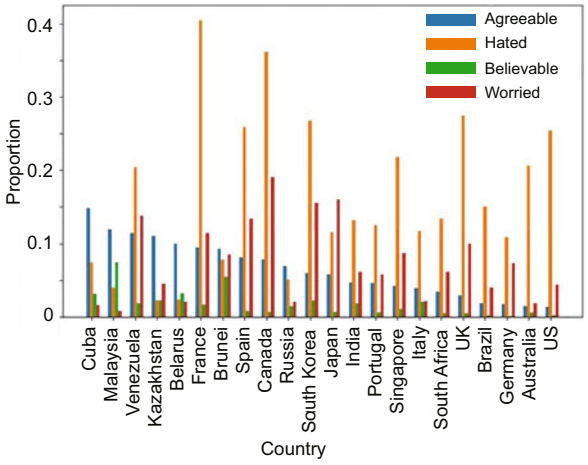

Fig. 15 Emotions in different countries

We rank the typical emotions in descending order for the convenience of comparison. As we can see in Fig. 16, the proportion of "agreeable" toward China is generally low in every country, and the highest is Cuba, followed by Malaysia, Venezuela, Kazakhstan, and Belarus, while the lowest is the United States. As for the "hated" emotion, the highest is France, while the lowest is Kazakhstan.

We run the k-means algorithm to cluster different countries, where the number of clusters is set to 3, and input data is the proportion of emotions except "objective." The clustering results are shown in Table 3. As far as we can see, cluster 2 shows more positive emotions than others, while cluster 3 shows more negative emotions than others.

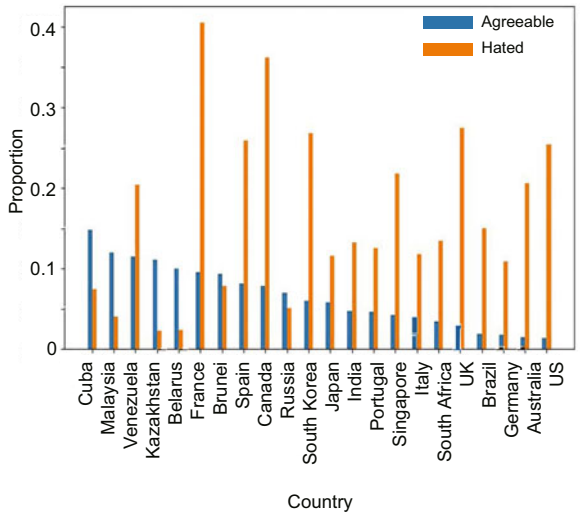

Fig. 16 Proportion of "agreeable" and "hated" emotions for each country

Table 3 Results of country clusters based on the proportion of emotions

Cluster	Countries
1	Japan, India, Portugal, Singapore, Italy, South Africa, Brazil, Germany, Australia, the United States
2	Cuba, Malaysia, Kazakhstan, Belarus, Brunei, Russia
3	Venezuela, France, Spain, Canada, South Korea, the United Kingdom

6 Prototype system

We developed and deployed a visual system (http://203.195.140.107) to show the whole news analysis process in this study. Fig. 17 shows the system framework. It consists of five modules:

1. Data collection

We crawled news from 57 news websites of mainstream media in 22 countries and updated the data automatically every day. Details are as given in Section 2.

2. Data preprocessing

We cleaned up the crawled data by strict standards and translated multilingual news into English. For better model learning, we annotated 5000 pieces of news with crowdsourcing technology. Details are as given in Section 2.

3. Data analysis

We aimed to answer three questions: (1) What has the international media focused on during the COVID-19 epidemic period? (2) What is the media's tone when they report China? (3) What is the media's attitude when talking about China? We deployed some modules that would be used to answer these questions, such as named entity recognition, topic classification, and topic clustering. Details are as given in Sections 3–5.

4. Data visualization

We displayed our system in a hierarchical manner, as shown in Fig. 18.

