Proceedings of Machine Learning Research vol 245:1–10, 2024 2024 International Conference on Machine Learning and Intelligent Computing

Characterization Study of Online Public Opinion Based on Natural Language Processing with Weibo Data

Zhiyuan An ANZHIYUAN@CAU.EDU.CN
China Agricultural University, Beijing 100083, China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

In the summer of 2022, "Ice cream assassin" has emerged as a prominent buzzword on the Chinese Internet. With its vast user base of 582 million monthly active users, Sina Weibo serves as an ideal platform for analyzing information disseminated within its ecosystem. This platform not only enables us to discern prevailing public sentiment but also facilitates governmental efforts in shaping and regulating the public opinion landscape. This study encompasses a collection of 60,228 Weibo pertaining to "ice cream assassin" posted on Sina Weibo. By employing sentiment analysis algorithm based on Bert's fine-tuned model, we analyze temporal shifts in emotional trends, summarize changes in public sentiment, and categorize variations in popularity levels. Furthermore, through semantic network analysis, we identify two distinct thematic segments within this realm of public opinion. This study has important implications for assessing the impact of online public opinion on the economy and even society.

Keywords: Natural Language Processing; Sentiment Analysis; Text Infomation Mining; Topic Analysis.

1. Introduction

In recent years, social media platforms have become an increasingly important arena for public discourse, rapidly amplifying and disseminating news and events. One such incident that gained widespread attention on China's Sina Weibo social network was the "ice cream assassin" phenomenon. This referred to the practice of some vendors mixing high-end ice cream products with cheaper alternatives, leading unsuspecting customers to be shocked by the significantly higher prices upon checkout. In the realm of big data, social media platforms not only function as conduits for public event information dissemination but also often obscure the emotional stance of in-formation publishers (Zhao et al., 2022). If left unregulated and allowed unrestricted growth, this phenomenon is likely to exert immense pressure on public opinion and have an ad-verse impact on long-term societal stability (Zhang et al., 2024).

This study aims to comprehensively evaluate the characteristics of the online dis-course surrounding the "ice cream assassin" incident on the Sina Weibo platform. Employing a combination of sentiment analysis algorithm based on Bert's fine-tuned model and topic extraction via TF-IDF techniques, the research seeks to uncover the emotional tone, thematic focus, and transmission dynamics of this widely discussed social event. the fine-tuning of BERT for sentiment analysis tasks offers significant advantages in terms of enhancing prediction accuracy, demonstrating flexibility across different languages and modalities, and improving fine-grained sentiment and aspect-based analyses. These advantages position BERT as a powerful tool for ad-vancing the field of sentiment analysis.

The findings contribute to a deeper understanding of how consumer-related controversies unfold and propagate on major social media channels in China. The implications of this research extend beyond the specific "ice cream assassin" case, offering valuable perspectives for businesses, policymakers, and communication scholars on navigating the complexities of online public opinion formation and crisis management in the digital age (Ravi and Ravi, 2015).

2. Literature Review

The proliferation of social media platforms has transformed the landscape of public discourse, allowing for the rapid dissemination and amplification of news and events (Peng et al., 2023). Scholars have extensively studied the mechanisms and implications of social media's influence on public opinion formation. Existing research has highlighted the complex interplay between online discussions, emotional responses, and the diffusion of information on social media (Thelwall et al., 2012).

These two theories are often used to examine the impact of social media platforms on the e-commerce economy. Agenda-Setting Theory: Your findings may indicate that topics heavily discussed online become important decision-making points for consumers (Han et al., 2024). This aligns with the Agenda-Setting Theory, which suggests that media shapes the public's reality by emphasizing what topics are important. In the context of e-commerce, products or services that receive more positive attention online may see a corresponding increase in sales (Damberg et al., 2022). Spiral of Silence Theory: This theory posits that people are less likely to express their opinions if they perceive them to be in the minority. In your study, this could explain a potential skew in online reviews or social media comments, where negative reviews might be underrepresented if the overall sentiment is overwhelmingly positive (Burnett et al., 2022). This discrepancy can have a significant impact on consumer perceptions and thus on economic outcomes (Lin et al., 2022).

