frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 10 November 2023
DOI 10.3389/fpubh.2023.1281259

OPEN ACCESS

EDITED BY
Yi-Ju Tseng,
National Yang Ming Chiao Tung University, Taiwan

REVIEWED BY
Soodamani Ramalingam,
University of Hertfordshire, United Kingdom
Xue Lan,
China Medical University, China

*CORRESPONDENCE
Lei Shi
✉ leiky_shi@cuc.edu.cn

RECEIVED 22 August 2023
ACCEPTED 23 October 2023
PUBLISHED 10 November 2023

CITATION
Luo J, Peng D, Shi L, El Baz D and Liu X (2023) A comparative analysis of the COVID-19 Infodemic in English and Chinese: insights from social media textual data.
Front. Public Health 11:1281259.
doi: 10.3389/fpubh.2023.1281259

A comparative analysis of the COVID-19 Infodemic in English and Chinese: insights from social media textual data

Jia Luo[1,2], Daiyun Peng[1], Lei Shi[3,4]*, Didier El Baz[5] and Xinran Liu[1]

[1]College of Economics and Management, Beijing University of Technology, Beijing, China, [2]Chongqing Research Institute, Beijing University of Technology, Chongqing, China, [3]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China, [4]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, China, [5]LAAS-CNRS, Université de Toulouse, Toulouse, France

The COVID-19 infodemic, characterized by the rapid spread of misinformation and unverified claims related to the pandemic, presents a significant challenge. This paper presents a comparative analysis of the COVID-19 infodemic in the English and Chinese languages, utilizing textual data extracted from social media platforms. To ensure a balanced representation, two infodemic datasets were created by augmenting previously collected social media textual data. Through word frequency analysis, the 30 most frequently occurring infodemic words are identified, shedding light on prevalent discussions surrounding the infodemic. Moreover, topic clustering analysis uncovers thematic structures and provides a deeper understanding of primary topics within each language context. Additionally, sentiment analysis enables comprehension of the emotional tone associated with COVID-19 information on social media platforms in English and Chinese. This research contributes to a better understanding of the COVID-19 infodemic phenomenon and can guide the development of strategies to combat misinformation during public health crises across different languages.

KEYWORDS

COVID-19, infodemic data, word frequency analysis, topic clustering analysis, sentiment analysis

1 Introduction

During the beginning of the COVID-19 pandemic, there was a surge in misinformation, false information, and rumors spreading rapidly across various media platforms. This phenomenon came to be known as the infodemic. The infodemic refers to the overwhelming abundance and rapid spread of misinformation, conspiracy theories, and unverified claims related to the pandemic (1). It accompanied the spread of the virus itself and was fueled by uncertainties, fear, and confusion during the early stages of the outbreak (2). Numerous falsehoods and conspiracy theories circulated globally, making it challenging for individuals to discern accurate information. Some examples include claims that the virus was intentionally created or released, that certain medications or alternative remedies could cure or prevent the virus, or that 5G networks were somehow linked to the spread of the disease. The infodemic had significant implications on public health, as it hindered effective pandemic response efforts. False information about prevention measures, symptoms, and treatments could potentially mislead the public and endanger lives. It also led to widespread panic, social unrest, and stigmatization of certain groups. Although the World Health Organization (WHO) declared an end to

COVID-19 as a global health emergency, it is important to note that combating the infodemic remains an ongoing challenge.

COVID-19 is a global pandemic, and misinformation knows no boundaries. Conducting a comparative analysis across different languages allows us to gain a comprehensive, global perspective on the infodemic. On one side, each language has its unique linguistic characteristics, cultural norms, and online behaviors. Analyzing social media data in different languages can uncover language-specific nuances that shape misinformation patterns and responses. On the other side, analyzing data from multiple languages uncovers common themes, misinformation tactics, and influential narratives. Understanding these cross-cultural trends allows for exchanging knowledge and implementing effective global strategies. According to Statista (3), English and Chinese are the top-2 most common languages used on the Internet. Therefore, this paper aims to conduct a comparative analysis of the COVID-19 infodemic in both English and Chinese languages by utilizing textual data extracted from social media platforms. The main contributions of our work are summarized as follows:

1. Two balanced infodemic datasets are introduced by adjusting previously collected social media textual data with annotations from healthcare workers where all records are classified into three distinct groups: true, false, and uncertain.
2. Word frequency analysis is conducted to identify the 35 most frequently occurring infodemic words to acquire knowledge on the prevalent patterns and trends of word usage in two languages.
3. Topic clustering analysis is executed to uncover thematic structures to gain insights into the similarities and differences between different topics or subject areas across two languages.
4. Sentiment analysis is performed to determine the percentage of positive, neutral, or negative sentiments within infodemic records to understand the emotional tone and attitudes expressed in two languages.
5. A discussion is held to grasp the language-specific nuances and cross-cultural trends of both the overall records and the records classified into three groups. The latter offers perspectives at a more refined level by incorporating the professional knowledge of healthcare workers.

The subsequent sections of this paper are organized as follows. Section 2 introduces related works. Section 3 displays the two balanced infodemic datasets. Section 4 provides the results of word frequency analysis, topic clustering analysis, and sentiment analysis, respectively. Afterward, a discussion is illustrated in Section 5. Finally, section 6 presents conclusions.

2 Related works

The majority of scholarly research about infodemic centers on addressing misinformation while trained models incorporating word embeddings stand out as the most commonly utilized methods (4). Glazkova et al. (5) proposed an approach using the transformer-based ensemble of COVID-Twitter-BERT models to detect COVID-19 fake news in English. Chen et al. (6) studied a novel transformer-based language model fine-tuning approach for English fake news detection during COVID-19. Paka et al. (7) set up a cross-stitch semi-supervised neural attention model for COVID-19 fake news detection which leverages the large amount of unlabelled data from Twitter in English. Chen et al. (8) used fuzzy theory to extract features and designed multiple deep-learning model frameworks to identify Chinese and English COVID-19 misinformation. Liu et al. (9) developed a deep learning based model and fine-tuned it to adapt to the specific domain context of COVID-19 news classification in English, Chinese, Arabic, and German. While these models have undoubtedly improved the efficacy in combatting misinformation during the COVID-19 pandemic, they often overlook the critical aspect of elucidating the underlying characteristics of the infodemic. Without being transformed into human-understandable knowledge, their outputs would have limited efficacy in aiding human efforts to combat the infodemic and develop targeted countermeasures and mitigation strategies.