5. Storage services

Storage and querying of the knowledge graph are the keys to the entire system. To persistently store and analyze the knowledge graph data, we used two different types of databases to store data at different stages in the data processing procedure. We used the document-based MongoDB to store the crawled data. In addition, for mining information and data at a deeper level, we leveraged Neo4j (www.neo4j.com) to store the knowledge graph data of entities, topics, and events.

7 Related work

7.1 Country image

Nimmo and Savage (1976) defined an image as "a human construct imposed on an array of perceived attributes projected by an object, event or person." The traditional analysis of national image often uses surface analysis based on related corpora and news content. Manheim and Albritton (1984) proposed two dimensions to describe the national image, visibility and valence, which represent the media's influence range and the degree of preference in the media content on the country, respectively.

In different media, China is portrayed with different national images. Wang (2003) compared China's national image as projected by Chinese media and American media based on a content analysis between 1958 and 2002. Peng (2004) studied the coverage of China in *New York Times* and *Los Angeles Times*. Zhang L (2010) explored the image of China in three international newspapers in Europe. Zhang L and Wu (2017) used critical discourse analysis to examine the representation of China by *China Daily*. To summarize all of these studies, it was found that China's national image was portrayed as a peace-loving country, a developing country, and an anti-hegemonic nation by Chinese media (Wang, 2003; Zhang L and Wu, 2017). The image of China

