Cross-cultural Sentiment Analysis of Social Media Responses to a Sudden Crisis Event

Zheng Hui♔, **Zihang Xu**♔, **John Kender**♔
♔ Columbia University
zh2483, zx2362@columbia.edu, jrk@cs.columbia.edu

Abstract

Although the responses to events such as COVID-19 have been extensively studied, research on sudden crisis response in a multicultural context is still limited. In this paper, our contributions are 1)We examine cultural differences in social media posts related to such events in two different countries, specifically the United Kingdom lockdown of 2020-03-23 and the China Urumqi fire[1] of 2022-11-24. 2) We extract the emotional polarity of tweets and weibos gathered temporally adjacent to those two events, by fine-tuning transformer-based language models for each language. We evaluate each model's performance on 2 benchmarks, and show that, despite being trained on a relatively small amount of data, they exceed baseline accuracies. We find that in both events, the increase in negative responses is both dramatic and persistent, and does not return to baseline even after two weeks. Nevertheless, the Chinese dataset reflects, at the same time, positive responses to subsequent government action. Our study is one of the first to show how sudden crisis events can be used to explore affective reactions across cultures.

1 Introduction

The COVID-19 pandemic has now been ongoing for three years, impacting significant events such as the Wuhan outbreak, vaccine roll-outs, and state of emergency declarations. Throughout these events, individuals have been expressing their viewpoints on various social media platforms, which have become integral to their lives. While polarity detection is well-studied (e.g.,Agarwal et al. (2011);Garcia-Garcia et al. (2017);Yadollahi et al. (2017);Zhang et al. (2020);Liu et al. (2024); Giorgi et al. (2021); Hu et al. (2023)), in sudden event contexts (Desai et al. (2020);Kruspe et al. (2020);Wang et al. (2024)), research on **crisis response in a multicultural context** is still limited (Imran et al. (2020)). This study aims to contribute to the understanding of how to guide and focus people's emotional responses during emergencies through the analysis of sentiment expressed on social media during sudden crisis events. Our research purpose is to investigate how individuals from different cultural and linguistic backgrounds respond to the COVID-19 pandemic in social medias, with a specific focus on crisis events in the United Kingdom and China. Cultural values and norms significantly influence people's behaviors (Kirk et al., 2024) during a crisis. Understanding these behaviors can help in tailoring public health messages that are culturally sensitive and more likely to be effective (Resnicow et al., 1999; Griffith et al., 2024). And the degree to which different cultures comply with and trust public health measures (such as social distancing, mask-wearing, and vaccinations) can provide insights into how these measures should be communicated and enforced. Addationaly, different cultures have unique ways of dealing with crisis and adversity. Studying these can offer valuable lessons in building resilience and mental health support systems (Hershcovich et al., 2022; Liu et al., 2025). Importantly, this work reflects a growing shift in NLP toward socially beneficial applications (Ai et al., 2024b; Hui et al., 2025), using language models not just for technical benchmarks but to understand real-world emotional responses in times of crisis. Each dataset covers a one-month period, spanning the two weeks before and after a sudden crisis event, and each collects manual crowd-sourced annotations of the polarity expressed in the posts. We have developed two language-specific transformer-based models to analyze the sentiment of these posts, classifying their polarity as negative, neutral, or positive. Compared to prior studies(Lee et al., 2022; White et al., 2024; Hui et al., 2024a) that consider only sentence-level or aspect-level texts, our work is more challenging, as it is Cross-cultural studies involve understanding

[1] https://wikipedia.org/wiki/2022_ÃœrÃmqi_fire

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 294–305
July 31, 2025 ©2025 Association for Computational Linguistics

and navigating diverse cultural norms, values, and communication styles. Analyzing sentiment in this context requires sensitivity to cultural nuances that influence how emotions are expressed and interpreted. Moreover, focusing on the before-and-after aspects adds a temporal dimension, demanding an examination of evolving emotional dynamics and how cultural factors shape these changes over time. This complexity makes the study more challenging but also more comprehensive in capturing the full spectrum of emotional responses to sudden crisis events with nation-wide impact.

2 Related Work

2.1 Impact of Covid-19 on Mental Health

During the COVID-19 pandemic, social distancing and city lockdowns significantly impacted people's emotional health. The relationship between social media use and emotional health has been studied by researchers such as Karim et al. (2020). In particular, Marshall et al. (2022) used natural language processing to gain mental health insights from UK tweets during the COVID-19 pandemic, and Zhang et al. (2022) presented a narrative review of the application of NLP in detecting mental illnesses.