Certain academic studies pay their attention to comprehending the patterns exhibited within the COVID infodemic through an in-depth analysis of its content. Gupta et al. (10) identified topics and key themes present in English COVID-19 fake and real news, compared the emotions associated with these records and gained an understanding of the network-oriented characteristics embedded within them. Wan et al. (11) described the prominent lexical and grammatical features of English COVID-19 misinformation, interpreted the underlying (psycho-)linguistic triggers, and studied the feature indexing for anti-infodemic modeling. Zhao et al. (12) used 1,296 COVID-19 rumors collected from an online platform in China, and found measurable differences in the content characteristics between true and false rumors. Zhou et al. (13) investigated both thematic and emotional characteristics of COVID-19 fake news at different levels and compared them in English and Chinese. All of the aforementioned works prioritize conducting analysis using a binary truth classification system, precisely distinguishing between true and false categories, to minimize discrepancies arising from truth labeling. However, it is incumbent upon us to acknowledge the inherent challenges faced when adjudicating the authenticity or veracity of certain statements during the labeling process.

The majority of collected records utilized in the analysis and detection of the infodemic phenomenon are typically categorized and labeled as either true or false (14). Nonetheless, a limited number of studies have undertaken an alternative approach by classifying these records into 3–5 categories to have a more comprehensive understanding of the infodemic and its impact at a finer level of granularity. Cheng et al. (15) built up an English COVID-19 rumor dataset by gathering news and tweets and manually labeling them as true, false, or unverified. Haouari et al. (16) proposed an Arabic COVID-19 Twitter dataset where each tweet was marked as true, false, or others. Luo et al. (17) collected widely spread Chinese infodemic during the COVID-19 outbreak from Weibo and WeChat while each record was indicated as true, false, or questionable after a four-time adjustment. Kim et al. (18) produced a dataset encompassing English claims and corresponding tweets, which were organized into four groups: COVID true, COVID fake, non-COVID true, and non-COVID fake. Dharawat et al. (19) released a dataset for health risk assessment of COVID-19-related social media posts. There are English tweets and tokens and all of them were classified into five categories: real news/claims, not severe, possibly severe misinformation, highly severe misinformation, or refutes/rebuts

misinformation. Given the profound interconnectedness between the infodemic and health records and its notable implications for public health, the active involvement of healthcare workers could help advance the comprehension of the infodemic. However, only (17) have considered this aspect while categorizing the collected records.

Considering the above-mentioned analysis, most studies have predominantly focused on English records. Therefore, it is valuable to conduct a comparative study of the COVID-19 Infodemic in multiple languages. The previously collected social media textual data offer an initial starting point while the integration of healthcare workers' professional knowledge serves to enhance insights at a more refined level. Additionally, conducting an analysis incorporating lexical, topical, and sentiment features would contribute to a comprehensive understanding of the underlying characteristics.

3 Data collection

English and Chinese records are chosen for this study because of their status as the two most prevalent languages used on the Internet (3). A summary of the encompassed data is presented in Table 1.

The English data is sourced from Patwa et al. (20). It collects 5,100 fake news from public fact-verification websites and social media. On the other side, there are 5,600 real news and they are tweets crawled from official and verified Twitter handles of the relevant sources using Twitter API. The dataset is split into train (60%), validation (20%), test (20%) and the training set has been selected for this study. The training set was published on October 1, 2020 (21) and consists of 3,360 real news and 3,060 fake news. We have invited three healthcare workers to manually classify these 6,420 records into three distinct groups: true, false, and uncertain. Their assessments rely exclusively on their judgments without any reference to external sources, and the assigned label for each record is determined by employing a majority agreement methodology. To address the limited number of instances in the real group (830 records), we randomly selected 830 records from both instances labeled as true and uncertain. Finally, a total of 2,490 records were kept, with an equal distribution for each group to mitigate any potential bias and to ensure fairness in representing various categories.

The Chinese data is derived from Luo et al. (17). This dataset gathers a total of 797 original records, which include manually verified Weibo posts from the Sina Community Management Center between January 21 and April 10, 2020, and specifically checked news from the WeChat mini-program "Jiaozhen" until March 31, 2020. All instances are classified into two types based on their content: strongly related health records and weakly related health records. The weakly related health records are further subdivided into specific categories, which include local measures, national measures, patient information, and others. Subsequently, four rounds of adjustments are conducted: (1) adjusting labels for instances classified as weakly related health records, (2) adjusting labels for records initially marked as partially true or conditionally true, (3) removing dummy records in the sub-group of local measures, (4) adding strongly related health records from authoritative sources to the true group. In the end, the dataset consists of 1,055 records overall, with 409 labeled as questionable, 276 as false, and 335 as true, ensuring that each group contains roughly an equal number of records. Since there is high intercoder reliability between the final labels and labels annotated by healthcare workers, we keep the classification results from Luo et al. (17) while simply replacing the label questionable with uncertain.

TABLE 1 A summary of the encompassed data.

Languages	Sources	Labels		
		True	False	Uncertain
English	Patwa et al. (20)	830	830	830
Chinese	Luo et al. (17)	335	276	409

4 Methods and results

4.1 Word frequency analysis

Weiciyun (22) is utilized in this section to conduct word frequency analysis for both English and Chinese records. It serves as a practical and user-friendly online tool for generating word clouds and visualizing text data. Before analysis, the built-in language-specific tokenization and stopword removal techniques provided by Weiciyun are leveraged to yield clean and meaningful text data. Afterward, content filtration based on part-of-speech is applied to retain only nouns, gerunds, and proper nouns. In terms of English text, only content with a word length of at least 3 and a frequency of at least 2 is selected. Similarly, for Chinese text, content with a character length of at least 2 is chosen. Finally, the 35 most frequent words are presented and they are illustrated with font size scaled to their frequencies while the detailed word frequencies of these words can be found in Table 2. To ensure translation consistency and reduce subjectivity, the word clouds maintain the original Chinese characters while providing a reference translation in Appendix 1 as needed.

The word clouds of all records in English and Chinese are presented in Figure 1. Firstly, it is noteworthy that the most frequently mentioned terms in both languages are the same, including "virus" (病毒), "pandemic" (疫情), "patient" (患者), and so on. Secondly, the term "mask" (口罩) is mentioned in both languages but holds greater prominence in the Chinese word cloud. Thirdly, the name "Wuhan" (武汉), which corresponds to the initial epicenter of the COVID-19 outbreak in China, appears in larger font size in the Chinese word cloud, while no specific city-related word is present in the English cloud. Fourthly, the term "death" appears with greater frequency in the English data than in the Chinese records where it is noticeably absent. Finally, the individual most frequently mentioned in English is President Donald Trump, whereas, in Chinese, it is Zhong Nanshan (钟南山), an esteemed academician in the field of healthcare.

The word clouds presented in Figure 2 categorize records into three groups in both English and Chinese. The true or false labeled groups primarily consist of common terms, which are predominantly derived from the expertise of healthcare professionals. These terms revolve around virus transmission methods, prevention measures, and treatment approaches. On the other hand, the uncertain group encompasses a diverse range of terms. Within this group, both English and Chinese records demonstrate an awareness of regional considerations. Notably, the Chinese word cloud places a greater emphasis on specific locations such as "Wuhan" (武汉), "Beijing" (北京), "Shanghai" (上海), "Canton" (广州), and "Chengdu" (成都). In

10.3389/fpubh.2023.1281259

TABLE 2 Word frequency of the 35 most frequent words displayed in Figures 1, 2.

