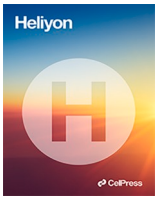
Heliyon 10 (2024) e35468

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

From hate to harmony: Leveraging large language models for safer speech in times of COVID-19 crisis

August F.Y. Chao [a,*], Chen-Shu Wang [b], Bo-Yi Li [c], Hong-Yan Chen [b]

[a] Department of Computer Science and Information Engineering, National Penghu University of Science and Technology, Taiwan
[b] Department of Information and Finance Management, National Taipei University of Technology, Taiwan
[c] Department of Management Information Systems, National Chengchi University, Taiwan

A B S T R A C T

This study investigates the rampant spread of offensive and derogatory language during the COVID-19 pandemic and aims to mitigate it through machine learning. Employing advanced Large Language Models (LLMs), the research develops a sophisticated framework adept at detecting and transforming abusive and hateful speech. The project begins by meticulously compiling a dataset, focusing specifically on Chinese language abuse and hate speech. It incorporates an extensive list of 30 pandemic-related terms, significantly enriching the resources available for this type of research. A two-tier detection model is then introduced, achieving a remarkable accuracy of 94.42 % in its first phase and an impressive 81.48 % in the second. Furthermore, the study enhances paraphrasing efficiency by integrating generative AI techniques, primarily Large Language Models, with a Latent Dirichlet Allocation (LDA) topic model. This combination allows for a thorough analysis of language before and after modification. The results highlight the transformative power of these methods. They show that the rephrased statements not only reduce the initial hostility but also preserve the essential themes and meanings. This breakthrough offers users effective rephrasing suggestions to prevent the spread of hate speech, contributing to more positive and constructive public discourse.

1. Introduction

The emergence and widespread adoption of social media platforms have facilitated an expansive forum wherein individuals can freely express their opinions and perspectives. However, this unrestricted environment has concurrently led to the proliferation of multifarious issues. Notably, one of the most prominent predicaments pertains to the dissemination and amplification of abusive and hateful language across these platforms, an occurrence that significantly jeopardizes societal well-being and harmony. For instance, at the societal level, the spread of racial hatred on social media following events such as the George Perry Floyd case in May 2020 led to significant turmoil in American society [1]. This led advertisers to endorse the #StopHateforProfit initiative, urging social media platforms to proactively address hate content. They advocated for measures including refunding advertising fees associated with hateful content and holding platforms accountable for facilitating racial discrimination and the propagation of hatred on their networks. At the individual level, events like the suicide of 14-year-old Molly Russell in 2017 in the UK, linked to posts on social platforms, underscore the grave consequences of abusive and hateful content. These occurrences underscore the escalating gravity of actions involving the propagation of racial, gender-based, and political hate speech, alongside instances of cyberbullying [2], particularly within the sphere of social media [3,2]. Within the discourse of regulating digital media, there is a growing recognition of the imperative to institute automated mechanisms capable of identifying and addressing hateful language and abusive behavior prevalent within digital media platforms. The dissemination of offensive speech and hate speech can have varying degrees of impact on

* Corresponding author.
E-mail addresses: augchao@gms.npu.edu.tw (A.F.Y. Chao), wangcs@ntut.edu.tw (C.-S. Wang), 107356508@nccu.edu.tw (B.-Y. Li), t110AB8024@ntut.org.tw (H.-Y. Chen).

https://doi.org/10.1016/j.heliyon.2024.e35468
Received 10 March 2024; Received in revised form 15 July 2024; Accepted 29 July 2024
Available online 31 July 2024
2405-8440/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

individuals and society. For example, in gender issues, Langton [4] points out that the spread of pornographic language has the potential to change public perception, unfairly portraying women as inferior, ultimately harming them. On an individual level, Matsuda [5] believes that racist speech causes psychological harm, stress, and diminished self-esteem. This conclusion is also consistent with findings in psychology [6]. Regarding public opinion, Maitra and McGowan [7] argue that hate speech can have a silencing effect on the targeted groups. From the above literature, it is evident that both offensive speech and hate speech cause harm to individuals, groups, and society. The boundary between offensive speech and hate speech is also a topic this study aims to explore.

Amid the worldwide COVID-19 pandemic in 2020, a surge in discriminatory attitudes and the proliferation of abusive and hateful language related to the pandemic became pervasive. Manifestations of racism and animosity directed towards Asian communities escalated to alarming levels, resulting in violent physical attacks. For instance, in Spain, an incident involved three individuals assaulting an Asian-American, rendering the victim unconscious for an extensive period. Similarly, in Texas, USA, a man perpetrated a knife attack against a Burmese family, attributing their Asian ethnicity as carriers of the virus. These instances of pandemic-related hate speech extend beyond singular occurrences, having been observed across diverse settings, including Taiwan's aviation industry and the specific Wanhua district in Taipei. As per Susan Sontag's analysis in "AIDS and Its Metaphors," diseases often become entwined with notions of malevolence and impurity, connecting these concepts with a sense of unfamiliarity and divergence [8]. Consequently, in times of pandemics, communities frequently encounter surges in exclusionary, fear-induced hatred. In the current age characterized by an overwhelming array of voices, depending solely on human moderation to navigate the extensive volume of hate speech pervasive on social media platforms has become ineffective. Hence, there is an exigent requirement for technological interventions to address this issue.

The ubiquity of hate speech on social media platforms presents a significant concern, and the limitations of manual moderation in effectively addressing this pervasive issue are increasingly evident. The widespread nature of hate speech, which encompasses various targeted groups, coupled with the sheer volume of content, poses substantial challenges for human-centric review processes to efficiently monitor and counteract these matters promptly. Hence, there arises an imperative need for the adoption of technological interventions to adequately confront and manage this prevailing challenge. The research landscape concerning hate speech detection demonstrates a robust presence within the English-speaking domain, notably exploring methodologies utilizing machine learning, deep learning, and Transformer models [9,10,11]. This research has been facilitated by the availability of a considerable array of research resources, including 25 extensive hate speech datasets accessible in English, alongside 6 datasets in Arabic and 5 in Italian [12]. In comparison, the collection of hate speech datasets in Traditional Chinese lacks an extensive, public collection of hate speech compared to the collection of English ones. This scarcity limits the scope of studies focusing on the identification of abusive language and hate speech within the Chinese language, rendering such investigations relatively limited in comparison [13,14].

Facebook recently publicized a proactive detection rate of 95.6 % for hate speech [15]. However, despite this high detection rate, there have been persistent user reports over the past two years detailing instances where regular posts were mistakenly flagged as violations of community guidelines. Such erroneous flagging has affected legitimate content, including cases where harmless content like three smiling face emojis triggered guideline violations. However, the reality on social media is not as such. Research indicates that social media, compared to other traditional media, shows a higher polarization index [16]. In addition to online division, social media has even fostered offline criminal behavior, such as the Mosque shootings in Christchurch Town New Zealand, where the perpetrator livestreamed on Facebook and published a 74-page manifesto on Twitter. In essence, efforts directed at identifying abusive language and hate speech within the Chinese context hold significant implications, not only for advancing natural language processing but also for shaping the dissemination of public discourse and contributing to overall societal well-being [17].