2.2 Cross-Cultural Differences of Sentiment

A major area of interest in the context of COVID-19 is how individuals react to critical events on social media. Dean et al. (2021) conducted cross-cultural comparisons of psychosocial distress during the early stages of COVID-19 in four countries with diverse public health strategies. The study identified varying magnitudes of psychological distress across regions, with Hong Kong experiencing the most significant decline in mental health, likely attributed to the imposition of stringent social distancing regulations and continuing political turmoil. There is limited research in this field.

2.3 Sentiment Polarity

Sentiment polarity analysis on social media has also been widely discussed, for example by Zhang et al. (2018), Yadollahi et al. (2017) , Hu and Collier (2024) and Ai et al. (2024a). Previous studies have mostly focused on utilizing lexicon-based sentiment polarity detection such as Musto et al. (2014), and use machine learning algorithms such as Samuel et al. (2020) to analyze social media posts for polarity assessment.

3 Methodology

3.1 Datasets

For our Weibo dataset, we builded a Weibo web crawler to collect data. In contrast, the UK Twitter data was obtained from COVID-19 Tweets Dataset (Banda et al., 2021) by location, time, and keyword filters. The kewyword using to extract data from Weibo and COVID-19 Tweets Dataset (Banda et al., 2021) are showed in Table 1.

CovidSEE and CovidSEC were then created by sampling 60 random minutes from each day of a four-week interval that bracketed, two weeks before and two weeks after, their sudden crisis event, which were the UK lockdown and the Urumqi fire, respectively. Then, through crowd-sourced manual annotation, we classified posts into three polarity categories: negative, neutral, and positive, encoded as -1, 0, +1, respectively. The distribution of the datasets according to keyword, and the numbers of collected posts are shown in Table 1. Examples from each dataset are shown in Appendix A. Annotation details are shown in Appendix B.

Dataset	Posts	Keyword
CovidSEE	49,810	covid, coronavirus
CovidSEC	47,681	新冠 (covid)

Table 1: Number of posts each dataset contains, and keywords used to filter them, based on the existing database or web crawler, respectively.Datasets are collected and formed in the posters' native language.

3.2 Transfer Learning for Sentiment Analysis

For sentiment analysis, COVID-Twitter-BERT (Müller et al., 2020) was used for the CovidSEE and Chinese-BERT (Cui et al., 2021) was used for the CovidSEC. The adaptations of these BERT-like models (Devlin et al., 2018) were relatively straightforward, and only involved a modification to the models' heads.

To begin, we tokenized input texts with nltk(Loper and Bird, 2002), Jieba and Fast WordPiece (Song et al., 2020), followed by prepending each tokenized input with a [CLS] classification token and feeding it through the BERT model. Finally, after the last layer, we linearly projected each [CLS] token into one of three categories: negative, neutral, or positive. We based our fine-tuning approach on the work of Sun et al. (2019), who used BERT for classification. The resulting fine-tuned BERT models were dubbed SaTwBERT ("Sentiment analysis Twitter BERT") and SaChBERT ("Sentiment analysis Chinese BERT").

In more detail: To fine-tune Covid-Twitter-BERT and Chinese-BERT on our collected dataset, we used the AdamW optimizer (Loshchilov and Hutter, 2017) with learning rate=2e-5, β_1=0.9, β_2=0.999, and weight decay=0.01. To aid in the optimization process, we used a learning rate warm-up for 10,000 steps and a batch size of 32. We used A100 GPU and conducted training for 5 epochs.

4 Experiments

We evaluated the performance of SaTwBERT for English and SaChBERT for Chinese by comparing their accuracy rates and macroF1 against baseline models. The dataset was divided into 80% training and 20% test data for five-fold cross-validation, for both languages. This ensured a comprehensive assessment of the models' effectiveness.

English models	**ACC**	**macroF1**	**Std**
FastText	80.4	67.0	0.052
ABCDM	83.9	80.1	0.047
T5-based	85.2	77.9	0.038
GPT-3*	88.5	79.8	0.043
GPT-4omini	89.0	82.5	0.033
SaTwBERT (ours)	**89.1**	**83.6**	0.050

Table 2: Average performance on CovidSEE using 5 random seeds. *GPT was trained multilingually.

Chinese models	**ACC**	**macroF1**	**Std**
BERT-base	81.0	70.8	0.045
SLCABG	86.2	79.7	0.043
T5-based-chinese	85.8	79.8	0.055
GPT-3*	87.9	**80.3**	0.032
GPT-4omini	87.5	80.1	0.046
SaChBERT (ours)	**88.5**	76.7	0.042

Table 3: Average performance on CovidSEC using 5 random seeds. *GPT was trained multilingually.