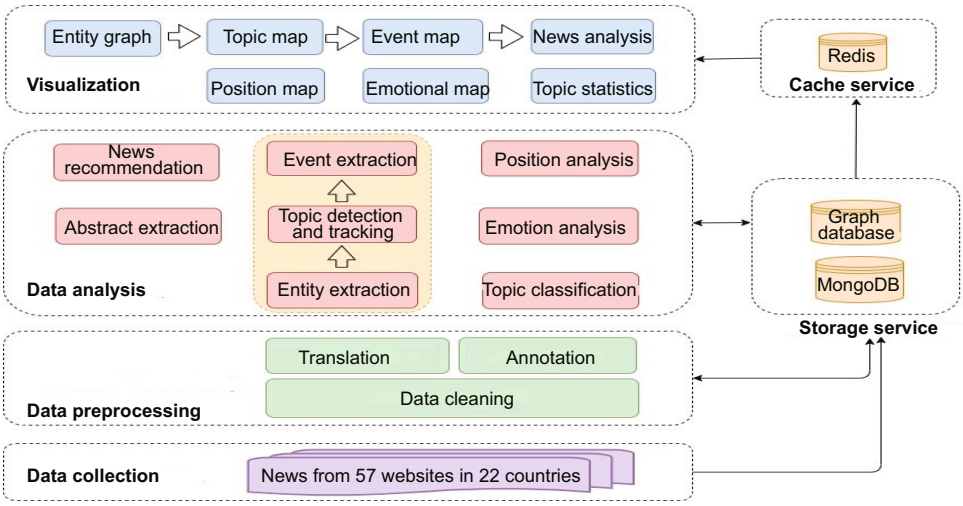

Fig. 17 System framework

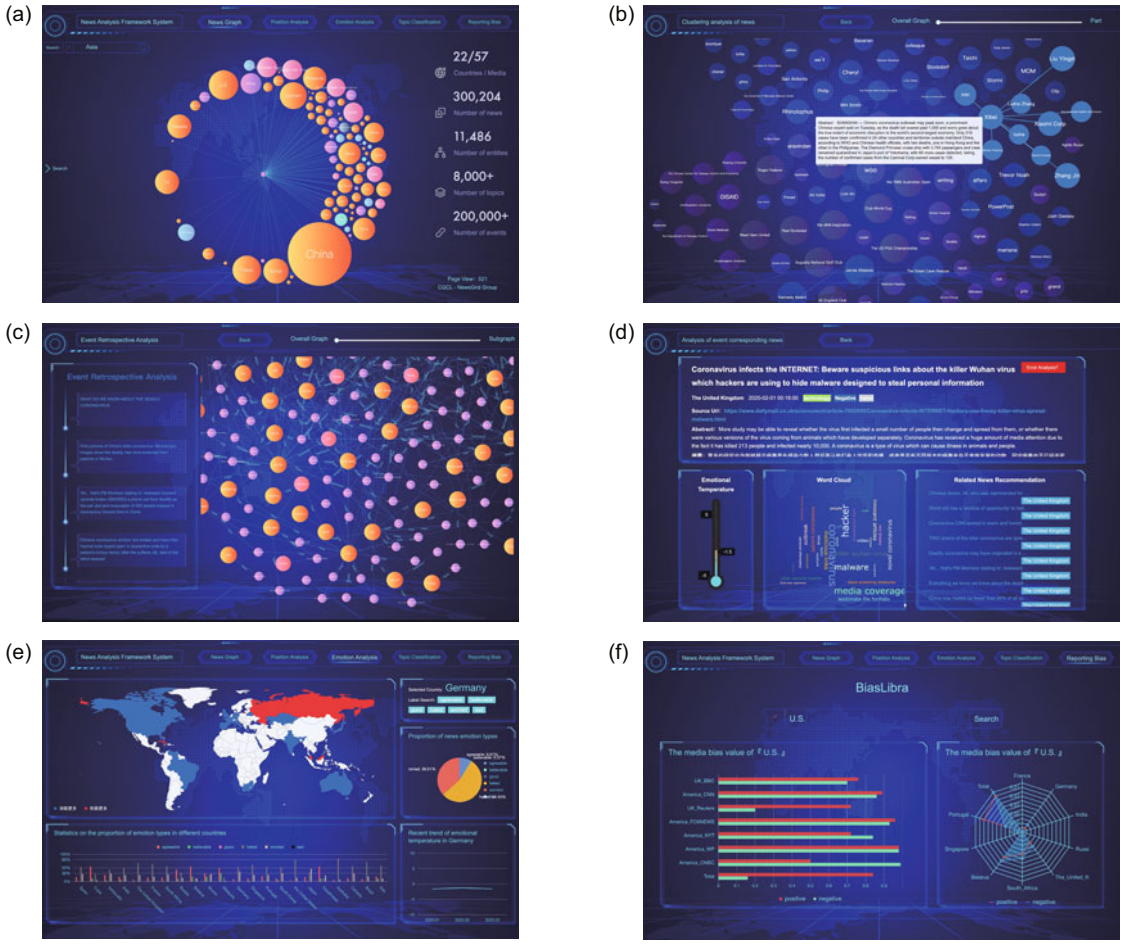

Fig. 18 System visualization

The first three figures (a–c) are visualizations of a multi-level media focus. The first layer (a) shows the entity graph extracted from all news corpora. In the graph, each node is an entity. When clicking one entity node, we can get into the second layer (b) and see a detailed graph related to that entity. Move the cursor to display fine-grained topics. Click one node in this graph to see the third layer (c) and detailed events associated with this node. The last three figures (d–f) show coarse-grained topics and the tone of news and emotion analysis

portrayed by foreign media was usually mixed and conflicted, such as a socialist country, a significant power, an authoritarian state, and a militant obstructive force (Zhang L and Wu, 2017). These methods prefer manual analysis and lack fine-grained analysis and scalability. Chen et al. (2021) investigated China's image during the COVID-19 pandemic with aspect-based sentiment analysis. Compared to our system, it is limited to sentiment analysis. Moreover, our system provides a multi-level and multi-view country image analysis.