In an era marked by the widespread availability of information and the democratization of media outlets, the rapid dissemination of abusive and hateful language across social media platforms has become increasingly prominent. The swift generation of such content during emotionally charged periods, notably evident during the COVID-19 pandemic, poses significant challenges for traditional reporting methods and manual review processes, stemming from human and time constraints. The complexity of regulating abusive and hateful language is further compounded by the fundamental human right of freedom of speech in democratic societies, sparking contentious debates. This study aims to academically contribute to understanding abusive language prevalent in the Chinese-speaking realm while assisting social media platforms in enhancing their detection capabilities for such content. Moreover, it endeavors to explore the potential role of generative AI in offering non-mandatory rephrasing suggestions, thereby aiding in curtailing the proliferation of hate speech. Therefore, this research initiative aims to compile public sentiments regarding the COVID-19 pandemic from prevalent social media platforms in Taiwan with the following specific objectives:

1. Examine the patterns of abusive and hateful language that emerged on social media platforms amid the COVID-19 pandemic. Augment the existing Chinese lexicon of abusive and hateful language by manually annotating additional terms, thereby enriching the reference for future research endeavors in this specific domain.
2. Develop a two-stage model architecture to compare and evaluate their effectiveness in detecting abusive and hateful language. This endeavor aims to enhance the accurate identification of such content.
3. Develop an AI-driven system leveraging generative algorithms to identify and rephrase hateful language in real-time, providing immediate alerts to users regarding the hateful nature of their content while suggesting alternative, more constructive phrasing. This approach aims to significantly contribute to fostering a healthier and more inclusive public discourse environment by proactively addressing the propagation of hate speech.

This study presents a thorough examination of current literature on abusive language and hate speech, defining key concepts and providing examples to demonstrate the scope of research in this area. It also explores previous studies that have identified perspectives

conveying negative emotions, emphasizing common themes and viewpoints, while offering a detailed overview of the design parameters used in the investigation. The study covers all methodological aspects, including the specific methods employed and experimental results presented through detailed data analysis and tables to clearly illustrate the outcomes. Additionally, it conducts an in-depth analysis of opinions generated by large language models, focusing on removing abusive and hate speech to deepen understanding of these societal issues, concluding with key findings and proposals for future research to address gaps in the current literature and advance comprehension of this critical matter.

2. Literature review

The delineation of hate speech remains a contentious issue within the realm of international human rights law [18], lacking a universally accepted definition, a fact acknowledged by the United Nations and Council of Europe [19]. While the United Nations defines hate speech as any expression targeting individuals or groups based on identifiable characteristics—such as religion, race, nationality, ethnicity, gender, or other identity facets—using derogatory or discriminatory language or conduct, scholarly contributions, exemplified by Yong [20], offer valuable insights by categorizing hate speech into discrete classifications. Yong's framework delineates four primary categories: targeted vilification, diffuse vilification, organized political advocacy endorsing exclusionary or eliminative policies, and statements passing adverse judgments on recognizable racial or religious groups. It has been found that there is a correlation between the quantity of hate speech and emotional/demographic variables, with surprise, fear, poverty, and unemployment rates being particularly significant [21]. Additionally, Vishwamitra and colleagues utilized BERT Attention to identify new hate words related to the COVID-19 pandemic, including 186 targeting the Asian community and 100 targeting the elderly [22].

The global surge in online hate speech has prompted governments to introduce legislative measures to regulate it. France's "National Action Plan" of 2015 targets anti-LGBT hate speech online, while Germany's "Network Enforcement Act" requires prompt removal of harmful or illegal content on platforms with over 2 million users. During critical events such as the COVID-19 pandemic, governmental interventions have demonstrated a degree of efficacy in regulating online public discourse [23]. However, the discovery regarding the increase in followers presents a counterpoint to the existing literature about the spread of misinformation during the COVID-19 pandemic [24]. Furthermore, the rise and circulation of hate-fueled sentiments have the potential to warp official directives, thus worsening the crisis by spreading misguided information and amplifying its impacts.

Debates persist on the effectiveness of regulatory measures in curbing hate speech online. Major social media platforms like Twitter and Facebook have instituted community guidelines explicitly prohibiting hate speech. For instance, Facebook's latest report disclosed 38.3 million instances of hate speech from Q1 to Q3 2023. However, this surge in harmful content poses challenges and substantial costs for moderation. Managing this volume necessitates a more efficient approach, possibly leveraging machine learning. Such technology could swiftly identify objectionable content, reducing manual intervention and operational expenses. Implementing an effective machine learning system is crucial not only for promptly addressing hate speech but also for optimizing resources and lessening the economic burden of content moderation efforts.

The identification of harmful language, including hate speech and abusive content, on online platforms is a significant challenge in contemporary society. Multiple research studies on hate speech detection models were gathered and referenced in Table 1. Appropriate research materials and methods were examined for this study. Waseem and Hovy [25] established an initial framework by proposing eleven principles to identify hate speech. Their criteria provided a foundational understanding for subsequent research. Complementing this, Nobata et al. [26] introduced methodologies aimed at detecting abusive language. Their approach incorporated diverse linguistic features, employing a regression model based on Vowpal Wabbit's Regression Model. Building upon these foundations, Niemann et al. [27] conducted an extensive analysis of abusive language, delineating its multifaceted aspects such as gender discrimination, racism, threats, insults, and profanity. This comprehensive analysis reinforced the notion that abusive language encompasses various forms of hate speech, aligning with the assertions made by Fortuna and Nunes [28].

Transitioning towards machine learning models, Davidson et al. [9] compared traditional approaches like Logistic Regression, Support Vector Machine (SVM), Naive Bayes (NB), Decision Trees, and Random Forests. Their findings favored Logistic Regression and Linear SVM models, showcasing superior performance in hate speech detection. Furthering the exploration of machine learning, Zhang

Table 1
Categorization and models used in recent hate speech detection research.

Authors	Data Source	Classification Task	Models Used
Waseem & Hovy, 2016 [25]	Twitter	3 Classes (Gender Discrimination, Racial Discrimination, None)	Logistic Regression
Davidson et al., 2017 [9]	Twitter	3 Classes (Hate, abusive, Other)	SVM-Bayes-Logistic Regression, Decision Trees – Random Forests
Badjatiya et al., 2017 [29]	Twitter	3 Classes (Gender Discrimination, Racial Discrimination, None)	CNN-LSTM, FastText
Swamy et al., 2019 [30]	Twitter	3 Classes (Gender Discrimination, Racial Discrimination, None)	SVM-LSTM BERT
Nikhil et al., 2018 [31]	Facebook	3 Classes (Openly abusive, Covertly abusive, None)	LSTM-Attention
Liu et al., 2020 [32]	Facebook	2 Classes (Gender Discrimination, None)	Three Classes (Openly abusive, Covertly abusive, None)
Wang et al., 2022 [33]	LINE Today	2 Classes (positive, negative)	Lexicon, BERT

et al. [11] demonstrated the effectiveness of Convolutional Neural Network (CNN) models combined with Gated Recurrent Units (GRU). Across seven hate speech datasets, their study highlighted the superior F1-Scores achieved by this architecture compared to SVM models, indicating the promising potential of deep learning methodologies. Recent advancements in deep learning, particularly Google's BERT model, have significantly impacted natural language processing tasks. Mozafari et al. [10] leveraged BERT's capabilities, refining it with CNN, showcasing superior predictive performance on distinct hate speech datasets. Synthesizing these studies reveals an evolution from foundational principles and traditional machine learning approaches towards sophisticated deep learning architectures. While recent advancements exhibit promising results, future research should focus on addressing the challenges posed by evolving online language patterns, ethical considerations in content moderation, and the development of robust models adaptable to diverse linguistic contexts. Following table provides a summary of the classification and methodologies employed in recent studies focusing on the identification of abusive and hate speech.