4.1 Performance

As baselines to compare with our model, for English we used fastText (Bojanowski et al., 2017); ABCDM, with modifications (Basiri et al., 2021); and T5 (Raffel et al., 2020). For Chinese we used BERT-base-uncased (Devlin et al., 2018) after including a linear layer to achieve the three polarities; SLCABG (Yang et al., 2020); and T5-based-chinese (Raffel et al., 2020). Additionally, we also benchmarked the multilingual large language model GPT-3 and GPT-4omini[2] (Brown et al., 2020), in order to explore the capabilities of multilingual models for polarity classification in the context of sudden crisis events.

Table 2 and Table 3 present the comparison between the ten different models, showing the average over the five random folds. Both of our models outperformed their baselines with statistically significant results in accuracy, and our English model did the same in macroF1. In addition, our models significantly outperformed the baselines(FastText) by 8.7% in overall polarity classification, even when polarities were not explicitly stated in the posts. For instance, in a sentence such as "I want to leave UK and never come back in my life", our model accurately inferred that the statement conveys negative polarity.

4.2 Error Analysis

We observed three common types of errors. The first occurs when a neutral sentence contains a non-emotive negation. The second involves complicated sentence structures in a single post that express more than one perspective. The third is triggered by sarcasm and irony. More detailed examples of these errors are shown in Appendix F.

4.3 Sentiment Analysis Result

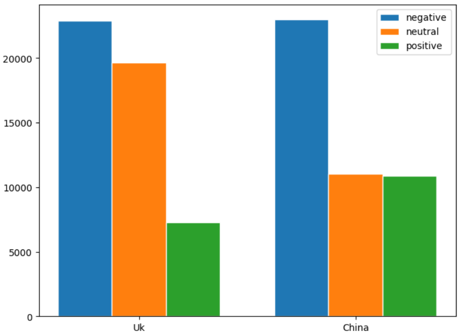


Figure 1: Statistics of polarity.

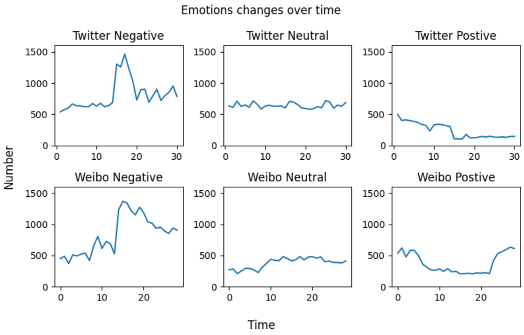


Figure 2: Individual polarity changes over time.

[2] See Appendix E for the GPT-3 and GPT4omini prompt.

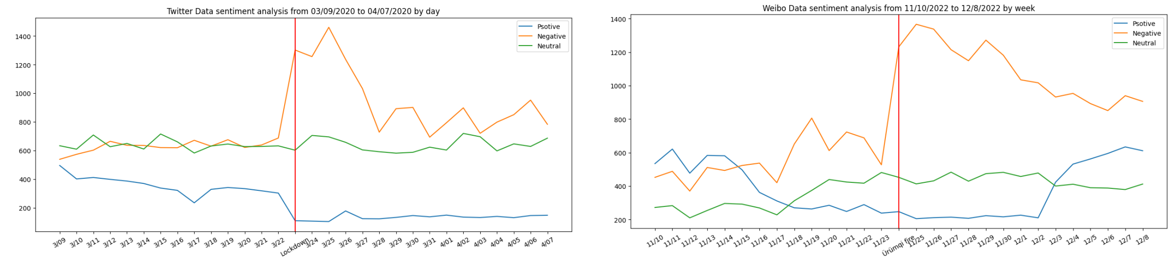


Figure 3: Effect of sudden crisis events over time. Left: UK, right: China.

Figure 1 and Figure 2 present the distributions of polarities, and how they change over time for both countries. From 2020-03-09 to 2020-04-07, a total of 22,884 negative tweets, 19,675 neutral tweets, and 7,251 positive tweets were recorded on UK Twitter. Similarly, Weibo data from 2022-11-10 to 2022-12-08 indicate 24,534 negative tweets, 11,663 neutral tweets, and 11,484 positive tweets.

Figure 3 shows polarity changes during the month-long span bracketing the UK lockdown and the Urumqi fire, respectively. The total amounts of negative posts between the two countries is roughly comparable. The Weibo data shows an additional late recovery in positive posts, which nevertheless did not appreciably affect the continued dominance of negative posts. These findings are in contrast to prior research which found that only positive news events tended to be long-lasting (Wu et al., 2011).