All records				Records labeled as true				Records labeled as false				Records labeled uncertain			
English	W.F.	Chinese	W.F.	English	W.F.	Chinese	W.F.	English	W.F.	Chinese	W.F.	English	W.F.	Chinese	W.F.
Covid	833	病毒	185	Covid	476	病毒	71	Coronavirus	353	病毒	71	Cases	336	肺炎	84
Coronavirus	617	肺炎	179	People	141	口罩	60	People	91	肺炎	59	Covid	274	武汉	48
Cases	430	口罩	110	Spread	126	肺炎	36	Covid	83	口罩	32	Coronavirus	162	病毒	43
People	319	疫情	56	Coronavirus	102	患者	28	Virus	79	疫情	13	Tests	123	疫情	39
Health	177	武汉	55	Health	97	消毒剂	22	Trump	62	患者	11	Deaths	103	医院	20
Tests	159	患者	54	Risk	87	症状	17	Pademic	51	美国	10	Number	101	美国	20
Spread	147	美国	30	Cases	76	医用	16	Cure	51	酒精	10	People	87	中国	19
Deaths	145	钟南山	26	Face	73	飞沫	14	President	49	钟南山	9	States	85	口罩	18
Virus	138	消毒剂	26	Others	67	建议	11	Vaccine	47	疫苗	8	India	78	钟南山	16
Testing	137	酒精	25	Testing	63	风险	10	Video	42	武汉	7	Today	71	患者	15
Pademic	125	疫苗	24	Symptoms	62	酒精	10	Government	38	大蒜	6	Testing	64	北京	15
Vaccine	119	医院	22	Patinets	52	疾病	10	Corona	37	大量	6	State	62	上海	14
Number	119	中国	22	Virus	50	证据	10	China	37	病人	5	Indiafightscorona	59	意大利	14
States	116	病人	20	Pandemic	47	感染者	9	News	33	日本	5	Health	47	病人	11
India	111	症状	20	Masks	46	人群	9	Health	33	抗体	4	Report	44	疫苗	10
Patients	109	医用	18	Mask	46	儿童	8	Claims	32	院士	4	Vaccine	43	病例	9
Risk	107	风险	17	Care	44	居家	8	Chinese	31	医生	4	Rate	42	湖北	9
Face	87	北京	17	Hands	43	通风	8	Masks	31	病毒感染	4	Case	41	人员	9
Test	87	人员	17	Use	42	人员	8	Bill	28	空气	4	Nigeria	39	成都	9
Trump	84	病例	16	Contact	42	物品	7	World	28	白酒	3	Data	38	院士	8
State	83	意大利	16	Distancing	40	效果	7	Gates	27	防病毒	3	Lakh	38	入境	7
Masks	83	建议	15	Home	39	传染性	7	Flu	26	小时	3	Lockdown	37	医生	7
Days	82	上海	14	CDC	39	人类	7	Novel	25	病情	3	Day	35	视频	7
Today	81	飞沫	14	Measures	37	距离	7	Donald	25	中国	3	Patients	35	全国	7
Symptoms	81	抗体	13	Cloth	35	核酸检测	6	Being	24	流鼻涕	3	Days	32	全部	6
Indiafightscorona	75	院士	13	Disease	34	疫苗	6	India	24	纸尿裤	3	Test	32	阳性	6
Others	73	感染者	13	Test	34	动物	6	Claim	23	气溶胶	3	Yesterday	31	员工	6
Home	72	阳性	12	Treatment	30	食品	6	Outbreak	22	二氧化氯	3	Week	31	印度	6
CDC	71	核酸检测	12	Days	29	情况	6	Home	22	消毒剂	3	First	30	国家	5
Government	71	医生	11	Vaccine	29	传播者	6	Lockdown	22	牛羊肉	3	Million	30	物资	5
Data	70	疾病	11	Data	29	重症	6	Patients	22	喉咙	3	Recoveries	29	酒精	5
Lockdown	70	湖北	11	Infection	29	手部	6	Disease	21	肥皂	3	Coronavirusupdates	28	特朗普	5
Care	69	空气	11	Countries	29	手套	6	Test	21	食品	3	Pandemic	27	风险	5
Video	68	证据	11	Deaths	27	传染病	6	Days	21	食用	3	Isolation	27	广州	5
Case	67	人类	10	Person	27	紫外线	5	Message	20	瘟疫	2	Numbers	27	医疗	5

W.F., Word Frequency.

FIGURE 1
Word clouds for all records.

contrast, the English word cloud labeled as uncertain indicates a temporal focus by frequently including terms like "Today," "Yesterday," "Days," and "Week." It is worth mentioning that these time-related terms are not explicitly included in the Chinese word cloud.

4.2 Topic clustering analysis

In this section, the Latent Dirichlet Allocation (LDA) topic model is implemented to uncover hidden topics and thematic structures from both English and Chinese records. LDA is a widely adopted technique in the field of natural language processing, wherein documents are represented as stochastic mixtures across latent topics, and each topic is characterized by a distribution over words (23). For enhanced comprehension of the clustered topics, we employ the LDAvis package (24) to visualize the results using multidimensional scale analysis. We set the initial range of topic numbers to [1, 15], and the final determination of the optimal number of topics relies on the highest coherence score. The step size is retained as 1, while α and β are maintained at their default values. Furthermore, language-specific tokenization and stopword removal techniques are employed to mitigate the influence of text analysis when applying LDA to analyze different languages. For English text, whitespace-based tokenization is employed, and the widely recognized Chinese word segmentation tool Jieba is utilized for Chinese tokenization. The built-in function from the Natural Language Toolkit (NLTK) library in Python is leveraged to access a collection of stopwords specifically for English, whereas the widely used cn_stopwords.txt file is applied to remove stopwords from Chinese text. Finally, in line with sub-section 4.1, the original Chinese characters are preserved in the visualization graphs, supplemented with a reference translation provided in Appendix 2.

The visualization graphs of all records in English and Chinese are presented in Figure 3. Firstly, the number of clustered topics in the English records is significantly fewer compared to the Chinese records. Specifically, there are only 4 topics identified in the English records, whereas the Chinese records encompass 13 topics. Secondly, the English topics are mutually exclusive with no overlap. The proximity between Topic 1 and Topic 2 is high, while the remaining topics exhibit considerable dissimilarity. Conversely, in the visualization graph of the Chinese records, the topics demonstrate interconnectedness. Notably, Topic 2 overlaps with Topic 9, as does Topic 8 with Topic 11. Thirdly, Topic 1 stands out in the English records as it covers a significant portion of the tokens, specifically 35.2% in the top 30 most relevant terms. On the other hand, Topic 1 has a comparatively smaller presence in the Chinese records, accounting for only 11% of the tokens in the top 30 most relevant terms. Its size is not as noticeable when compared to Topic 2 and Topic 3, where the difference is not considered significant. Finally, there are shared terms that appear in the top 30 most relevant terms of Topic 1 in both languages, indicating a mutual focus from both sides.