3. Method

The primary objective involves the establishment of a sufficiently comprehensive dataset that captures public sentiments regarding the COVID-19 pandemic. This entails the development of machine learning mechanisms for detecting hate speech, as well as the evaluation of opinion intentions subsequent to the utilization of large language models in rephrasing hate speech. The research flow diagram, as shown in Fig. 1, depicting these processes is delineated in Graph 1, can be separated into four steps (1) data collection (green background color), (2) data preparation (blue background color), (3) 2 stage model building (yellow background color), (4) rewrite and evaluation (pink background color).

In the first step, our data compilation for constructing models emanates from the PTT platform, employing specific keywords linked to significant domestic occurrences associated with the pandemic. PTT stands out as one of the foremost terminal-based bulletin board systems, boasting an extensive user base of 3.3 million individuals in Taiwan [34]. The temporal parameters for the collection of public discourse data are predicated upon instances when these events garnered substantial attention from the populace. This approach ensures the inclusion of public opinions and concurrent instances of abusive and hateful language during periods of heightened interest. Following requisite data preprocessing, this study employs the Chinese lexicons delineating profane terminology formulated by Yang and Lin [35] and the compendium of Chinese political hate speech vocabulary established by Wang et al. [33] to filter out offensive language categorized as abusive or indicative of hate speech.

At the second step, the remaining textual content undergoes separate annotation by three autonomously recruited annotators specializing in identifying abusive language and hate speech. This process facilitates the creation of a dataset specifically focused on Chinese abusive language and hate speech detected by both lexical databases and human annotators. To ensure balance within the dataset, an undersampling technique is implemented, subsequently leading to the development of models geared towards the detection of Chinese abusive language and hate speech. In the model building step, the initial model concentrates on identifying abusive

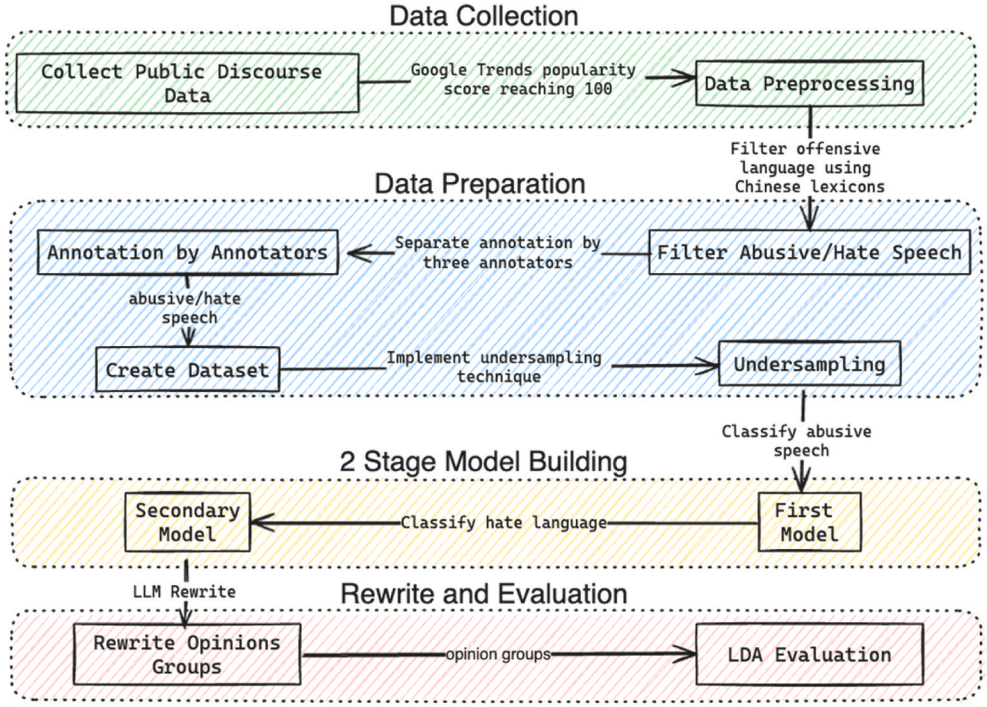


Fig. 1. Research flow.

language, encompassing instances of hate speech. Textual data categorized as abusive language are subsequently fed into a secondary model for predictive analysis. And at the final step, to mitigate the deleterious effects of hate speech prevalent in social media platforms, leveraging the generative artificial intelligence capabilities of “hidden Who" [36], within the ChatGPT framework, offers a potential solution. This approach aims to rephrase opinions in a manner that attenuates the heightened emotional tone while maintaining the essence and intended meaning [14], thereby facilitating smoother communication. An evaluation of these rephrased opinions is conducted through Latent Dirichlet Allocation (LDA) [37] topic modeling to discern and assess the alterations in discourse patterns.

3.1. Data collection and preparation

The research methodology employed in this study involves a systematic approach to curate and analyze abusive and hate speech data pertinent to the pandemic. The primary objective is to ascertain patterns of offensive language in relation to societal responses during this critical period. To accomplish this, a methodical process is undertaken for data collection and thematic classification. Firstly, three principal thematic categories directly associated with the epidemic are discerned through the analysis of Google Trends search metrics, each representing distinct facets: (1) large-scale epidemic events, (2) vaccine related issues, and (3) pandemic policy related issues. These categories serve as a foundational framework for dataset classification, listed in Table 2.

In Table 2, the dataset is meticulously compiled from the ’Gossiping’ and ’nCoV2019’ boards within the PTT platform, chosen for their active engagement and relevance during the pandemic. A temporal criterion is implemented, gathering opinions within a 14-day window to capture the dynamic sentiment landscape surrounding the discussed events. Following data aggregation, a rigorous manual annotation process ensues to ascertain sentiment within each thematic enclave. This process involves comprehensive evaluation to identify and categorize instances of abusive and hate speech. Rigorous measures are implemented to validate data reliability and representativeness. It’s important to note the study’s reliance on established methodologies in sentiment analysis and data validation to ensure robustness and credibility. The approach underscores the significance of understanding the nuances of abusive language within the context of evolving societal responses during a crisis like the pandemic.