In the context of consumer-related controversies, studies have examined how inci-dents involving product quality, pricing, or business practices can quickly become the focus of intense online scrutiny and debate. The "ice cream assassin" case represents a prime example of such a socially charged event, where the perceived deception of customers has the potential to elicit strong emotional reactions and shape public perceptions.

3. Research Method

In this study, firstly, a crawler program was used to obtain relevant data on Weibo platform, followed by data cleaning. Next, an algorithm based on the Bert fine-tuning model is used to assess the trend of emotional color changes. TF-IDF was used to assess the topic distribution of related topics. The research framework is shown in Figure 1.

3.1. Data Acquisition and Preprocessing

The data collection process involved using the official API provided by Weibo and the crawler program to simulate user searches for keywords on Weibo, retrieving 50 microblogs at a time. Specifically focusing on the keyword "ice-cream assassin," the study collected a total of 60,228 related pieces of information posted on Sina Weibo between June 13, 2022, and July 27, 2022, including user-related details, content body, posting time, and other key information. Since the body of the crawled Weibo contains a large number of useless characters, such as "@XXX" and "#XXX",

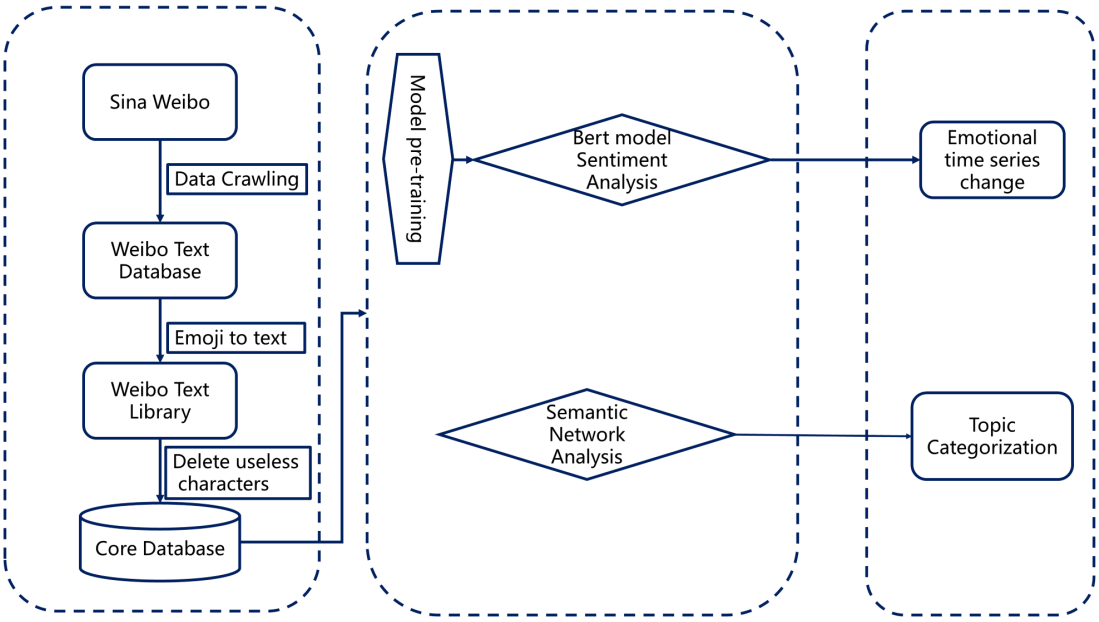

Figure 1: Research framework

the Chinese text is regularized to remove the non-Chinese characters and the related content after the @ symbol, so as to achieve preliminary cleaning. The distribution of data is shown in Figure 2.