The visualization graphs in Figure 4 categorize records into three groups in both English and Chinese. The annotation of each visualization graph remains the same as shown in Figure 3. Due to space limitations, they are not included in Figure 4. Firstly, the pattern of topic numbers remains consistent across the groups labeled as true and uncertain. However, in the group labeled as false, the English records show a significantly larger number compared to the Chinese records. Secondly, within the groups labeled as true, the percentage of tokens in the top 30 most relevant terms of Topic 1 is similar in both languages while there exists a difference of more than 10% in the other two groups. Finally, the groups labeled as true or false primarily consist of common terms in the top 30 most relevant terms in both languages. Nevertheless, the uncertain group encompasses a diverse range of terms. This observation further supports the conclusion mentioned in sub-section 4.1.

4.3 Sentiment analysis

Monkeylearn (25) is utilized in this section to conduct sentiment analysis on English records. The platform offers a user-friendly graphical interface that enables users to create personalized text classification and extraction analyzes by training machine learning models. In the analysis of Chinese records, ROST_CM6 (26), a widely used Chinese social computing platform, is employed to generate the results. ROST_CM6 enables various text analyzes, including microblog, chat, and web-wide analyzes. It is important to note that Monkeylearn generated multiple emotions for 145 instances due to the length or complexity of certain English records. To maintain consistency, these instances were manually annotated by three annotators, and the emotional tone was determined based on the majority agreement. Finally, each record was broken down into positive, negative, or neutral categories.

FIGURE 2
Word clouds for records classified into three groups.

The pie charts and bar charts in Figure 5 present the sentiment proportion of all records in English and Chinese. It indicates that over 50% of the information, in both English and Chinese, is characterized as negative. Specifically, English records have a negative proportion of 59.96%, while Chinese records have a negative proportion of 50.71%. This implies that regardless of the language system, over half or more of the infodemic being disseminated to the public carries a negative tone. In addition, the distribution of positive and neutral information differs between the two language systems. Within the Chinese records, there is a balance between positive (25.69%) and neutral (23.60%) information. On the other hand, in the English records, the proportion of positive information exceeds that of neutral information significantly, with 31.12% being positive and only 8.92% being neutral. These findings suggest that individuals within the English language system tend to adopt a more positive attitude when confronted with the infodemic during the COVID-19 pandemic. Conversely, individuals in the Chinese language system lean toward a more neutral and conservative stance.

The pie charts and bar charts in Figure 6 categorize records into three groups in both English and Chinese. In the true group, the sentiment proportions for Chinese records have remained relatively stable compared to the results displayed in Figure 5. However, for English records, there has been an increase in the proportions of positive and neutral information. In the false group, it is observed that the proportions of negative information have increased in both languages and remain relatively consistent. Additionally, false

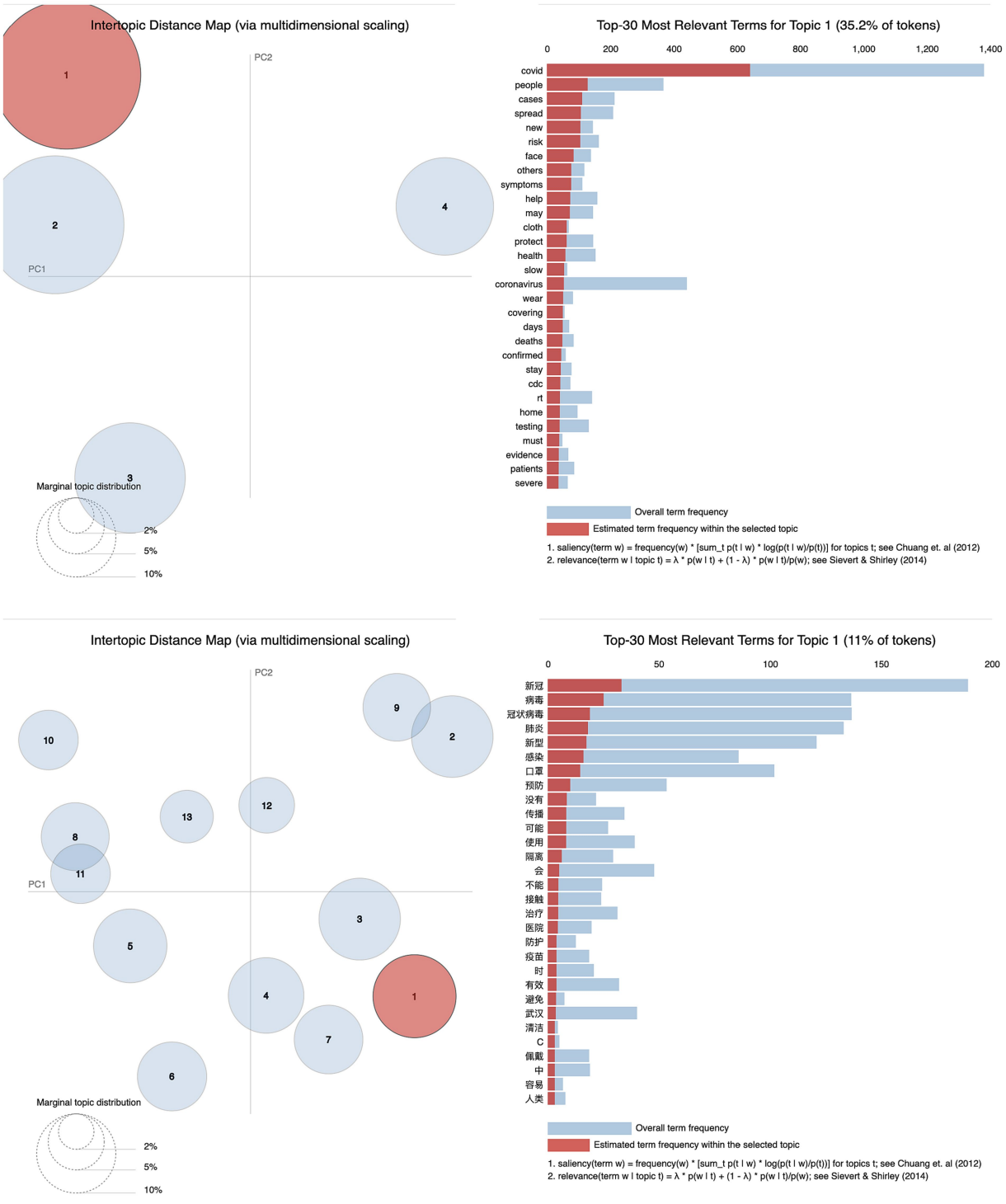

FIGURE 3
Visualization graph of LDA topic modeling for all records.

records exhibit the highest proportion of negative sentiment among the three groups. Moving on to the uncertain group, the sentiment proportions for English records have not shown significant changes compared to the false group. However, for Chinese records in the uncertain group, the proportion of negative sentiment has decreased, resulting in a relatively balanced distribution of the three sentiment categories.