In the course of compiling the dataset for this study, it is important to note that the textual content within the posts has been excluded from analysis. This exclusion is attributed to the prevalent inclusion of news reports or the instigation of discussion topics within these texts, consequently reducing the likelihood of encountering instances of abusive or hateful language. Furthermore, distinctive features inherent to the PTT platform, namely the "推" (upvote) and "噓" (downvote) mechanisms, have been deliberately omitted from consideration in this analysis. The main emphasis of this research is on examining the occurrence of hate speech in comments, rather than assessing the semantic orientation of comments. Users might employ hate speech irrespective of their choice to upvote or downvote a certain topic. As a result, the evaluation of hate speech in comments could be impacted by considering upvotes and downvotes information from opinions.

3.2. Two-stage model

The main goal of this study is to sort opinions into three specific types: general comments, abusive language, and hate speech. This creates a complex challenge for classification because there’s an imbalanced distribution of these types in the dataset. Abusive language and hate speech are less common compared to general comments, making it harder to accurately classify them. Fernández et al.’s [38] research highlights how classifying multiple types of comments is difficult, especially when there are unclear boundaries between the categories. When there’s an imbalance in the data, it makes this classification task even more complex. The challenges go beyond just the overlap between categories; they also affect how accurate and dependable the classification models are. The authors

Table 2
Theme and associated event for compiling public opinion datasets.

Theme	Event Code	Duration	Description
Large-Scale Epidemic Events	EVA Air	2021/4/25–2021/5/9	Outbreak among EVA Air flight attendants
	Novotel	2021/4/25–2021/5/9	Outbreak at Novotel hotel
	Tea Art House	2021/5/9–2021/5/23	Outbreak linked to a tea house
	Lions Club	2021/5/9–2021/5/23	Outbreak among members of a Lions Club chapter
	Wanhua District	2021/5/9–2021/5/23, 2021/5/23–2021/6/6	Outbreaks in the Wanhua District
	China Airlines	2021/9/5–2021/9/19	COVID-19 infections among China Airlines pilots
Vaccine Related Issues	BNT	2021/5/23–2021/6/6	Comments related to BNT vaccine
	GoodWill Clinic	2021/6/6–2021/6/20	Comments related to GoodWill Clinic vaccine
	UbiAsia	2021/8/15–2021/8/29	Comments related to UbiAsia vaccine, Taiwan local company
	Medigen Vac	2021/8/22–2021/9/5	Comments related to Medigen Vac vaccine, Taiwan local company
Pandemic Policy Related Issues	Mass Screening Policy	2020/8/23–2020/9/6	Policy requiring mass COVID-19 screening
	Retrospective Adjustment	2021/5/23–2021/6/6	Retrospective adjustment for confirmed case statistics
	3 + 11 Policy	2021/5/23–2021/6/6	3 + 11 pandemic border quarantine policy

explore how imbalanced data significantly hampers accurate classification, especially in situations with multiple categories. They suggest using binarization schemes as a potential approach to analyze datasets with uneven distributions.

To overcome the challenges associated with managing a multi-class problem and rectifying data imbalance, this research employs a two-stage model. This strategy involves the conversion of the initial multi-class problem into a two-stage binary classification. The two-stage model research related to public opinion includes the work of Faisal & Mahendra, who utilized a two-stage model for detecting false information related to the COVID-19 pandemic in Indonesian tweets. In the first stage, the model determines whether a tweet is related to the COVID-19 pandemic, filtering out unrelated data before feeding the relevant data into the second stage for false information detection [39]. The model framework proposed in this study is inspired by this literature and introduces a two-stage model for detecting offensive and hate speech. According to studies [40,28,26], hate speech is a subset of offensive speech. This study adopts this standard for labeling, using the first stage of the model to filter out non-offensive speech, and then feeding the offensive speech data identified by the first stage into the second stage for hate speech detection. The visual representation of this model's configuration can be found in Fig. 2. Building upon prior studies that suggest hate speech falls under the umbrella of abusive language [40,28,26,41], this study adopts this correlation for classification purposes. In the initial stage (blue background color), the model pinpoints instances of abusive language, thereby filtering out non-hate speech instances. Subsequently, the data identified as abusive language in the first stage are inputted into the second-stage model (green background color), specifically crafted for identifying hate speech.

In the development of models aimed at identifying abusive language and hate speech in Chinese texts during the COVID-19 pandemic, various sophisticated techniques including Support Vector Machine (SVM), Long Short-Term Memory (LSTM) models [42], Bidirectional LSTM [36] models, and BERT models were employed within a two-stage framework.

3.3. Rewriting and evaluation

Understanding and tailoring messages to specific audiences is crucial in effective communication. Large language models like ChatGPT enhance this by embodying various personas, aiding in crafting relatable messages for different audience segments [43]. In this study, we have employed the ChatGPT service to facilitate the rewriting of Chinese hate speech detected through the two-stage model, with the aim of fostering a more amicable atmosphere within the realm of public discourse. By inputting the original hate speech into the ChatGPT model, we have generated rewritten statements. Such rewriting mitigates abusive and hateful language, thereby fostering more harmonious conversations and discussions. Crafting communication strategies that prioritize transparency while steering clear of propagating hateful speech is crucial. To ensure that the input hate speech is not flagged as a violation by ChatGPT, prompting techniques are employed to provide suitable cues. We follow Zhao et al. 's [44] sequential instructions when submitting hate speech to ChatGPT for rewriting purposes. This includes clearly defining the task's objective, dividing it into smaller, detailed sub-tasks, providing brief examples for substituting hate speech with latent semantic alternatives, and organizing the content.

In this study, our initial prompt involves a specified task: "Rewrite Hate Speech." The task is subsequently outlined as follows: "Numerous hate speech instances targeting the COVID-19 pandemic and Taiwanese politics have spread on the internet. The goal is to rewrite these hate speech statements to reduce harm to others." After the task is designated, ChatGPT is instructed as follows: "The following sentences will require rewriting. Please endeavor to retain the original meaning while modifying hateful or abusive words or sentences to eliminate their harmful nature. If encountering new vocabulary related to current events, explanations will be provided." Once this process is completed, the hate speech categorized by the two-stage model can be inputted for rewriting by ChatGPT.

Finally, for an in-depth examination of the linguistic disparities between the initial and modified assertions, we utilize Latent Dirichlet Allocation (LDA) [37] topic modeling. LDA constitutes a statistical framework employed to unveil underlying themes within textual information. The categories of hate speech, abusive language, and the revised iterations of hate speech undergo individual LDA topic modeling procedures, facilitating a comparative analysis of the thematic compositions across these three delineated classifications.