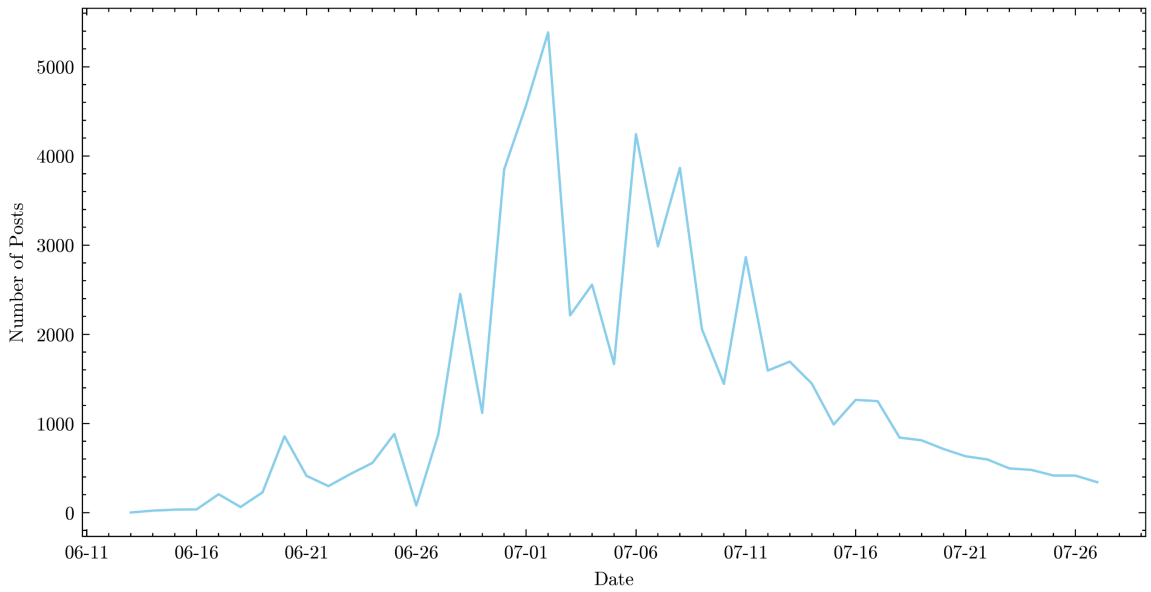

Figure 2: Distribution of data

3.2. Sentiment Analysis

In the domain of Chinese sentiment analysis, the application of a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) model follows a systematic approach. Initially, a pretrained BERT model, which has already been trained on extensive language corpora to grasp basic language patterns, is prepared.

The choice between naive Bayes and BERT for sentiment analysis (or any other text classification task) ultimately depends on the specific requirements and constraints of the application. Naive Bayes offers simplicity, efficiency, and adaptability, making it a suitable choice for small-scale or resource-limited scenarios. BERT, on the other hand, excels at capturing complex linguistic relationships and can leverage large-scale language models to achieve state-of-the-art performance, particularly in more sophisticated or domain-specific tasks. Careful consideration of the trade-offs between these approaches is crucial in selecting the most appropriate solution for a given problem.

While the BERT fine-tuning model is highly regarded for its effectiveness in various sentiment analysis tasks, it faces distinct challenges when applied to short Chinese texts. The architecture's token processing limit, designed for a maximum of 510 to-kens, necessitates truncation methods that may compromise sentiment capture in longer sequences, impacting overall effectiveness. Additionally, the irregular structure and feature sparseness typical of Chinese short comments pose significant classification challenges, requiring more sophisticated models and techniques to accurately grasp emotional tendencies. Furthermore, online Chinese buzzwords, which often flout syntactic norms and lack clear semantic structures, necessitate deeper contextual analysis for accurate sentiment polarity determination). The linguistic character-istics of Chinese also demand effective word segmentation in preprocessing, which directly influences model performance; a limitation when not using a character-level approach.