5 Discussion

Regarding word frequency analysis, the distinctions between the English and Chinese word clouds reflect some unique perspectives. Firstly, the term "mask" holds particular significance in the Chinese context, reflecting the country's proactive approach to mask-wearing as a preventive measure against the virus. This cultural aspect is not as

10.3389/fpubh.2023.1281259

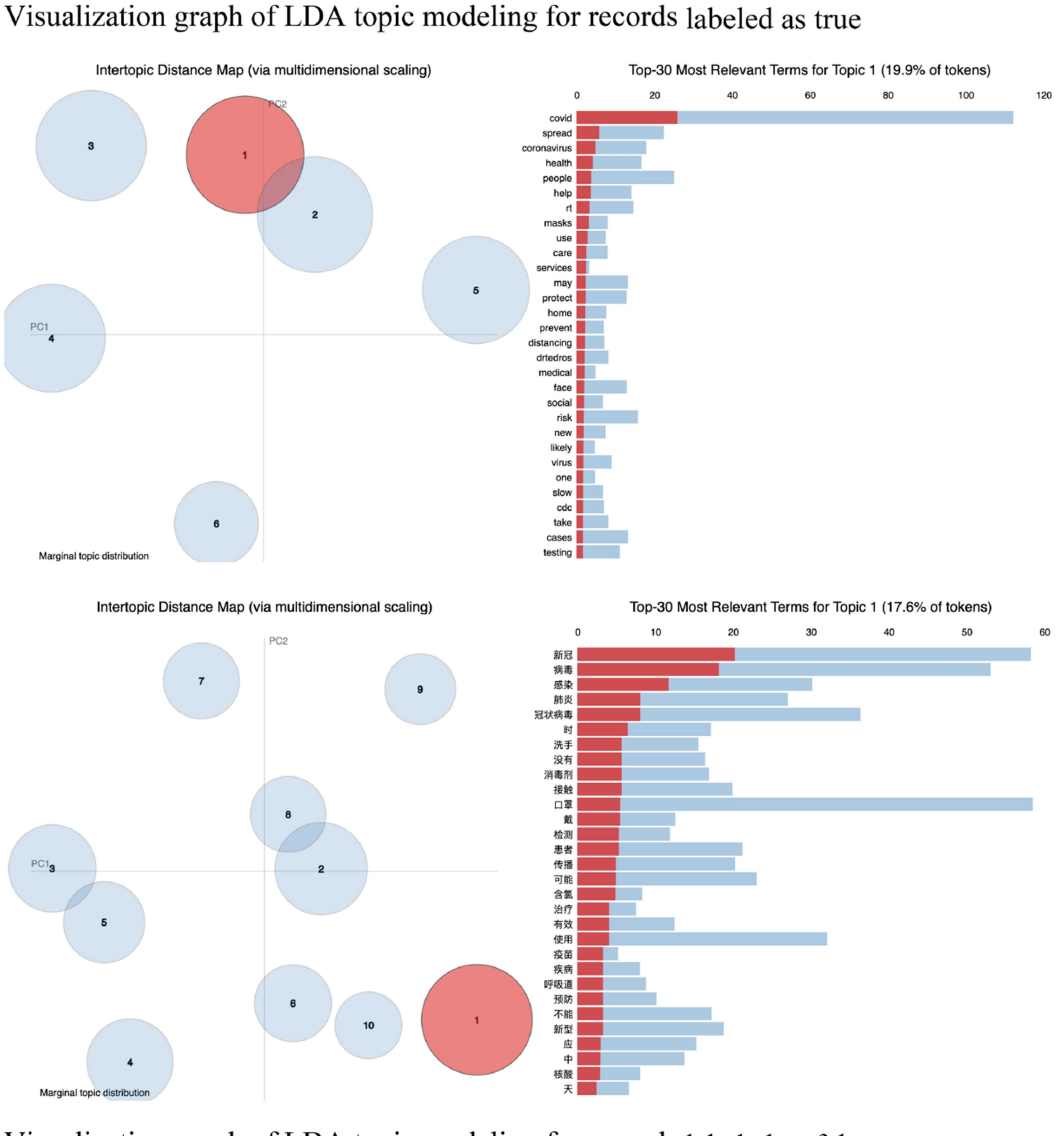

FIGURE 4 (Continued)