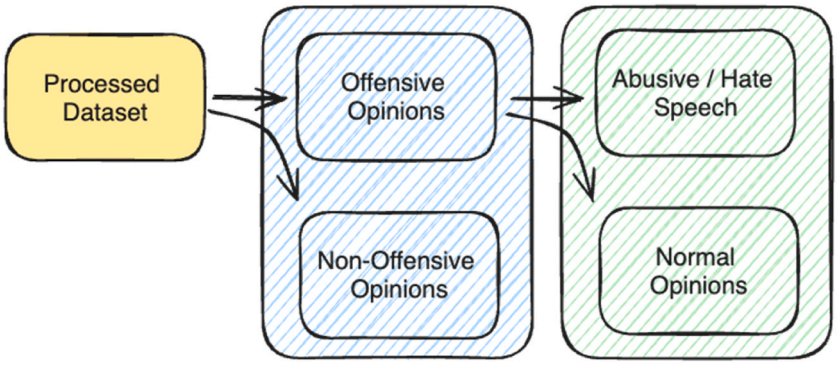


Fig. 2. Categorizing hate speech using a two-stage model in a hierarchical structure.

4. Experiment and analysis results

4.1. Dataset preparation

This research utilized a consolidated set of keywords sourced from the TOCP (NTOU Chinese Profanity) [35] in conjunction with a political hate speech lexicon developed by Wang et al., in 2022 [33] (please refer to the Appendix for detailed information). This amalgamated lexicon served as a tool to extract data from PTT within a specific timeframe for the purpose of data collection. The objective was to employ these two lexicons as a standard for evaluating abusive language and manifestations of hatred within comments, aiming to improve the efficacy of identifying and filtering offensive and hate speech. The process encompassed various stages: initially, a comprehensive examination of dictionaries to ensure their alignment with the definitions of offensive and hate speech pertinent to this study. Following validation, the offensive and hate terms referenced in the two dictionaries employed in this research were found to align with the definitions used in this context. Consequently, the integration of these dictionaries aimed to prevent the duplication of identification of offensive and hate terms already present in established dictionaries. This method aimed to conserve significant resources that would otherwise be required for manual annotation.

However, the manual annotation of offensive and hate speech still relies significantly on subjective judgments, necessitating caution to avoid issues stemming from singular viewpoints in manual labeling. To mitigate this concern, the study intends to enlist the assistance of three additional annotators to aid in the annotation process of offensive and hate speech. This initiative aims to construct a reliable training dataset. The Fleiss' Kappa coefficient [45] will be utilized to evaluate the inter-annotator agreement among multiple annotators. In the segment concerning the experimental outcomes of consistency testing, the computed Fleiss' Kappa coefficient was found to be 0.759. Nichols et al. suggest that a Kappa coefficient between 0.80 and 0.61 indicates substantial agreement among multiple annotators, revealing significant consistency in hate speech labeling among the three annotators in this study [46]. The dataset has been meticulously curated for the purpose of this study. Subsequently, this report provides comprehensive insights into the quantity of collected data (see Table 3), the ratio between abusive and non-abusive language, and the prevalence of abusive language in relation to instances of hate speech samples.

During the annotation process, it was observed that certain terms were not previously enlisted in past research endeavors. However, these terms emerged frequently in comments during the pandemic period and carried deterministic sentiment in opinions. Subsequently, during the manual annotation process, these terms were integrated into the compiled lexicon. The inclusive list is presented as follows (see Table 4):

4.2. Two-stage model construction

This experiment employs a two-stage model for detecting abusive language and hate speech. In the first stage, abusive language detection is performed, followed by hate speech detection using the data predicted by the first-stage model. Four models, including (1) Support Vector Machine (SVM), (2) Long Short-Term Memory (LSTM), (3) Bidirectional LSTM (Bi-LSTM), and (4) BERT, are used in both stages. The best parameter combinations from the first stage are selected for inputting into the second-stage model. The best-performing model is determined in the second stage based on its predictive performance. The optimal model parameter combinations for both stages are identified. To evaluate across 4 models, we employed collected dataset from PPT in the model development process. Initially, a partition of 60 % for training and 40 % for testing facilitated a substantial dataset retention for subsequent modeling stages while preserving acceptable performance in the initial model. Subsequently, in the second stage, an 80 % training and 20 % testing split was utilized to ensure ample foundational data availability for refining the model. We use following performance matrices to evaluate the performance of the models:

Table 3
Collected dataset after annotation.

Theme	Event Code	Positive Annotated Samples	Non-Positive Annotated Samples	All Samples	Positive Sample Ratio
Large-Scale Epidemic Events	EVA Air	6	970	976	0.61 %
	Novotel	6	288	294	2.04 %
	Tea Art House	9	409	418	2.15 %
	Lions Club	36	856	892	4.04 %
	Wanhua District	65	541	606	10.73 %
	China Airlines	11	54	986	1.12 %
Theme-1 Subtotal		**133**	**3118**	**4172**	**3.19 %**
Vaccine Related Issues	BNT	96	1921	2017	4.76 %
	GoodWill Clinic	137	2830	2967	4.62 %
	UbiAsia	0	374	374	0.00 %
	Medigen Vac	66	2908	2974	2.22 %
Theme-2 Subtotal		299	8033	8332	**3.59 %**
Pandemic Policy Related Issues	Mass Screening Policy	5	284	289	1.73 %
	Retrospective Adjustment	13	1229	1242	1.05 %
	3 + 11 Policy	34	2898	2932	1.16 %
Theme-3 Subtotal		52	4411	4463	**1.17 %**

Table 4
Collected dataset after annotation.

Group	Terms	Description/Translation
1	菸粉, 塔綠班, 綠共蟑螂腦, 垃圾民盡擋	Derogatory term for DPP (Democratic Progressive Party) and followers
2	側翼蟑螂, 綠蟑螂, 綠狗, 側翼網軍, 綠畜蟑螂	
3	狗屎中, 范雲病毒, 柯糞	Insulting nicknames aimed at politicians that are considered contemptible
4	支那仔, 中共同路人	Insults directed at those perceived as pro-China
5	台派黑道賤畜, 失智列車	The derogatory terms derived from current affairs.

- True Positive (TP): The model correctly predicted offensive/hate speech.
- False Negative (FN): The model incorrectly predicted non-offensive/non-hate (normal) speech when it was offensive speech.
- False Positive (FP): The model incorrectly predicted offensive/hate speech when it was non-offensive speech.
- True Negative (TN): The model correctly predicted non-offensive/non-hate (normal) speech.

The confusion matrix categorizes the model's performance by presenting results into four cases: True Positives (TP), False Negatives (FN) or Type-II error, False Positives (FP) or Type-I error, and True Negatives (TN). A perfect model would exhibit a substantial number of TPs and TNs, accurately distinguishing between offensive/hate speech and non-offensive/normal speech. Conversely, a high quantity of FNs (Type-II error) indicates challenges in recognizing offensive/hate speech, resulting in missed detections. Likewise, an elevated number of FPs (Type-I error) signifies misidentifying non-offensive/normal speech as offensive/hate speech, leading to false alarms.