5 Discussion

5.1 Cultural Context and Emotional Response

The impact of COVID-19 is evident from the analysis from both countries, where individuals were more inclined towards negative polarity. The proportion of negatives was 45.8% in the UK, and 49.1% in China. However, the neutrals in the UK were significantly higher than China's, where the figures were 39.7% and 27.7%, respectively. This could be attributed to the severity of the events in the two countries. The lockdown imposed in the UK only affected everyday convenience, social communication, and business profitability. UK individuals' negative profile experienced a less severe increase, and more rapidly reverted to something closer to its pre-crisis levels. On the other hand, the tragic loss of lives in the Urumqi fire in China was much more severe. Chinese individuals tended to respond more radically and more persistently, even as the Chinese government took some steps to salvage public sentiment. A further analysis of the data from China reveals a somewhat more complicated picture. Although negative sentiments showed only a slow decline in the days following the fire, the expressions of positive sentiment in China witnessed a notable improvement to well above baseline levels on the days after 2022-12-02. This was only about a week after the fire. Upon research, we found that the Chinese government announced on that day the cancellation of the health code policy. Nonetheless, there was also a continuation of negative sentiment, most likely attributable to concerns about premature re-opening while still amid the COVID-19 pandemic. This suspicion finds support through subsequent reports that people's infection rate then started to rise drastically, leading to an increased number of fatalities.

5.2 Temporal Dynamics and Persistence of Negative Sentiment

Another key insight from our study is the temporal persistence of negative sentiment following each crisis event. We observed that surges in negative emotions did not subside immediately after the initial shock; instead, elevated levels of anger, fear, and sadness persisted for an extended period in both contexts. In the UK, public anxiety and frustration remained high for weeks after the lockdown announcement, with sentiment trends showing only gradual normalization as people adapted to restrictions. This prolonged negativity suggests a sustained psychological impact – a collective stress that could influence compliance and mental health long after the policy was introduced. In China, the wave of anger triggered by the Urumqi fire similarly showed a lasting presence on social media in the days following the incident. Despite swift official efforts to control the narrative, internet users continued to voice skepticism and anger. The persistence of negative sentiment in China had palpable societal implications: it helped fuel rare public protests and demands for policy change, indicating that when grievances remain unaddressed, online negativity can translate into real-world action. From a psychological perspective, the enduring nature of these negative emotions in both countries points to potential long-term effects on public trust and wellbeing (Yuan et al., 2023; Liu et al., 2024). If left unmanaged, sustained collective anger or fear may erode confidence in authorities and hinder recovery from the crisis. Thus, the temporal dynamics we uncovered, particularly the lingering tail of negative sentiment, carry important implications. Crisis managers and public health officials must recognize that the public's emotional recovery often lags behind the immediate crisis response. Interventions such as ongoing mental health support, transparent communication to address persisting fears, and visible responsiveness to public concerns are essential to help dissipate negative sentiment over time.

6 Conclusion

We presented two transformer-based models for sentiment analysis that were tailored to sudden crisis events. Our models demonstrated stable and superior performance compared to baseline models. They enabled cross-cultural comparisons of people's responses, showing a notable persistence of negative responses to sudden crisis events.

We aim to further enhance our models by examining further sudden crisis events, and by expanding our multi-cultured analysis to events that are synchronous across countries. We are also exploring ways of reliably extending our tripartite division of sentiment to one of a five-way scale, in order to better accommodate extreme sentiments (e.g., "very positive"). Social media plays a crucial role in public health emergencies, enabling the public to access important information and express their emotions. However, there are significant differences in social media usage patterns and public sentiment responses across different countries and regions. Our research results can aid agencies in developing effective response strategies for public health emergencies and promoting better public mental health.

7 Future Research Directions

Future Research Directions: Building on this work, we see several avenues to broaden and deepen the analysis. First, a multi-event, multi-lingual approach should be pursued. Analyzing additional crisis events across different countries and languages would test the generalizability of our findings and models. Comparing sentiment patterns from diverse crises – from natural disasters to public health emergencies – could reveal whether certain emotional trajectories are universal or culture-specific. Second, future studies should employ more nuanced sentiment scales and emotion categories. Rather than relying on coarse sentiment polarity or a few basic emotions, researchers could incorporate fine-grained emotions (e.g. distinguishing anger from disappointment, or fear from anxiety) and even measure sentiment intensity. This would capture subtler shifts in public mood and provide a richer picture of the crisis impact on society's psyche. Third, exploring real-time sentiment analysis and response modeling is a promising direction. Developing systems that continuously track social media sentiment during an unfolding crisis would enable dynamic feedback – for instance, alerting officials to spikes in negative emotion so they can adjust messaging in the moment. Real-time models, possibly integrated with geo-spatial or network analyses, could help identify not only when and what emotions surge, but also where misinformation or distress is propagating. Finally, ongoing refinement of transformer-based sentiment models is needed to address the limitations highlighted in our discussion. This includes improving handling of sarcasm, context, and multilingual inputs, as well as ensuring ethical use of these technologies. By pursuing these future directions, researchers and practitioners can enhance the power of cross-cultural sentiment analysis as a tool for understanding and navigating the complex emotional landscape of crisis events. Ultimately, our study shows that tracking and interpreting public sentiment across cultures is not only feasible with advanced NLP models, but also invaluable for guiding compassionate and effective crisis management on a global scale.