3.3. Sentiment Analysis

The semantic network is mainly based on the TF-IDF algorithm. TFIDF, also called word frequency inverse document frequency, combines the formula for calculating word frequency and the formula for calculating inverse document frequency (Aizawa, 2003). The solution is to distinguish the importance of these words to a document when there are different words with the same word frequency in a document (Qin, 2015). The key advantages of using TF-IDF for topic modeling are its simplicity and interpretability, computational efficiency, and robustness to noise and variations in text (Simon et al., 2023).

$$\begin{aligned} \mathrm{TF}-\mathrm{IDF}(t,d,D) = & \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in docu-ment } d} \\ & \times \log\left(\frac{\text{Total number of documents in corpus } D}{1+\text{ Number of documents containing term } t}\right) \end{aligned} \tag{1}$$

TF-IDF is a straightforward algorithm that is easy to understand and implement, and the resulting topic representations are relatively interpretable as they are based on the frequency of individual words. The tf-idf algorithm helps in ranking the relevance of words in documents relative to a query or set of documents. It is a widely used feature vector representation for text-related tasks in machine learning, such as document classification and clustering. In recent years, newer techniques

like BERT (Bidirectional Encoder Representations from Transformers) have gained popularity for topic modeling and other natural language processing tasks (Abedi and Sacchi, 2024). BERT is a deep learning-based language model that can capture more nuanced and contextual representations of text, which can lead to more accurate and sophisticated topic representations, particularly for ambiguous or polysemous words (Wan et al., 2024). BERT learns its word representations from a large corpus of text, allowing it to capture complex linguistic features and patterns, and it can be fine-tuned on specific domains or tasks, making it more flexible and adaptable than TF-IDF (Jáñez Martino et al., 2023).

The utilization of the TF-IDF algorithm for topic clustering presents several inherent limitations and challenges, particularly when applied to extensive datasets and diverse text forms. Furthermore, the loss of word semantics is another critical issue, as the TF-IDF algorithm, focusing solely on the term frequency and inverse document frequency, fails to incorporate the semantic relationships between words, thereby overlooking significant semantic information during feature extraction. This challenge underscores the need for integrating word vector models to capture semantic features within the corpus. Additionally, the unique difficulties presented by short texts, such as social media posts, highlight the inadequacy of traditional TF-IDF methods due to high-dimensional sparsity, necessitating the exploration of novel text representation techniques to mitigate semantic matrix sparsity and improve clustering accuracy.

4. Research Result

4.1. Semantic emotion classification

The obtained data were split according to the time series, and the relevant Sina Weibo data with the posting time from June 13, 0:00 to July 27, 24:00 were extracted, and a total of 60,228 Sina Weibo data were obtained. The obtained data were brought into the pretrained Bert fine-tuned model for analysis. The distribution of affective values is from 0 to 1. The closer to 0, the more pronounced the negative affect; the closer to 1, the more pronounced the positive affect; and when it tends to 0.5, it is considered close to neutral affect (Drus and Khalid, 2019).

In order to further analyze the number and proportion of tweets in different emotional color intervals, the number of tweets located in different intervals were counted separately, and the following results were obtained. The number of Weibo in the interval '0.0-0.2' is 46,425; in the interval '0.2-0.4' is 3,249; in the interval '0.4-0.6' is 2,019; in the interval '0.6-0.8' is 2,004; in the interval '0.8-1.0' is 6,526. After calculating the respective percentages, the results can be obtained as shown in the following Table 1.

Table 1: Distribution of each emotional color

Emotion score	Emotion tendency	Number of Weibo	Percentage of Weibo
0-0.2	Strongly negative	46425	77.088%
0.2-0.4	Negative	3249	5.395%
0.4-0.6	Neutral	2019	3.353%
0.6-0.8	Positive	2004	3.328%
0.8-1.0	Strongly Positive	6526	10.836%

4.2. Thematic analysis results

The crawler crawled 60,228 Sina Weibo texts, and took out the 200 keywords with the highest word frequency, and built a semantic network graph, forming 14,130 connection lines, average degree is 176.820, average path length is 1.11145 and then we can get the semantic network graph as shown in Figure 3.