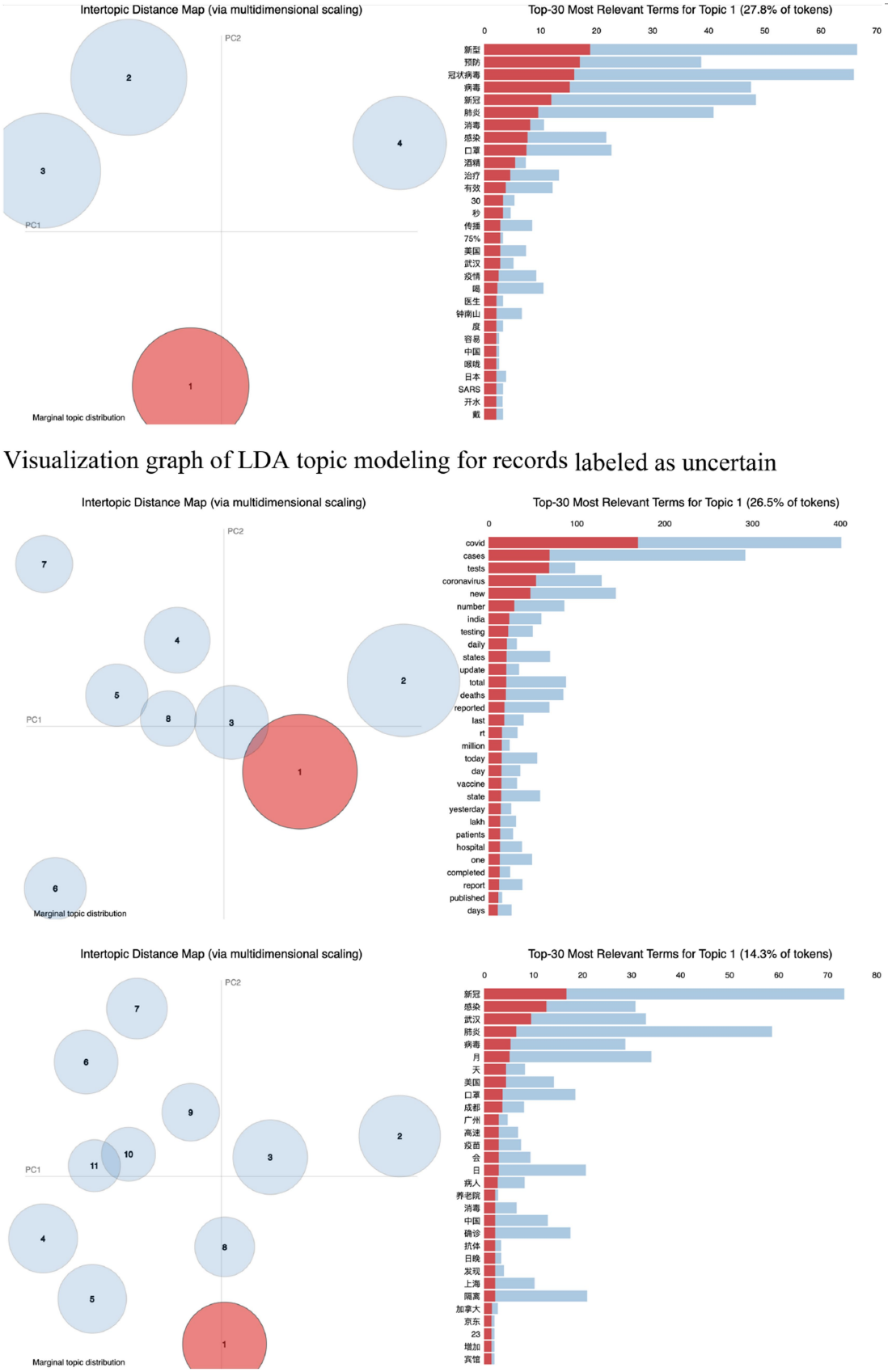

FIGURE 4
Visualization graph of LDA topic modeling for records classified into three groups.

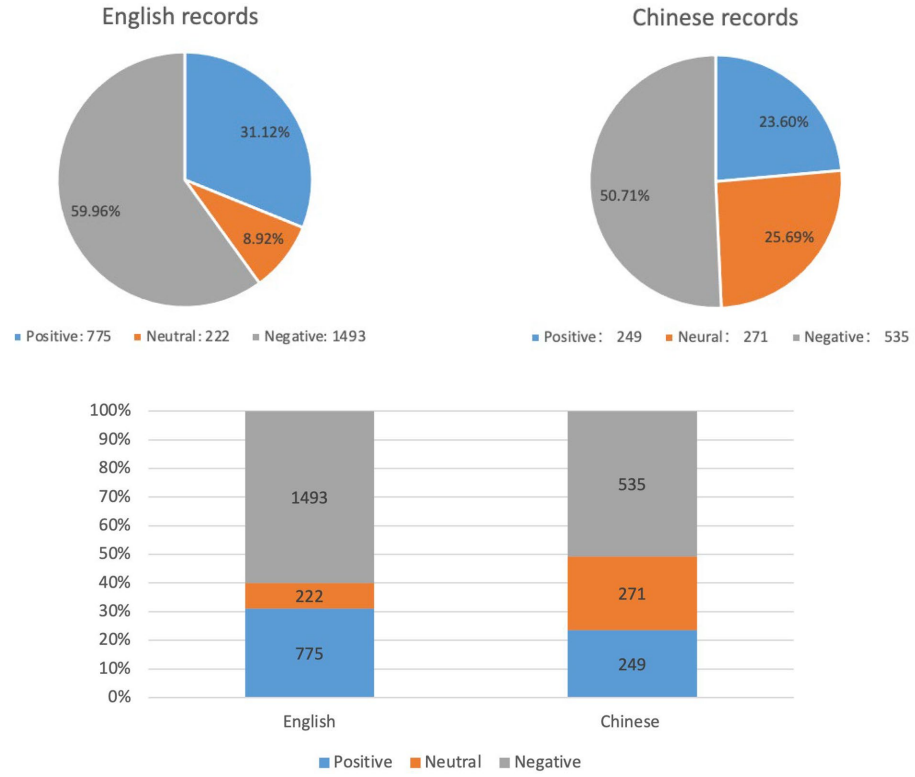

FIGURE 5
Sentiment analysis results for all records.

prominent in the English word cloud, indicating potential differences in the adoption and perception of this protective measure. Secondly, the variation in the frequency of the term "death" between the English and Chinese word clouds sheds light on the different tones and focuses within each language. The higher occurrence in the English cloud may indicate a greater emphasis on the global loss of life and the severity of the situation, whereas its absence in the Chinese cloud might suggest a more limited or sensitive discussion surrounding this aspect. Thirdly, the individuals most frequently mentioned, President Donald Trump in English and Zhong Nanshan, an esteemed healthcare academician in Chinese, further exemplify the contrasting perspectives. It highlights the significance of political figures in English discussions and the recognition of medical experts and authoritative voices in the Chinese discourse. Finally, the region-specific emphasis in the Chinese cloud and the temporal focus in the English cloud showcase the nuances and contextual factors shaping the discussions in each language. These city names suggest a focus on regional impact and potential localized concerns within China while these time-related terms reflect the need to stay updated with real-time information within English conversations.

The topic clustering analysis highlights the distinct characteristics and priorities within the English and Chinese discussions on COVID-19. Firstly, the English records have a lower number of clustered topics compared to the Chinese records in most cases. This discrepancy suggests that the English discussions on COVID-19 may exhibit a more focused and limited scope while the Chinese records suggest a wider range of perspectives and a more nuanced understanding of various aspects. Secondly, the group labeled as false stands out as an exception to this pattern, with the English records displaying a significantly larger number of clustered topics compared to the Chinese records. This stark difference may indicate a higher prevalence of diverse false narratives and misinformation spread across various sources within the English language. Thirdly, the presence of shared terms in the top 30 relevant terms of Topic 1 signifies a shared focus between both languages, particularly within the group labeled as true and false. This common attention highlights the significance of specific themes or concerns in the global discourse surrounding COVID-19, transcending linguistic and cultural boundaries. Finally, there are noticeable differences in the top 30 relevant terms of Topic 1 within the uncertain group between the Chinese and English languages. These variations emphasize disparities in how uncertain records are conceptualized and discussed within the Chinese and English language communities, which can likely be attributed to variances in cultural, linguistic, and contextual factors.