7.2 News system

Several systems have been built to analyze news from multiple perspectives. For news aggregation and analysis, Google News (www.news.goole.com) is the largest news aggregation system and monitored more than 5000 news sources worldwide as of 2013 (Filloux, 2013). It performs topic detection, tracking, and clustering of news. It also uses algorithms to offer personalized news to users. Lloyd et al. (2005) built a news analysis system named Lydia to track temporal and spatial information of entities in the news. Hou et al. (2015) built NewsMiner (www.newsminer.net), which is a news-mining system framework. They proposed a three-level representation of news and formalized news-mining tasks as link predictions in the heterogeneous network. For fake news detection, Emergent (www.emergent.info) is a rumor tracking tool developed by Columbia University to study the spreading mechanism of rumors. For news and event detection, Liu et al. (2016) built a tool that can help journalists discover news on social media more quickly and assess news authenticity. There are also news detection and tracking systems and services provided by some companies or organizations, such as People's Daily Online (www.peopleyun.cn/yuqing.html), Baidu (www.yuqing.baidu.com/saas/intro/newindex), and Nielsen (www.nielsen.com). These systems usually monitor forums, blogs, news media, and social networks to find content relevant to a particular topic or keyword. Statistical analysis is performed based on the collected content to provide consulting services. The above works focused mainly on different aspects of news analysis, thus different from our system at different granularities.

7.3 Media analysis

Media analysis is also known as media content analysis, which is a part of content analysis. Macnamara (2005) interpreted content analysis as a technique for describing what is said at a given place and a given time with objectivity and accuracy. However, slanted news coverage always exists in real life.

In content analysis, researchers first define analysis questions or assumptions that need to be studied. Then they collect relevant news data, systematically read the news text, and annotate the text with examples of media bias associated with the ongoing analysis. After that, researchers use the annotated findings to accept or reject their hypothesis (McCarthy et al., 2008; Oelke et al., 2012). There are two types of content analysis: quantitative and qualitative (Vaismoradi et al., 2013). Qualitative analysis attempts to find "all" instances of media bias, including some that require human interpretation. Quantitative analysis measures news by determining the frequency of a particular word or phrase, the number of articles that include that word or phrase, or the size and location of an article in a printed newspaper (D'Alessio and Allen, 2000). They also use computer software to aid analysis, for example, by analyzing how often terms, topics, or words appear together (Lowe, 2002).

In addition to content analysis, media bias can be analyzed through public opinion polls or public votes, such as through the Gallup/Knight Foundation (John and James, 2018) and MBFC (www.mediabiasfactcheck.com). In analyzing media sentiment and content analysis, computational methods can be used for sentiment analysis. For example, Hutto and Gilbert (2014) presented VADER, a simple rule-based model for general sentiment analysis. Neri et al. (2012) used linguistic and semantic approaches to analyze sentiment about newscasts. In analysis of the media's tone, opinion mining has mostly focused on polarity detection of reviews by classifying the given text as positive, negative, or neutral. There are several existing tone detection models, including both neural models and classical classifier-based models (Ghosh et al., 2019). Zhang Q et al. (2018) defined this problem as a ranking one and proposed a ranking-based method to maximize the differences among different tones.

8 Conclusions

In this study, we focused on how international news media portrayed China during the COVID-19 epidemic period. We answered three questions using big data techniques: (1) What has the international media focused on during the COVID-19 epidemic period? (2) What is the media's tone when they report China? (3) What is the media's attitude when talking about China? Specifically, we crawled more than 280 000 pieces of news from 57 mainstream news media entities in 22 countries and made a detailed analysis. We found that during the second wave of the COVID-19 epidemic period, mainstream media cared more about medical scientists. Also, during the COVID-19 epidemic period, Singapore and Malaysia were more concerned about China's economy, whereas Canada and France were more concerned about Chinese politics. In March and April, "progress of Chinese vaccines," "specific drugs and treatments," and "virus outbreak in U.S." became the topics that most concerned the media. In terms of news emotion toward China, Cuba, Malaysia, and Venezuela had a positive attitude, while France, Canada, and the United Kingdom had a negative one. Our study can help understand China's image in the eyes of the international media and provide a sound basis for image analysis.

Contributors

Hong HUANG and Xuanhua SHI designed the research. Zhexue CHEN, Chenxu WANG, and Zepeng HE processed the data. Hong HUANG, Zhexue CHEN, and Chenxu WANG drafted the manuscript. Hai JIN, Mingxin ZHANG, and Zongya LI helped revise the manuscript. Hong HUANG finalized the paper.