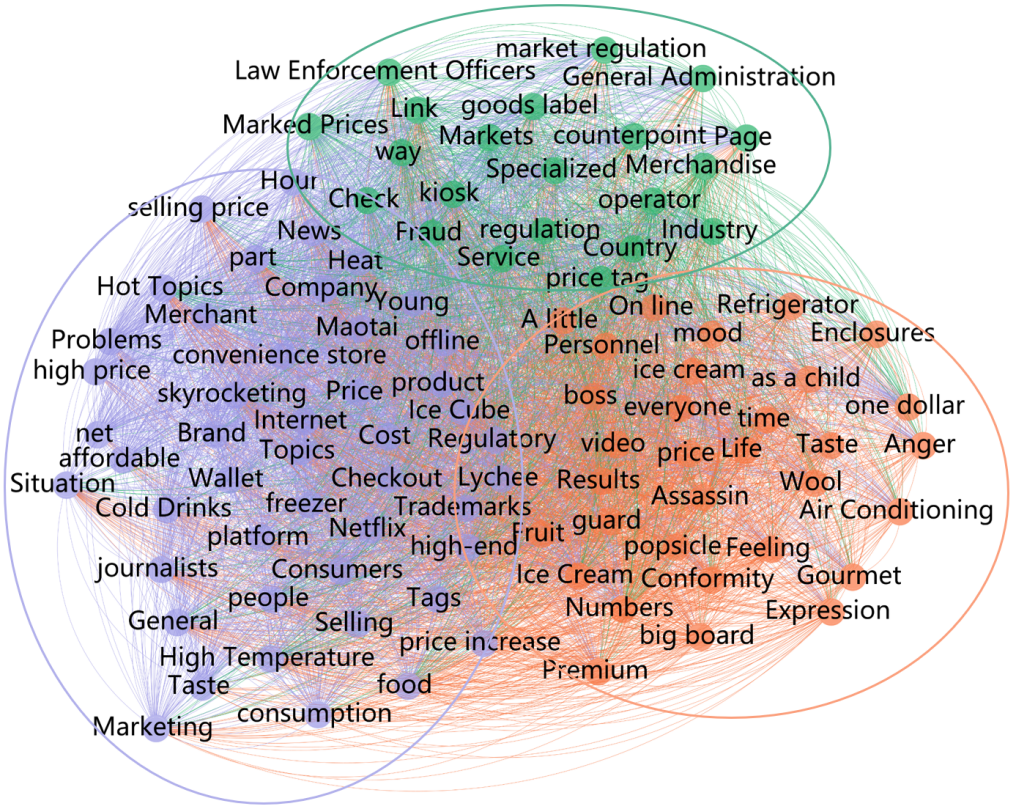

Figure 3: Semantic network diagram of related Sina Weibo

The first theme, characterized by its green color, delves into the investigative efforts by government bodies and market supervisory authorities into the burgeoning issue of "ice-cream assassins." This terminology refers to the phenomenon of excessively priced ice-cream products that have sparked considerable scrutiny. The theme explores the series of investigations and the subsequent regulatory measures that have been implemented to address this issue, highlighting the government's proactive stance in safeguarding consumer interests and ensuring fair market practices (Pu et al., 2023).

The second theme, marked by its orange color, focuses on the public's endeavor to understand and analyze the underlying reasons behind the problem of high-priced ice-cream. It represents a collective inquiry into the factors that lead to such exorbitant pricing, ranging from production costs, market dynamics, to potential manipulative practices within the industry. This theme captures the essence of consumer curiosity and the quest for transparency in how products are priced.

The third theme, represented in purple, sheds light on the personal experiences and reactions of individuals when faced with high-priced ice-cream during their purchasing moments. It encapsu-

lates the array of emotions, from surprise to dismay, that consumers feel when encountering such unexpected costs for a typically accessible treat. This theme resonates with the broader discourse on consumer rights and the psychological impact of pricing strategies on the buying behavior.

Together, these themes paint a comprehensive picture of the societal, regulatory, and individual dimensions surrounding the issue of high-priced ice-cream. They reflect a multifaceted approach to understanding and addressing a concern that touches upon economic, ethical, and emotional facets of daily life.