When it comes to sentiment analysis, the comparative results in English and Chinese records offer valuable insights into emotional trends. Firstly, it is evident that in most cases, over 50% of the information in both languages skews toward negativity, indicating a prevalent negative sentiment in the collected infodemic data. This is likely influenced by the nature of the discussed topics, the tone employed, and the general sentiment of those generating the records. Secondly, English records typically demonstrate a notably higher proportion of positive sentiment with a substantial margin compared to their Chinese counterparts. This disparity can be attributed to various factors, such as cultural contexts, linguistic nuances, or even the diverse user demographics associated with each language. Thirdly,

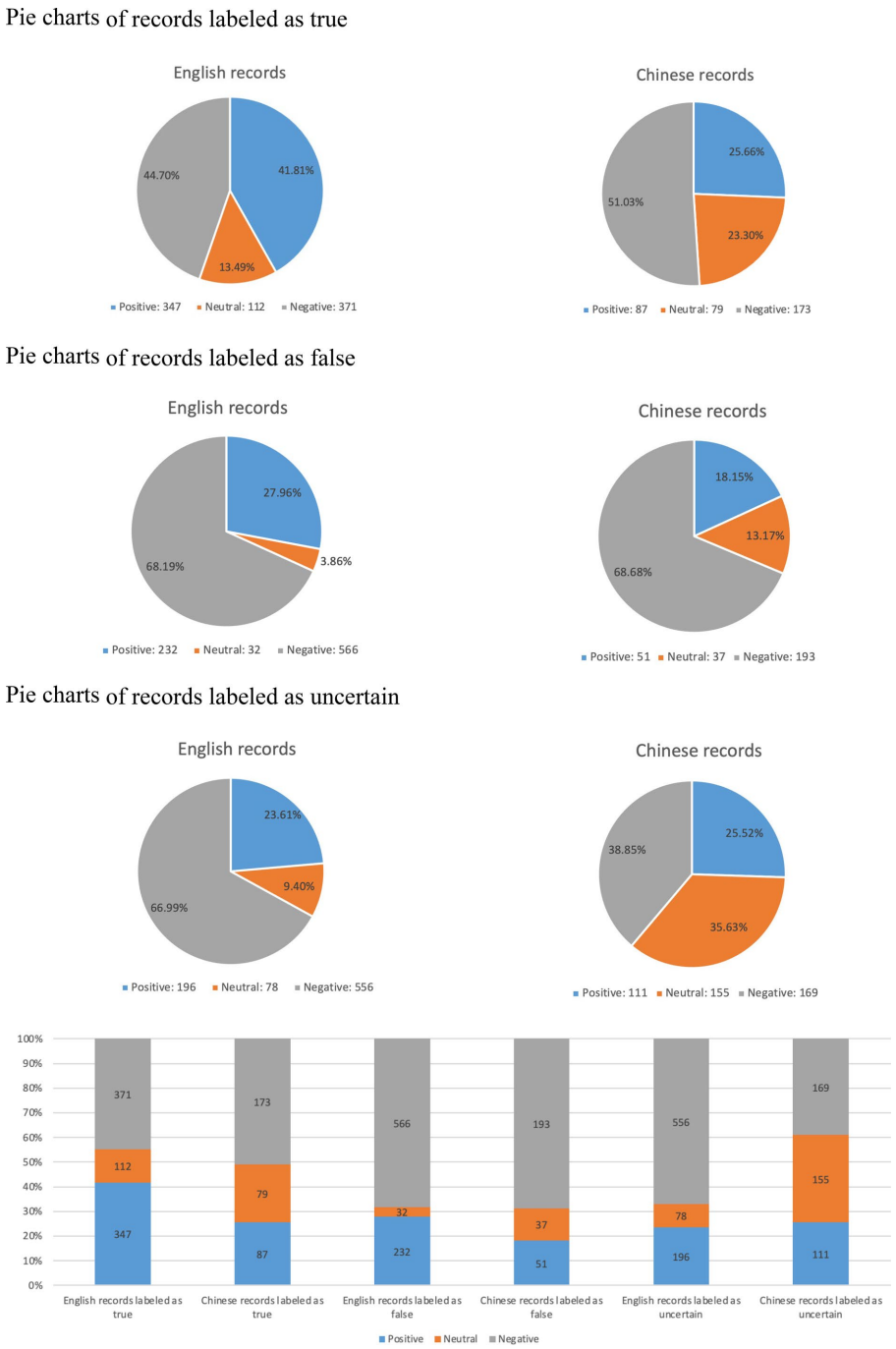

FIGURE 6
Sentiment analysis results for records classified into three groups.

false records consistently manifest the highest proportion of negative sentiment among the three groups in both languages. This observation implies a strong association between misinformation and the generation of negative sentiment among readers. As a result, there is a critical need to actively combat the spread of false records since misleading content not only deceives individuals but also significantly impacts their emotional well-being. Finally, English records in the uncertain group display a nearly identical proportion of negative sentiment, while Chinese records show a decline in negative sentiment. This divergence implies a potential shift toward increased clarity or certainty in the Chinese records classified as uncertain, suggesting that Chinese sources may provide more conclusive or reliable content in this group compared to their English counterparts.

There are several limitations to this study. Firstly, most data collected from the two datasets were obtained from authoritative and representative channels that specifically focus on gathering and presenting valuable information related to popular online topics. However, relying on these sources can introduce biases as the selection of sources and editorial decisions may influence the representation of different perspectives and prioritize certain

viewpoints. Secondly, the English records' labels are determined by invited healthcare workers' judgments using a majority agreement methodology. This approach can lead to variations in labeling due to individual differences in interpretation, knowledge, and biases. The absence of clear guidance or standardized criteria for healthcare workers further contributes to potential inconsistencies in labeling decisions. Thirdly, the retention rate for English data sourced from (20) is low. After manual classification, only 830 records were included in the real group, which is the smallest group out of the three. Ultimately, a total of 2,490 records were retained for equal distribution among each group. Considering the initial count of 6,420 records, the overall retention rate is only 38.78%.

6 Conclusion and future works

This paper presents a comparative analysis of the COVID-19 infodemic in English and Chinese languages, utilizing textual data extracted from social media platforms. Firstly, to ensure a balanced representation and a fair assessment, two infodemic datasets were introduced through the augmentation of previously collected social media textual data with annotations provided by healthcare workers. Secondly, word frequency analysis was conducted, revealing the 35 most frequently occurring infodemic words in both English and Chinese. This comparison offers valuable insights into the prevalent discussions surrounding the COVID-19 infodemic. Thirdly, topic clustering analysis was performed to identify thematic structures present in both languages. This exploration provides a deeper understanding of the primary topics related to the COVID-19 infodemic within each language context. Finally, sentiment analysis was carried out to evaluate the distribution of positive, neutral, and negative sentiments. This investigation helps comprehend the overall emotional tone associated with COVID-19 information shared on social media platforms in the English and Chinese languages.