Limitations

Our study has several limitations. First, social media data may not be fully representative, as usage patterns and sentiment expression vary across cultures, regions, and demographics. Certain groups may be underrepresented or self-censor due to platform moderation or political concerns, particularly when expressing harmful or sensitive views

(Hui et al., 2024b). Second, sentiment annotation is inherently subjective. Annotators may interpret emotions differently based on cultural or personal perspectives, which can lead to inconsistencies—especially in posts involving sarcasm or implicit harmful speech. Third, limited resources constrained the size and depth of our annotation process, potentially affecting label quality. Addressing these limitations through broader data collection, refined annotation protocols, and explicit handling of harmful content would strengthen future cross-cultural sentiment analysis.

Ethics Statement

Copyright Compliance and Data Anonymization: For the Twitter dataset: We utilized an open-source compendium of tweets and annotations, ensuring that all data were fully anonymized to safeguard user privacy. For the Weibo dataset: We collected data in strict accordance with Weibo's copyright terms of use, using our proprietary scraper to ensure compliance. Furthermore, all collected Tweets and Weibo content underwent thorough anonymization before being made available for annotation.

Annotator Recruitment: Our annotators were recruited through the networks of two student coauthors via platforms such as WeChat and student WhatsApp campus group chats .Annotator volunteers were required to commit to a minimum of 300 tri-valued annotations, covering negative, neutral, and positive sentiments. They were also provided with a clear set of instructions and agreements to follow. All annotators underwent testing on a smaller dataset to assess their qualifications. It's important to note that annotators participated voluntarily and without any form of money compensation.

Annotator Selection: Annotators were selected based on their language expertise and their ability to commit to a minimum annotation workload of 300 items. Annotators were only rejected if they did not meet the commitment requirements or if they did not pass the initial qualification test. The selection process prioritized language proficiency rather than considering the annotators' country of origin, ensuring a diverse perspective. We believe that these measures sufficiently address the ethical concerns raised, ensuring that our research adheres to ethical principles and practices. We are committed to transparency and accountability in our work and welcome any further inquiries or clarifications regarding the ethical aspects of our research.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 30–38.

Lin Ai, Sameer Gupta, Shreya Oak, Zheng Hui, Zizhou Liu, and Julia Hirschberg. 2024a. Tweetintent@ crisis: A dataset revealing narratives of both sides in the russia-ukraine crisis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1872–1887.

Lin Ai, Tharindu Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, et al. 2024b. Defending against social engineering attacks in the age of llms. In *In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12880–12902, Miami, Florida, USA. Association for Computational Linguistics*.

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324.

Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U Rajendra Acharya. 2021. Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis. *Future Generation Computer Systems*, 115:279–294.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Derek J Dean, Ivy F Tso, Anne Giersch, Hyeon-Seung Lee, Tatiana Baxter, Taylor Griffith, Lijun Song, and Sohee Park. 2021. Cross-cultural comparisons of psychosocial distress in the usa, south korea, france, and hong kong during the initial phase of covid-19. *Psychiatry Research*, 295:113593.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jose Maria Garcia-Garcia, Victor MR Penichet, and Maria D Lozano. 2017. Emotion detection: a technology review. In *Proceedings of the XVIII international conference on human computer interaction*, pages 1–8.

Salvatore Giorgi, Vanni Zavarella, Hristo Tanev, Nicolas Stefanovitch, Sy Hwang, Hansi Hettiarachchi, Tharindu Ranasinghe, Vivek Kalyan, Paul Tan, Shaun Tan, et al. 2021. Discovering black lives matter events in the united states: Shared task 3, case 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 218–227.

Louis A Gottschalk and Goldine C Gleser. 1979. *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press.

Derek M Griffith, Caroline R Efird, Monica L Baskin, Monica Webb Hooper, Rachel E Davis, and Ken Resnicow. 2024. Cultural sensitivity and cultural tailoring: lessons learned and refinements after two decades of incorporating culture in health communication research. *Annual Review of Public Health*, 45(1):195–212.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*.

Tiancheng Hu, Manoel Horta Ribeiro, Robert West, and Andreas Spitz. 2023. Quotatives indicate decline in objectivity in us political news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 363–374.

Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, Lin Ai, Yinheng Li, Julia Hirschberg, and Congrui Huang. 2024a. Can open-source llms enhance data augmentation for toxic detection?: An experimental study. *arXiv preprint arXiv:2411.15175*.

Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, and Congrui Huang. 2024b. Toxicraft: A novel framework for synthetic generation of harmful information. In *In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Zheng Hui, Xiaokai Wei, Yexi Jiang, Kevin Gao, Chen Wang, Frank Ong, Se eun Yoon, Rachit Pareek, and Michelle Gong. 2025. Matcha: Can multi-agent collaboration build a trustworthy conversational recommender?

Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, and Rakhi Batra. 2020. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. *Ieee Access*, 8:181074–181090.

Fazida Karim, Azeezat A Oyewande, Lamis F Abdalla, Reem Chaudhry Ehsanullah, and Safeera Khan. 2020. Social media use and its connection to mental health: a systematic review. *Cureus*, 12(6).

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344. Curran Associates, Inc.

Anna Kruspe, Matthias Häberle, Iona Kuhn, and Xiao Xiang Zhu. 2020. Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Chia-Hsuan Lee, Aditya Siddhant, Viresh Ratnakar, and Melvin Johnson. 2022. DOCmT5: Document-level pretraining of multilingual language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 425–437, Seattle, United States. Association for Computational Linguistics.

Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, May Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. 2024. Propainsight: Toward deeper understanding of propaganda in terms of techniques, appeals, and intent. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Zizhou Liu, Ziwei Gong, Lin Ai, Zheng Hui, Run Chen, Colin Wayne Leach, Michelle R Greene, and Julia Hirschberg. 2025. The mind in the machine: A survey of incorporating psychological theories in llms. *arXiv preprint arXiv:2505.00003*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Christopher Marshall, Kate Lanyi, Rhiannon Green, Georgina C Wilkins, Fiona Pearson, Dawn Craig, et al. 2022. Using natural language processing to explore mental health insights from uk tweets during the covid-19 pandemic: infodemiology study. *Jmir Infodemiology*, 2(1):e32449.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Cataldo Musto, Giovanni Semeraro, and Marco Polignano. 2014. A comparison of lexicon-based approaches for sentiment analysis of microblog posts. In *DART@ AI* IA*, pages 59–68. Citeseer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ken Resnicow, Tom Baranowski, Jasjit S Ahluwalia, and Ronald L Braithwaite. 1999. Cultural sensitivity in public health: defined and demystified. *Ethnicity & disease*, 9(1):10–21.

Jim Samuel, GG Md Nawaz Ali, Md Mokhlesur Rahman, Ek Esawi, and Yana Samuel. 2020. Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, 11(6):314.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2020. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2024. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. In *Advances in Neural Information Processing Systems*, volume 37, pages 58118–58153. Curran Associates, Inc.

Isadora White, Sashrika Pandey, and Michelle Pan. 2024. Communicate to play: Pragmatic reasoning for efficient cross-cultural communication. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12201–12216, Miami, Florida, USA. Association for Computational Linguistics.

Shaomei Wu, Chenhao Tan, Jon Kleinberg, and Michael Macy. 2011. Does bad news go away faster? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 646–649.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.

Li Yang, Ying Li, Jin Wang, and R Simon Sherratt. 2020. Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE access*, 8:23522–23530.

Yue Yuan, Shuting Yang, Xinying Jiang, Xiaomin Sun, Yiqin Lin, Zhenzhen Liu, Yiming Zhu, and Qi Zhao. 2023. Trust in government buffers the negative effect of rumor exposure on people's emotions. *Current Psychology*, 42(27):23917–23930.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multimodal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3584–3593.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

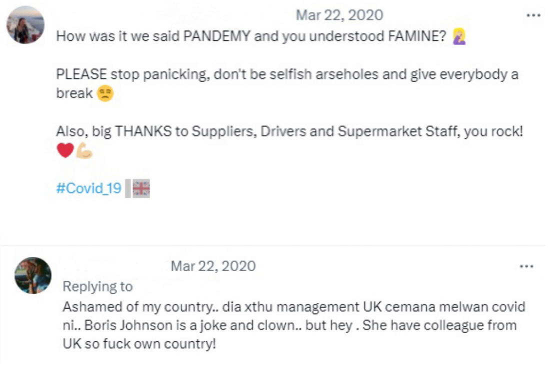
Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.

A Dataset Examples

We show some data examples from both CovidSEE and CovidSEC, in Figure 4 and Figure 5, respectively.

The details of the tweets were obtained by reverse-searching our collected database on Twitter and Weibo. Regrettably, due to the dynamic nature of social media platforms, some of the data initially recorded in CovidSEE and CovidSEC could no longer be located on Twitter and Weibo. This can be attributed to various reasons, such as post deletions or account suspensions by the original users. As a result, we acknowledge the limitations in the availability of the complete dataset.