5. Discussion

The rapid spread of negative information on social media platforms has amplified the consequences of the "ice cream assassin" incident. The consumer concerns over food safety raised by this event could lead to a decline in trust towards related e-commerce platforms and online food ordering services. Consumers' hesitation to purchase from these platforms may adversely affect their sales performance. Furthermore, these platforms may implement restrictive measures, such as enhanced screening processes, which could in turn impact the operations of the businesses using these platforms. This chain reaction could potentially spread to the broader e-commerce market. Schools must strengthen their food safety management, while e-commerce platforms should improve their review and verification standards to regain consumer trust. Only through a collaborative effort can the negative consequences of such incidents be minimized.

To further understand the dynamics and implications of the "ice cream assassin" discussion on social media platforms, it would be valuable to conduct a sentiment analysis using a tool like Bert fine-tuned models . By applying sentiment analysis to the relevant social media posts and comments, researchers can gain deeper insights into the overall public sentiment surrounding this issue. This analysis could reveal the prevalence of negative, positive, or neutral sentiments, as well as identify any notable shifts in sentiment over time . Understanding the emotional undercurrents of this discussion could provide crucial context for assessing its potential impact on the school consumer goods economy and the e-commerce market.

There are limitations in our study due to the regulatory reasons for Weibo platforms and the various limitations and challenges faced by existing technologies. Using data exclusively from Weibo may limit the generalizability of your findings to other social media platforms or digital communities. Weibo users may not represent the broader internet population, especially outside of Chinese-speaking communities. The BERT model, while powerful, comes with inherent biases based on its training data. If the original training set does not well represent the specific context or slang used on Weibo, the sentiment analysis might be less accurate. The TF-IDF algorithm, while effective for identifying relevant terms, does not capture the semantics of the language or the context in which terms are used. This could result in less meaningful topic clusters if subtle nuances are critical. The choice of posts, timing, and volume can all impact the results. Limited or biased sampling may not accurately reflect broader trends or sentiments. Social media data can include bots, advertisements, and spam, which can distort analyses unless carefully cleaned and filtered.

6. Conclusion

In this study, we utilized tweets from Sina Weibo containing the keyword "ice-cream assassin" to conduct a Bert fine-tuned model-based sentiment analysis. The cleaned data was used to generate a

temporal change trend graph of sentiment intensity related to the keyword, and its correlation with major events at corresponding time points was analyzed. Our findings revealed that significant peaks in sentiment were associated with the release of relevant events by media and government entities. Furthermore, topics were categorized using the TF-IDF algorithm, and high-frequency words in the microblog content were analyzed through semantic network analysis. A relevant semantic network graph was constructed to identify three key topics: government measures in response to the incident, public discussion on its causes, and expressions of public sentiments about the incident.

Negative reviews and comments can directly deter potential buyers. Studies have shown that a significant percentage of consumers trust online reviews as much as personal recommendations, and negative information can disproportionately affect their purchase decisions due to negativity bias, where individuals give more weight to negative aspects of a product or service. Negative sentiment can lead to increased brand switching. Customers who encounter negative reviews are more likely to switch to competitor brands that do not have such associations. This behavior is intensified in highly competitive markets, such as electronics or fashion e-commerce. Continual negative feedback can erode customer loyalty, even among previously satisfied customers. The public nature of online reviews can create a bandwagon effect, leading even satisfied customers to reconsider their perceptions and loyalty to a brand.

It is important for e-commerce businesses to invest in sophisticated monitoring tools that can detect shifts in public sentiment early. Early detection can be crucial for mitigating damage and strategically managing public relations crises. I would like to suggest practical steps for businesses to engage with their customers transparently and authentically, especially during crises. This includes maintaining a consistent communication channel and being proactive in resolving customer issues. In addition, e-commerce platforms can build resilience against the negative impacts of adverse public opinion by fostering a loyal customer base and developing a strong, positive brand image that can withstand occasional negative bursts.

In addition, the rapid implementation of pertinent laws and regulations far surpasses the typical process for general policies. It can be inferred that the amplification of public sentiment on microblogs has driven the enactment of relevant laws and regulations, serving as a significant demonstration of public opinion monitoring governmental governance and widespread political participation.