Acknowledgements

Special thanks to Yi FENG, Mingqi LAI, Rui ZHANG, Mohan ZHANG, Yu LI, and Haoshuang CAO for their efforts devoted to this project.

Compliance with ethics guidelines

Hong HUANG, Zhexue CHEN, Xuanhua SHI, Chenxu WANG, Zepeng HE, Hai JIN, Mingxin ZHANG, and Zongya LI declare that they have no conflict of interest. This study is GDPR (General Data Protection Regulation) compliant.

Data availability

The data that support the findings of this study are openly available at http://203.195.140.107/dataset/download.

References

Chen HM, Zhu ZY, Qi FC, et al., 2021. Country image in COVID-19 pandemic: a case study of China. *IEEE Trans Big Data*, 7(1):81-92. https://doi.org/10.1109/TBDATA.2020.3023459

D'Alessio D, Allen M, 2000. Media bias in presidential elections: a meta-analysis. *J Commun*, 50(4):133-156. https://doi.org/10.1111/j.1460-2466.2000.tb02866.x

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics, p.4171-4186.

Filloux F, 2013. Google News: the Secret Sauce. Monday Note. https://mondaynote.com/google-news-the-secret-sauce-3f1cec521209 [Accessed on Feb. 23, 2021].

Ghosh S, Singhania P, Singh S, et al., 2019. Stance detection in web and social media: a comparative study. Int Conf of the CLEF Association, p.75-87. https://doi.org/10.1007/978-3-030-28577-7_4

Honnibal M, Montani I, 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. GitHub. https://github.com/xtrancea/spaCy [Accessed on Feb. 23, 2021].

Hou L, Li J, Wang Z, et al., 2015. NewsMiner: multifaceted news analysis for event search. *Knowl-Based Syst*, 76:17-29. https://doi.org/10.1016/j.knosys.2014.11.017

Hutto C, Gilbert E, 2014. Vader: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf on Web and Social Media, p.216-225.

John S, James L, 2018. Perceived Accuracy and Bias in the News Media. Knight Foundation. https://knightfoundation.org/wp-content/uploads/2020/03/KnightFoundation_AccuracyandBias_Report_FINAL.pdf [Accessed on Feb. 23, 2021].

Jurafsky D, Martin JH, 2000. Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, New Jersey, USA.

Kim Y, 2014. Convolutional neural networks for sentence classification. Proc Conf on Empirical Methods in Natural Language Processing, p.1746-1751.

Liu X, Li Q, Nourbakhsh A, et al., 2016. Reuters tracer: a large scale system of detecting & verifying real-time news events from Twitter. ACM Int Conf on Information and Knowledge Management, p.207-216. https://doi.org/10.1145/2983323.2983363

Lloyd L, Kechagias D, Skiena S, 2005. Lydia: a system for large-scale news analysis. Int Conf on String Processing and Information Retrieval, p.161-166. https://doi.org/10.1007/11575832_18

Lowe W, 2002. Software for Content Analysis—a Review. Harvard University. https://dl.conjugateprior.org/preprints/content-review.pdf [Accessed on Apr. 23, 2021].

Macnamara JR, 2005. Media content analysis: its uses, benefits and best practice methodology. *Asia Pacif Publ Rel J*, 6(1):1-34.

Manheim JB, Albritton RB, 1984. Changing national images: international public relations and media agenda setting. *Am Pol Sci Rev*, 78(3):641-657. https://doi.org/10.2307/1961834

McCarthy J, Titarenko L, McPhail C, et al., 2008. Assessing stability in the patterns of selection bias in newspaper coverage of protest during the transition from communism in Belarus. *Mob Int Q*, 13(2):127-146. https://doi.org/10.17813/maiq.13.2.u45461350302663v

Neri F, Aliprandi C, Capeci F, et al., 2012. Sentiment analysis on social media. IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining, p.919-926. https://doi.org/10.1109/ASONAM.2012.164

Nimmo DD, Savage RL, 1976. Candidates and Their Images: Concepts, Methods, and Findings. Goodyear Publishing Company, Pacific Palisades, USA.