- Accuracy is the ratio in all data where the actual offensive/hate speech is predicted as offensive/hate speech, and the actual normal speech is predicted as normal speech. It is calculated as: Accuracy = (TP + TN)/(TP + FP + FN + TN)
- Precision is the ratio of actual offensive/hate speech among the data predicted by the model as offensive/hate speech. It is calculated as: Precision = TP/(TP + FP)
- Recall is the proportion of actual offensive/hate speech data that is predicted by the model as offensive/hate speech. It is calculated as: Recall = TP/(TP + FN)
- F1-score is the harmonic mean of precision and recall, which provides a combined score for these two metrics to prevent the case where precision is too high and recall is too low, or precision is too low, and recall is too high. It is calculated as: F1-score = 2 * (precision * recall)/(precision + recall)

(1) Support Vector Machine

In this experiment, the two-stage model was employed for model construction utilizing Support Vector Machine (SVM). The parameters utilized included the Gaussian kernel (*RBF*) with a *gamma* value ranging from *0.1* to *0.3*. Additionally, the regularization parameter (*C*) was selected within the range of *1* to *3* for comparative analysis. The comparison results show in Table 5.

After a thorough comparison of the parameter set used in constructing a two-stage model, it has been determined that the combination of $C = 2$ and $gamma = 0.2$ in the first stage, along with $C = 1$ and $gamma = 0.1$ in the second stage, will yield the optimal accuracy and f1-score combination in the results. Specifically, the precision is registered at *74.87 %* and the f1-score is *77.24 %* for the first stage, while the precision is *54.05 %* and the f1-score is *65.31 %* for the second stage. In this experimental study, it becomes clear that the representative (SVM) model might misclassify certain non-hateful expressions as hate speech due to their similar features. In spite of that, the model showcases a relatively high recall rate, which implies its ability to effectively identify instances of hate speech.

In Figs. 3 and 4, SVM model with parameters set *kernel* = 'rbf', $C = 2$, $gamma = 0.1$ has 70 false positive samples and 29 false negative samples in first stage, and 15 false positive samples and 2 negative samples. From the results of the confusion matrix, the model's precision is relatively low, resulting in more Type I errors. This indicates that the model may have misjudged some features of non-hate speech and incorrectly classified them as hate speech. However, the model's recall rate is relatively high, showing that its ability to identify hate speech is relatively strong.

Table 5
Experiment Results: Two-Stage Model (SVM).

First-stage Results

Parameter Setting	Accuracy	Precision	Recall	F1-score
kernel = 'rbf', C = 1, gamma = 0.1	70.30%	65.04%	87.82%	74.73%
kernel = 'rbf', C = 1, gamma = 0.2	67.51%	63.42%	82.74%	71.81%
kernel = 'rbf', C = 1, gamma = 0.3	63.96%	60.62%	79.70%	68.86%
kernel = 'rbf', C = 2, gamma = 0.1	74.87%	70.59%	85.28%	77.24%
kernel = 'rbf', C = 2, gamma = 0.2	71.57%	68.09%	81.22%	74.07%
kernel = 'rbf', C = 2, gamma = 0.3	67.26%	64.17%	78.17%	70.48%
kernel = 'rbf', C = 3, gamma = 0.1	74.11%	70.21%	83.76%	76.39%
kernel = 'rbf', C = 3, gamma = 0.2	70.56%	67.23%	80.20%	73.15%
kernel = 'rbf', C = 3, gamma = 0.3	67.01%	63.90%	78.17%	70.32%

Second-stage Results

Parameter Setting	Accuracy	Precision	Recall	F1-score
kernel = 'rbf', C = 1, gamma = 0.1	54.05%	51.61%	88.89%	65.31%
kernel = 'rbf', C = 1, gamma = 0.2	62.16%	75.00%	33.33%	46.15%
kernel = 'rbf', C = 1, gamma = 0.3	59.46%	71.43%	27.78%	40.00%
kernel = 'rbf', C = 2, gamma = 0.1	62.16%	66.67%	44.44%	53.33%
kernel = 'rbf', C = 2, gamma = 0.2	64.86%	72.73%	44.44%	55.17%
kernel = 'rbf', C = 2, gamma = 0.3	64.86%	72.73%	44.44%	55.17%
kernel = 'rbf', C = 3, gamma = 0.1	64.86%	69.23%	50.00%	58.06%
kernel = 'rbf', C = 3, gamma = 0.2	67.57%	75.00%	50.00%	60.00%
kernel = 'rbf', C = 3, gamma = 0.3	64.86%	72.73%	44.44%	55.17%

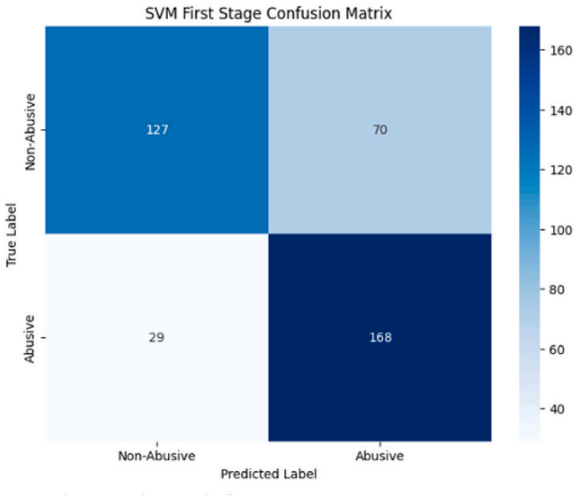


Fig. 3. Confusion matrix for SVM in first stage.

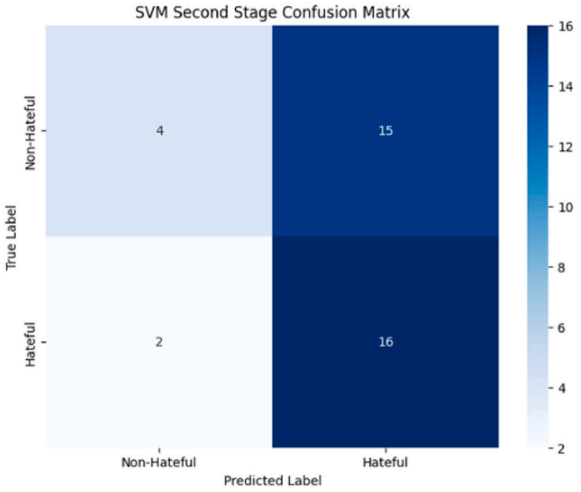


Fig. 4. Confusion matrix for SVM in second stage.

(2) Long Short-Term Memory

In the experiment of the LSTM two-stage model, the Dropout mechanism is utilized to address the problem of overfitting [47]. Prior studies by Khan et al. [48] utilized a Dropout rate of *0.3*, whereas Baldi and Sadowski [37] recommended setting this value to *0.5*. To facilitate comparative LSTM experimentation within a two-stage model, a Dropout rate of *0.5* was employed in this study. Additionally, the optimizer plays a crucial role in adjusting neural network parameters to minimize error. Specifically, this study adopts the widely utilized Adam optimizer within the LSTM model framework. The learning rate (*lr*) parameter governs the pace of the model's learning process. Excessively high learning rates can impede model convergence, resulting in gradient explosions, while very low rates

Table 6
Experiment results: Two-stage Model (LSTM).