To maintain strict adherence to privacy and ethical standards, we have taken precautions to conceal any identifiable user information from both Twitter and Weibo. This ensures that the individuals behind the collected data remain anonymous and their privacy is protected throughout the analysis process.



Mar 22, 2020 ···

How was it we said PANDEMY and you understood FAMINE? 🤦‍♀️

PLEASE stop panicking, don't be selfish arseholes and give everybody a break 😒

Also, big THANKS to Suppliers, Drivers and Supermarket Staff, you rock! ❤️💪

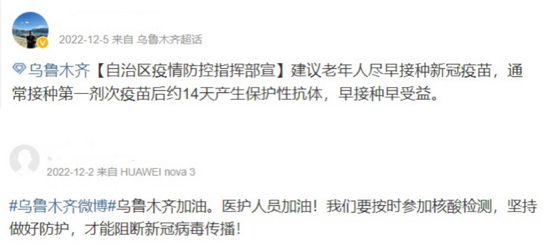
#Covid_19 🇬🇧

Mar 22, 2020 ···

Replying to

Ashamed of my country.. dia xthu management UK cemana melwan covid ni.. Boris Johnson is a joke and clown.. but hey . She have colleague from UK so fuck own country!

Figure 4: Twitter dataset example



2022-12-5 来自 乌鲁木齐超话

乌鲁木齐【自治区疫情防控指挥部宣】建议老年人尽早接种新冠疫苗，通常接种第一剂次疫苗后约14天产生保护性抗体，早接种早受益。

2022-12-2 来自 HUAWEI nova 3

#乌鲁木齐微博#乌鲁木齐加油。医护人员加油！我们要按时参加核酸检测，坚持做好防护，才能阻断新冠病毒传播！

Figure 5: Weibo dataset example

B Annotation Detail and Instruction

129 Annotators were recruited from friendship networks and social media contacts. Annotators total annotated 90k posts(English and Chinese). Due to limited resources, not all annotators possessed bilingual proficiency in both English and Chinese. However, every annotator had at least one of these languages as their native language.

Annotators were provided with clear guidelines (Gottschalk and Gleser, 1979) on how to annotate the polarity of the given text, with options for negative, neutral, or positive sentiment. Annotators were instructed to only choose one option, and were not allowed to make multiple selections; however, if the text exhibited two polarities simultaneously, annotators had the option to select "none of the above" as an alternative. Annotators used the numerical scale of -1, 0, +1 to denote negative, neutral, and positive sentiment, respectively.

A random subset(15%) of annotations were cross-checked against other annotators, the agreements between annotators is 95% and a different subset was explicitly checked by the authors. Table 4 presents some illustrative examples.

C Anotation Agrement

Scope of Work: The annotator agrees to annotate sentiment polarity labels for social media posts collected from various platforms such as Twitter, Weibo. The posts will be related to specific sudden crisis events, and the annotator will be responsible for accurately labeling the sentiment as positive, negative, or neutral based on the content of the posts.

Guidelines for Annotation: The annotator agrees to follow the provided annotation guidelines which including examples, which include specific criteria for determining the sentiment polarity of each social media post. These guidelines will outline the key indicators for identifying positive, negative, and neutral sentiment in the context of crisis events, taking into consideration the cultural and linguistic nuances of the target audience.

Quality and Consistency: The annotator agrees to maintain a high level of quality and consistency throughout the annotation process. This includes ensuring that each labeled sentiment reflects the actual sentiment expressed in the social media post accurately. Any uncertainties or ambiguities encountered during the annotation process will be immediately brought to the attention of the project supervisor for clarification.

Confidentiality and Data Security: The annotator acknowledges the sensitive nature of the data being handled and agrees to maintain strict confidentiality throughout the annotation process. The annotator will not disclose any information or data related to the project to any unauthorized individuals or third parties.

Timelines and Deliverables: The annotator agrees to adhere to the agreed-upon timelines and deliverables for the completion of the annotation tasks. The annotator will provide timely updates on the progress of the annotation work and notify the project owner of any potential delays or issues that may arise during the process.

Both parties acknowledge that they have read and understood the terms of this agreement and agree to abide by its provisions.

D Sample of Guidelines for Annotation

Contextual Understanding: The annotator must have a thorough understanding of the context in which the social media posts were made, including the specific crisis events and the cultural and linguistic nuances associated with the target audience. This contextual understanding will help in accurately assessing the sentiment expressed in the posts.

Language Considerations: The annotator should be proficient in the languages used in the social media posts to accurately interpret the sentiment. They should be aware of any colloquial expressions, slang, or language variations that might affect the overall sentiment conveyed in the posts.

Tone and Emotive Language: The annotator should pay close attention to the tone and emotive language used in the social media posts. They should consider factors such as the use of emoticons, exclamation marks, and other linguistic markers that indicate the emotional intensity of the content.