Oelke D, Geißelmann B, Keim DA, 2012. Visual analysis of explicit opinion and news bias in German soccer articles. Int Workshop on Visual Analytics, p.49-53. https://doi.org/10.2312/PE/EuroVAST/EuroVA12/049-053

Peng Z, 2004. Representation of china: an across time analysis of coverage in the *New York Times* and *Los Angeles Times*. *Asian J Commun*, 14(1):53-67. https://doi.org/10.1080/0129298042000195170

Sayyadi H, Raschid L, 2013. A graph analytical approach for topic detection. *ACM Trans Intern Technol*, 13(2):4. https://doi.org/10.1145/2542214.2542215

Sun Y, Qiu H, Zheng Y, et al., 2020. SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access*, 8:10896-10906. https://doi.org/10.1109/ACCESS.2020.2965087

Vaismoradi M, Turunen H, Bondas T, 2013. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. *Nurs Health Sci*, 15(3):398-405. https://doi.org/10.1111/nhs.12048

Wang H, 2003. National image building and Chinese foreign policy. *China Int J*, 1(1):46-72. https://doi.org/10.1142/S0219747203000050

Zhang L, 2010. The rise of China: media perception and implications for international politics. *J Contemp China*, 19(64):233-254. https://doi.org/10.1080/10670560903444199

Zhang L, Wu D, 2017. Media representations of China: a comparison of China Daily and Financial Times in reporting on the belt and road initiative. *Crit Arts*, 31(6):29-43. https://doi.org/10.1080/02560046.2017.1408132

Zhang Q, Yilmaz E, Liang S, 2018. Ranking-based method for news stance detection. The Web Conf, p.41-42. https://doi.org/10.1145/3184558.3186919

Hong HUANG, first author of this invited paper, is an associate professor at Huazhong University of Science and Technology (HUST), Wuhan China. She received her PhD degree in computer science from the University of Göttingen, Germany in 2016, and her ME degree in electronic engineering from Tsinghua University, Beijing, China in 2012. Her research interests include social network analysis, social influence, and data mining.

Xuanhua SHI, corresponding author of this invited paper, is currently a professor with the School of Computer Science and Technology, HUST, Wuhan, China. He is the deputy director of the National Engineering Research Center for Big Data Technology and System (NERCBDTS). He published more than 100 peer-reviewed papers in conferences and journals such as ASPLOS, VLDB, *ACM Trans Comput Syst*, and *IEEE Trans Parall Distr Syst*. He is a corresponding expert of *Front Inform Technol Electron Eng*. He received research supports from several governmental and industrial organizations, such as the National Natural Science Foundation of China, Ministry of Science and Technology, Ministry of Education, and the European Union. His current research interests include cloud computing, big data processing, and AI systems.

Hai JIN received his PhD degree in computer engineering from HUST, Wuhan, China, in 1994. He worked at the University of Hong Kong from 1998 to 2000, and was a visiting scholar at the University of Southern California, Los Angeles, CA, USA from 1999 to 2000. He is currently the Cheung Kung Scholars Chair Professor of Computer Science and Engineering with HUST. He has coauthored 15 books, and published over 700 research articles. He is now serving as an editor of *Front Inform Technol Electron Eng*. He was awarded the German Academic Exchange Service Fellowship to visit the Technical University of Chemnitz, Germany in 1996, and was supported by the National Natural Science Foundation of China for Distinguished Young Scholars in 2001. He is a Fellow of China Computer Federation (CCF) and a Life Member of Association for Computing Machinery (ACM). His research interests include computer architecture, virtualization technology, cluster computing and cloud computing, peer-to-peer computing, network storage, and network security.