First-stage Results

Parameter Setting	Accuracy	Precision	Recall	F1-score
Optimizer = Adam, *lr* = 0.1,	65.48%	63.20%	74.11%	68.22%
Optimizer = Adam, *lr* = 0.01	69.54%	74.21%	59.90%	66.29%
Optimizer = Adam, *lr* = 0.001	74.87%	72.07%	81.22%	76.37%

Second-stage Results

Parameter Setting	Accuracy	Precision	Recall	F1-score
Optimizer = Adam, *lr* = 0.1	50.00%	50.00%	31.25%	38.46%
Optimizer = Adam, *lr* = 0.01	53.12%	54.55%	37.50%	44.44%
Optimizer = Adam, *lr* = 0.001	65.62%	63.16%	75.00%	68.57%

can cause slow convergence and potential overfitting. Prior experiments [49] have suggested scholars adopting a learning rate of *0.01*. Nevertheless, considering the constrained nature of the training dataset and the specific context of hate speech detection, our experiment in this study involves an examination and comparative analysis of learning rates (*lr*) spanning from *0.1*, *0.01*, to *0.001*.

The optimal parameter set identified in the second stage remained consistent with the model established in the first stage, indicating its effectiveness in classifying offensive and hate speech (optimizer = Adam, $lr = 0.001$). Furthermore, experimental outcomes demonstrated that the implementation of the LSTM model, in comparison to the SVM model, exhibited improved discrimination capabilities for differentiating offensive and hate speech. Specifically, the results showed a sustained accuracy of 74.87 % and an F1-score of 76.37 % in the first stage, while in the second stage, the accuracy remained at 65.62 % with an F1-score of 68.57 %, (see Table 6).

In Figs. 5 and 6, LSTM model with parameters set optimizer = Adam, $lr = 0.001$ has fewer Type-I errors with 62 samples and higher Type-II errors 37 samples compared to SVM model. That indicates higher precision and lower the recall ratio, and almost the performance in F1-socre in both first stage and second stage.

(3) Bidirectional LSTM

The Bi-directional Long Short-Term Memory network (Bi-LSTM) [36] represents a variant of the preceding experiment, incorporating two Long Short-Term Memory (LSTM) layers. This architectural configuration enables the language model to engage in both forward and backward semantic learning concerning the sequence of text, specifically aimed at detecting offensive and hate speech within opinions. The Bi-LSTM structure, compared to a singular LSTM layer, is notably more intricate, resulting in increased computational time. Nonetheless, in the course of experimental procedures, the parameters employed remain consistent with those used for a single-layer LSTM, encompassing a dropout rate of *0.5*, Adam optimizer, and a learning rate ranging (*lr*) from *0.1*, *0.01*, and *0.001*. These parameter settings allow for a more precise comparative analysis of experimental outcomes against the performance of the single-layer LSTM. The findings are presented in Table 7.

The performance of the bidirectional long short-term memory (Bi-LSTM) model closely resembled that of previous LSTM model, indicating that employing a bidirectional framework did not yield significant advantages in this context. One speculated reason for this observation could be attributed to the potentially smaller size of the dataset, resulting in an insufficient number of samples corresponding to various forms of abusive language and hate speech.

Moreover, the Bi-LSTM, utilizing two LSTM layers, is inherently more intricate than the single-layer LSTM, and in cases of limited training samples, the use of a more complex model might lead to overfitting. Another contributing factor is the shorter length of the textual sequences; the comments related to the pandemic collected in this study, owing to platform-specific characteristics and a prevalence of expressing anger or discontent through hate speech, typically contain comparatively less information. Consequently, the utilization of Bi-LSTM may not confer substantial advantages in such contexts.

According to the performance metrics above, in Figs. 7 and 8, we know that the performance of the Bi-LSTM and LSTM models are closely resembled, but the Bi-LSTM model produces fewer Type-I errors and more Type-II errors than the LSTM model in the first stage, shown in Fig. 5. Additionally, the Bi-LSTM model has a better ability to identify normal speech (more true negative samples) than the LSTM model. However, in the second stage, we can see that the confusion matrices are identical.

(4) Bidirectional Encoder Representations from Transformers, BERT

The pre-trained model utilized in BERT is bert-base-chinese [50], comprising 12 hidden layers, each containing 768 hidden units.

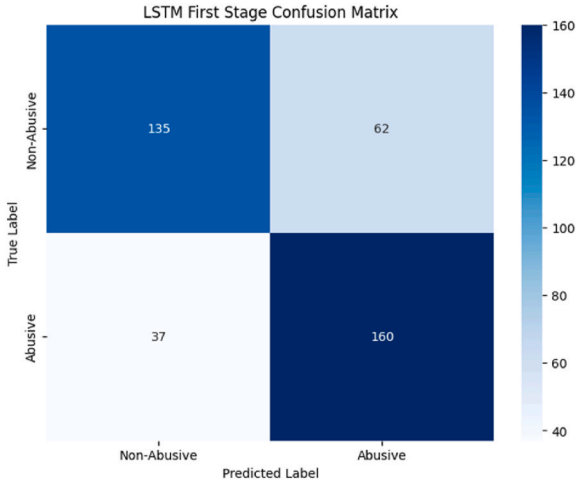


Fig. 5. Confusion matrix for LSTM in first stage.

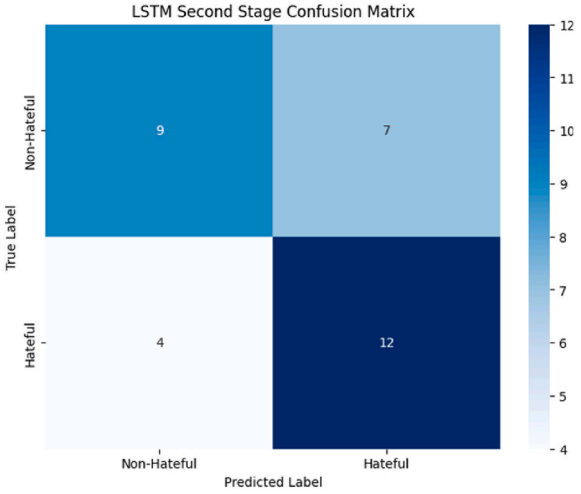


Fig. 6. Confusion matrix for LSTM in second stage.

Table 7
Experiment results: Two-stage Model (Bi-LSTM).

First-stage Results

Parameter Setting	Accuracy	Precision	Recall	F1-score
Optimizer = Adam, *lr* = 0.1	66.75%	64.60%	74.11%	69.03%
Optimizer = Adam, *lr* = 0.01	68.78%	68.32%	70.05%	69.17%
Optimizer = Adam, *lr* = 0.001	75.63%	77.90%	71.57%	74.60 %

Second-stage Results

Parameter Setting	Accuracy	Precision	Recall	F1-score
Optimizer = Adam, *lr* =0.1	59.38%	55.17%	100.00%	71.11%
Optimizer = Adam, *lr* = 0.01	53.12%	52.38%	68.75%	59.46%
Optimizer = Adam, *lr* = 0.001	65.62%	63.16%	75.00%	68.57%

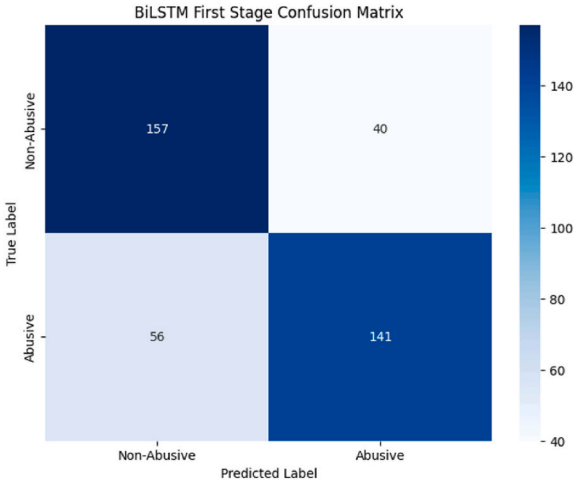


Fig. 7. Confusion matrix for Bi-LSTM in first stage.