References

Mohammad Majid Abedi and Emanuele Sacchi. A machine learning tool for collecting and analyzing subjective road safety data from twitter. *Expert Systems with Applications*, 240:122582, 2024. doi: 10.1016/j.eswa.2023.122582.

Akiko Aizawa. An information-theoretic perspective of tf—idf measures. *Inf. Process. Manage.*, 39(1):45–65, jan 2003. doi: 10.1016/S0306-4573(02)00021-3.

Alycia Burnett, Devin Knighton, and Christopher Wilson. The self-censoring majority: How political identity and ideology impacts willingness to self-censor and fear of isolation in the united states. *Social Media + Society*, 8(3):20563051221123031, 2022. doi: 10.1177/20563051221123031.

Sarah V. Damberg, Julia Hartmann, and H. Sebastian Heese. Does bad press help or hinder sustainable supply chain management? an empirical investigation of us-based corporations. *International Journal of Production Economics*, 249:108504, 2022. doi: 10.1016/j.ijpe.2022.108504.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, June 2019. doi: 10.18653/v1/N19-1423.

Zulfadzli Drus and Haliyana Khalid. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Comput. Sci.*, 161(C):707–714, jan 2019. doi: 10.1016/j.procs.2019.11.174.

Shuihua Han, Zhenyuan Liu, Ziyue Deng, Shivam Gupta, and Patrick Mikalef. Exploring the effect of digital csr communication on firm performance: A deep learning approach. *Decis. Support Syst.*, 176(C), mar 2024. doi: 10.1016/j.dss.2023.114047.

Francisco Jáñez Martino, Rocío Alaiz-Rodríguez, Víctor González-Castro, Eduardo Fidalgo, and Enrique Alegre. Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. *Appl. Soft Comput.*, 139(C), may 2023. doi: 10.1016/j.asoc.2023.110226.

Chen Lin, Dugang Liu, Hanghang Tong, and Yanghua Xiao. Spiral of silence and its application in recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2934–2947, 2022. doi: 10.1109/TKDE.2020.3013973.

Fengying Peng, Runmin Wang, Yiyun Hu, Guangyi Yang, and Ying Zhou. Feature fusion-based text information mining method for natural scenes. *Demonstratio Mathematica*, 56(1):20220255, 2023. doi: doi:10.1515/dema-2022-0255.

ZHANG Pu, ZHANG Hao, KONG Feng, and KONG Yunlong. A study on public opinion characteristics of rainstorm flooding disasters based on sina weibo data: take the three rainstorm flooding disasters in china in 2021 as an example. *Water Resources and Hydropower Engineering*, pages 47–59, 2023. doi: 10.13928/j.cnki.wrahe.2023.02.005.

Wang Qin. Project keyword lexicon and keyword semantic network based on word co-occurrence matrix. *Journal of Computer Applications*, 2015.

Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015. doi: 10.1016/j.knosys.2015.06.015.

Vivian Simon, Neta Rabin, and Hila Chalutz-Ben Gal. Utilizing data driven methods to identify gender bias in linkedin profiles. *Inf. Process. Manage.*, 60(5), sep 2023. doi: 10.1016/j.ipm.2023.103423.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012. doi: 10.1002/asi.21662.

Qifeng Wan, Xuanhua Xu, and Jing Han. A dimensionality reduction method for large-scale group decision-making using tf-idf feature similarity and information loss entropy. *Appl. Soft Comput.*, 150(C), apr 2024. doi: 10.1016/j.asoc.2023.111039.

Pu Zhang, Hao Zhang, and Feng Kong. Research on online public opinion in the investigation of the "7–20" extraordinary rainstorm and flooding disaster in zhengzhou, china. *International Journal of Disaster Risk Reduction*, 105:104422, 2024. doi: 10.1016/j.ijdrr.2024.104422.

J. Zhao, H. He, X. Zhao, and J. Lin. Modeling and simulation of microblog-based public health emergency-associated public opinion communication. *Inf Process Manag*, 59(2):102846, 2022. doi: 10.1016/j.ipm.2021.102846.