In the future, we intend to conduct a study considering the contextual factors. The two proposed datasets in this paper solely consist of original posts from social media, excluding reposts and replies. Additionally, certain records were sourced from official handbooks, authoritative webpages, and fact-verification websites, which lack propagation information. Therefore, the first issue is to collect the user social engagements from the social platform based on infodemic content, including the timestamp of who engages in the records dissemination process. The second line of interest is to conduct a comprehensive understanding of how infodemic spreads within the online community by effectively analyzing users' interactions and their engagement records. Finally, an in-depth analysis will be implemented to seek valuable insights into the mechanisms and dynamics of infodemic propagation, aiming to uncover why and how infodemics occur.

Data availability statement

The original datasets can be found: https://www.dropbox.com/scl/fo/1qug1snyu49bsiuj53hty/h?rlkey=kew7715ubl83jhmtvcroxv7uj&dl=0. Further inquiries can be directed to the corresponding author.

Author contributions

JL: Conceptualization, Writing – original draft. DP: Visualization, Writing – review & editing. LS: Writing – review & editing, Methodology. DEB: Supervision, Writing – review & editing. XL: Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work is supported by the Natural Science Foundation of Chongqing, China (Grant No. CSTB2023NSCQ-MSX0391), the National Natural Science Foundation of China (Grant No. 72104016), the R&D Program of the Beijing Municipal Education Commission (Grant No. SM202110005011), and the Guangxi Key Laboratory of Trusted Software (Grant No. KX202315).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1281259/full#supplementary-material

References

1. Zarocostas J. How to fight an infodemic. *Lancet*. (2020) 395:676. doi: 10.1016/S0140-6736(20)30461-X

2. Xu J, Liu C. Infodemic vs. pandemic factors associated to public anxiety in the early stage of the COVID-19 outbreak: a cross-sectional study in China. Frontiers. *Public Health*. (2021) 9:723648. doi: 10.3389/fpubh.2021.723648

3. Statista. (2023). Available at. www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet.

4. Sanaullah AR, Das A, Das A, Kabir MA, Shu K. Applications of machine learning for COVID-19 misinformation: a systematic review. *Soc Netw Anal Min*. (2022) 12:94. doi: 10.1007/s13278-022-00921-9

5. Glazkova A, Glazkov M, Trifonov T. g2tmn at constraint@ aaai2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection In: S Akhtar, editor. *International workshop on combating online hostile posts in regional languages during emergency situation*. Cham: Springer International Publishing (2021). 116–27.

6. Chen B, Chen B, Gao D, Chen Q, Huo C, Meng X, et al. Transformer-based language model fine-tuning methods for COVID-19 fake news detection In: T Chakraborty, K Shu, HR Bernard, H Liu and MS Akhtar, editors. *Combating online hostile posts in regional languages during emergency situation: First international workshop, CONSTRAINT 2021, collocated with AAAI 2021, virtual event, February 8, 2021, revised selected papers 1*. Berlin: Springer International Publishing (2021). 83–92.

7. Paka WS, Bansal R, Kaushik A, Sengupta S, Chakraborty T. Cross-SEAN: a cross-stitch semi-supervised neural attention model for COVID-19 fake news detection. *Appl Soft Comput*. (2021) 107:107393. doi: 10.1016/j.asoc.2021.107393

8. Chen MY, Lai YW. (2022). *Using fuzzy clustering with deep learning models for detection of COVID-19 disinformation*. Transactions on Asian and Low-Resource Language Information Processing.

9. Liu J, Chen M. (2023) *COVID-19 fake news detector*. In: 2023 international conference on computing, networking and communications (ICNC). IEEE, pp. 463–467.

10. Gupta A, Li H, Farnoush A, Jiang W. Understanding patterns of COVID infodemic: a systematic and pragmatic approach to curb fake news. *J Bus Res*. (2022) 140:670–83. doi: 10.1016/j.jbusres.2021.11.032

11. Wan M, Su Q, Xiang R, Huang CR. Data-driven analytics of COVID-19 'infodemic'. *Int J Data Sci Anal*. (2023) 15:313–27. doi: 10.1007/s41060-022-00339-8

12. Zhao J, Fu C, Kang X. Content characteristics predict the putative authenticity of COVID-19 rumors. *Front Public Health*. (2022) 10:920103. doi: 10.3389/fpubh.2022.920103

13. Zhou L, Tao J, Zhang D. Does fake news in different languages tell the same story? An analysis of multi-level thematic and emotional characteristics of news about COVID-19. *Inf Syst Front*. (2023) 25:493–512. doi: 10.1007/s10796-022-10329-7

14. Murayama T. Dataset of fake news detection and fact verification: a survey. *arXiv*. (2021) 2021:03299. doi: 10.48550/arXiv.2111.03299

15. Cheng M, Wang S, Yan X, Yang T, Wang W, Huang Z, et al. A COVID-19 rumor dataset. *Front Psychol*. (2021) 12:644801. doi: 10.3389/fpsyg.2021.644801

16. Haouari F, Hasanain M, Suwaileh R, Elsayed T. ArCOV19-rumors: Arabic COVID-19 twitter dataset for misinformation detection. *arXiv*. (2020) 2020:08768. doi: 10.48550/arXiv.2010.08768

17. Luo J, Xue R, Hu J, El Baz D. Combating the Infodemic: a Chinese Infodemic dataset for misinformation identification. *Healthcare*. (2021) 9:1094. doi: 10.3390/healthcare9091094

18. Kim J, Aum J, Lee S, Jang Y, Park E, Choi D. FibVID: comprehensive fake news diffusion dataset during the COVID-19 period. *Telematics Inform*. (2021) 64:101688. doi: 10.1016/j.tele.2021.101688

19. Dharawat AR, Lourentzou I, Morales A, Zhai C. (2020). *Drink bleach or do what now? Covid-hera: A dataset for risk-informed health decision making in the presence of COVID 19 misinformation*.

20. Patwa P, Sharma S, Pykl S, Guptha V, Kumari G, Akhtar MS, et al. (2021). *Fighting an infodemic: Covid-19 fake news dataset. In combating online hostile posts in regional languages during emergency situation: first international workshop, CONSTRAINT 2021, collocated with AAAI 2021, virtual event, February 8, 2021, revised selected papers 1*. Berlin: Springer International Publishing, pp. 21–29.

21. CONSTRAINT. (2021). Available at: https://constraint-shared-task-2021.github.io.

22. (n.d.). Available at: https://www.weiciyun.com.

23. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res*. (2003) 3:993–1022.

24. Sievert C, Shirley K, Davis L (2014). *A method for visualizing and interpreting topics*. In: Proceedings of Workshop on Interactive Language Learning, Visualization, and Interfaces, Association for Computational Linguistics, pp. 63–70.

25. Monkeylearn. (n.d.). Available at: https://monkeylearn.com.

26. Zhang B, Jiang Y, Zhou J. Analysis of the contents of the "draft of the preschool education law of the People's republic of China (draft for solicitation of comments)" based on the ROST CM6. 0 content mining system. *Chin Educ Soc*. (2021) 54:1–20. doi: 10.1080/10611932.2021.1949208