Appendix: News statistical information of source media

Table A1 All media sources and their statistical information in our dataset

Country	Media	URL	COVID19 + China		COVID19	
Australia	Sydney Morning Herald	https://www.smh.com.au	967	1875	8240	12 983
	The Age	https://theage.com.au	908		4743	
Belarus	belta	https://www.belta.by	510	804	2194	4116
	naviny	https://naviny.by	294		1922	
Brazil	Agência Brasil	https://agenciabrasil.ebc.com.br/	552	2302	3857	16 425
	Folha de Sao Paulo	https://www.folha.uol.com.br	1750		12 568	
Brunei	borneobulletin	https://borneobulletin.com.bn	520	520	3692	3692
Canada	nationalpost	https://nationalpost.com/	1477	1668	9202	12 446
	theglobeandmail	https://www.theglobeandmail.com/	191		3244	
Cuba	Granma	https://www.granma.cu/	247	247	1048	1048
France	RFI	https://www.fi.fr/	1440	2598	1545	8033
	Le Nouvel Observateur	https://www.nouvelobs.com/	206		2612	
	Le Monde	https://www.lemonde.fr/	660		3388	
	Le Figaro	https://www.lefigaro.fr/	292		488	
Germany	Süddeutsche Zeitung	https://www.sueddeutsche.de/	1642	3888	20 791	31 046
	Frankfurter Allgemeine Zeitung	https://www.faz.net/aktuell/	118		533	
	BLID	https://www.bild.de	271		934	
	WESER-KURIER	https://www.weser-kurier.de	684		4264	
	Der Spiegel	https://www.spiegel.de	1173		4524	
India	IndiaToday	https://www.indiatoday.in	2218	2902	2350	2986
	Rediff	https://www.rediff.com	684		636	
Italy	lastampa	https://lastampa.it/	907	907	6553	6553
Japan	読売新聞	https://www.yomiuri.co.jp	511	1971	650	9759
	毎日新聞	https://mainichi.jp	1226		8875	
	朝日新聞	https://www.asahi.com	234		234	
Kazakhstan	kazpravda	www.kazpravda.kz	86	86	1637	1637
Malaysia	bernama	https://www. bernama.com/	313	355	1319	3208
	The Star	https://www.thestar.com.my/	42		1889	
Portugal	CMJORNAL	https://www.cmjornal.pt/	1616	3537	7134	14 850
	EXPRESSO	https://expresso.pt/	1219		6176	
	RTP	https://www.rtp.pt/	276		291	
	PUBLICO	https://expresso.pt/	426		1249	
Russia	RIA	https://ria.ru	1672	11 155	3341	56 210
	TASS	https://tass.com	7680		35 552	
	RG.RU	https://rg.ru	1803		17 317	
Singapore	TODAY	https://www.todayonline.com	2678	9350	16 714	45 652
	Channel News Asia	https://www.channelnewsasia.com/	4136		7671	
	Straits Times	https://www.straitstimes.com	2536		21 267	
South Africa	News24	https://www.news24.com	342	427	342	427
	Sport24	https://www.sport24.co.za/	85		85	
South Korea	조선뉴스	https://www.chosun.com	1786	3677	1998	4776
	중앙일보	https://news.joins.com	1127		1773	
	동아일보	https://www.donga.com	541		559	
	경향신문	https://khan.co.kr	223		446	
Spain	La Vanguardia	https://www.lavanguardia .com/	524	1239	839	13 701
	El Mundo	https://www.elmundo.es/	281		3181	
	EL Pais	https://elpais.com/	434		9681	
The United Kingdom	Daily Mail	https://www.dailymail.co.uk	2349	4770	2604	5258
	BBC	https://www.bbc.com/news	274		282	
	Reuters	https://www.reuters.com	2147		2372	
The United States	CNN	https://www.cnn.com/	1547	13 158	4308	28 698
	CNBC	https://www.cnbc.com/	2888		6014	
	NYT	https://www.nytimes.com/	3291		10 886	
	WP	https://www.washingtonpost.com/	3109		4913	
	FOX NEWS	https://www.foxnews. com/	2323		2577	
Venezuela	El Nacional	https://www.elnacional.com/	358	519	841	2316
	EI Universal	https://www.eluniversal.com/index.html	161		1475	