Objective Assessment: The annotator must approach the task with objectivity and impartiality, ensuring that personal biases or opinions do not influence the annotation process. The sentiment labels should reflect the general sentiment expressed by the majority of the posts rather than the annotator's individual viewpoint.

Ambiguity Resolution: In cases where the sentiment expressed in a social media post is ambiguous or unclear, the annotator should consult the provided guidelines or seek clarification from the project supervisor. It is essential to resolve any ambiguities to ensure consistent and accurate annotation across all posts.

Labeling Consistency: The annotator should strive for consistency in labeling sentiment across different social media posts. Similar content with comparable emotional expressions should receive the same sentiment label, maintaining uniformity throughout the annotation process.

Annotation Tools and Procedures: The annotator should utilize the designated annotation tools and follow the prescribed procedures for recording and documenting the sentiment labels. Any specific requirements regarding data entry, formatting, or tagging should be strictly adhered to for streamlined data management and analysis.

E GPT Prompt

The prompt used for GPT-3 is the following: "Given this text, it is important to consider the overall context, specific keywords, and the presence of any sentiment indicators to determine the sentiment conveyed. Pay attention to the tone, language, and any explicit expressions of emotions or opinions within the text. Analyze the text carefully, considering both the explicit and implicit sentiments expressed, to make an accurate judgment of the sentiment conveyed, choosing from negative, neutral, or positive. Text: {sentence}."

F Error Analysis Case Study

In Table 5, we present five illustrative examples that highlight common errors made by our sentiment analysis model. The first two examples exemplify instances where neutral sentences containing non-emotive negations result in incorrect predictions. The third example shows the challenges posed by complex sentence structures within a single post. The fourth and fifth examples demonstrate that the models may have difficulty in accurately classifying sentiment in the presence of sarcasm.

Posts	**Language**	**Platform**	**Annot. 1**	**Annot. 2**
1. [Expletive] working all your days only to find yourself setting an alarm and getting up early to go to Asda coz the UK population are greedy, stockpiling [expletive]s! Maybe some afternoon shopping hours for the elderly and vulnerable too no? #Covid_19 so many things about this making me sad	En	Twitter	-1	-1
2. To the Doctors, Police / Army, Government Officers on duty, Pilots, Aircraft staff, Train / Bus Drivers, Food / Courier Deliver Person and most importantly Garbage Pickers / Sweepers - Thank You So Much	En	Twitter	1	1
3. I really hope the #Covid_19 crisis in the UK really makes people ’wake up!’ to some of the realities in our society and the indoctrinated BS that’s amongst other things led to #panicbuying >:(	En	Twitter	-1	-1
4. 真的不懂健康码是怎么赋黄码的，在学校待着哪也没去，核酸检测每隔一天学校组织做一次，刚刚变黄码了，马上拉去隔离。全世界干脆和疫情一起毁灭算了#新冠	Zh	Weibo	-1	-1
5. 因为新冠居家隔离三天了，今天天气很好，明天该是怎样的状况呢?	Zh	Weibo	0	0
6. 【现在播报】北京今日新增44例本土感染者11月11日，在北京市新型冠状病毒肺炎疫情防控工作第410场新闻发布会上，市疾控中心副主任刘晓峰介绍，8日0时至15时，本市新增本土新冠肺炎病毒感染者49例，其中，隔离观察人员40例、社会面筛查人员4例。	Zh	Weibo	0	0

Table 4: Anotation Example

Posts	Language	Human Classification	Model Classification
1. 今天的天气好差，隔离餐也不好吃只有馒头配青菜#隔离日记#新冠	Zh	Netural	Negative✗
2. Between COVID-19 and the upcoming weather this week I for one **don't** want to go to work this week	En	Netural	Negative✗
3. 网友来信：你好，在家封控了将近一个月，今天是复工第一天，我不在新疆，但我男朋友在新疆喀什，他是去工作，所以住在酒店里面。我了解的情况是：10.8日通知，说是静默七天，之后就一直静默到现在。喀什没有报告一例新增、无症状，但是有人莫名被拉去隔离。我男朋友在酒店，盒饭30一份、没有肉。中午晚上都一样。现在泡面吃不上了，今天没有米饭, 明天的盒饭还不知道有没有。很想去看望我男朋友，但是又怕出去了回不来(隔离啥的)怕了怕了。	Zh	Negative	Netural✗
4. I love Covid-19, Covid-19 is my friend	En	Negative	Positive✗
5. 学校隔离每天都能吃到一顿肉呢，真的太幸福, 简直天堂	Zh	Negative	Positive✗

Table 5: Error Example, ✗ indicates incorrect prediction