For the task of classification, the *BertForSequenceClassification* model was employed to detect abusive and hate speech. In hyper-parameter configuration, batch sizes of *16* and *32* were utilized, coupled with learning rates (*lr*) set at $2e^{-5}$, $3e^{-5}$, and $5e^{-5}$, respectively, to conduct a comparative analysis of the two-stage model, aimed at evaluating the experimental outcomes among various algorithms. To circumvent overfitting while ensuring adequate model convergence, the experimental setup encompassed *4* iterations (*epoch*). The parameter configurations were proposed in accordance with the fine-tuning parameters recommended in the study conducted by Devlin et al. [50].

In the comparative analysis, see results in Table 8, the first-stage achieved an accuracy of 94.42 % and attained the highest F1-score with hyper-parameter settings of batch size $= 16$ and $lr = 2e^{-5}$. Diverging from the outcomes observed in the second-stage, previous experimentation indicates a decrease in both accuracy and F1-score relative to the first-stage. Nonetheless, employing the BERT

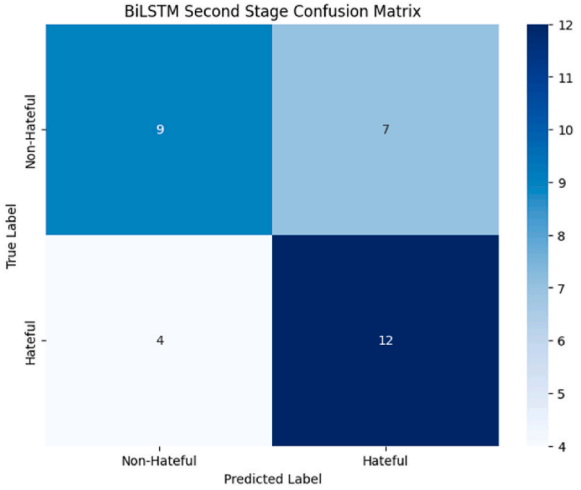


Fig. 8. Confusion matrix for Bi-LSTM in second stage.

Table 8
Experiment results: Two-stage Model (BERT).

First-stage Results

Parameter Setting	Accuracy	Precision	Recall	F1-score
Batch size=16, *lr*=2e-5	94.42%	95.79%	92.86%	94.30%
Batch size=16, *lr*=3e-5	90.86%	90.82%	90.82%	90.82%
Batch size=16, *lr*=5e-5	89.85%	86.79%	93.88%	90.20%
Batch size=32, *lr* =2e-5	90.86%	85.71%	97.96%	91.43%
Batch size=32, *lr* =3e-5	86.29%	81.42%	93.88%	87.20%
Batch size=32, *lr* =5e-5	86.29%	97.33%	74.49%	84.39%

Second-stage Results

Parameter Setting	Accuracy	Precision	Recall	F1-score
Batch size=16, *lr* =2e-5	74.07%	83.33%	66.67%	74.07%
Batch size=16, *lr* =3e-5	66.67%	71.43%	66.67%	68.97%
Batch size=16, *lr* =5e-5	81.48%	85.71%	80.00%	82.76%
Batch size=32, *lr* =2e-5	62.96%	66.67%	66.67%	66.67%
Batch size=32, *lr* =3e-5	62.96%	64.71%	73.33%	68.75%
Batch size=32, *lr* =5e-5	51.85%	55.00%	73.33%	62.86%

algorithm for classification demonstrates sustained performance with an accuracy of 84.48 % and an F1-score of 84.76 % (using batch size $= 16$ and $lr = 5e^{-5}$). The experimental findings substantiate the BERT model's robustness in effectively distinguishing between offensive and non-offensive language in a two-stage classification process, exhibiting high accuracy and commendable performance.

From Figs. 9 and 10, the BERT model performs well in classifying hateful speech in the first stage and hateful speech in the second stage among all models, with very low Type-I and Type-II errors.

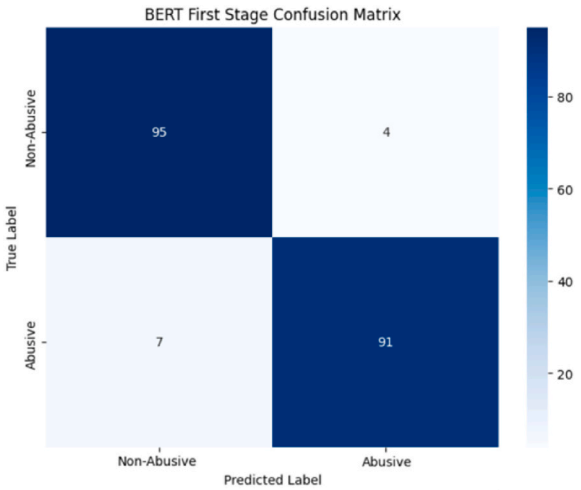


Fig. 9. Confusion matrix for BERT in first stage.

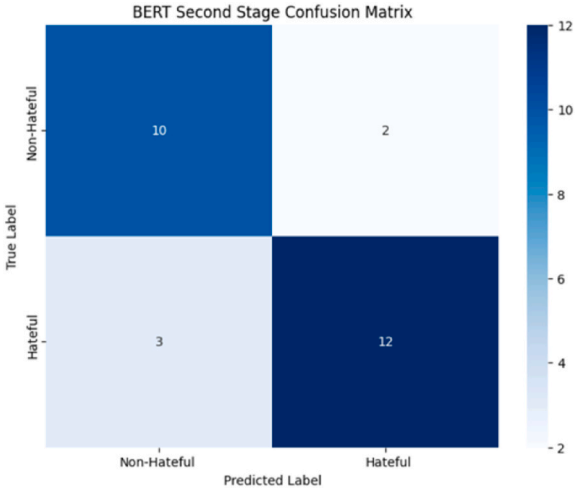


Fig. 10. Confusion matrix for BERT in second stage.

In summary, this study explored four models - SVM, LSTM, Bi-LSTM, and BERT - within a two-stage framework for the identification of offensive language and hate speech. The experimental results revealed that the BERT model achieved the highest level of performance. In contrast, the SVM, LSTM, and Bi-LSTM models exhibited lesser predictive capabilities, suggesting the dominance of BERT's pre-trained language representations in identifying offensive language and hate speech. These findings offer valuable insights into the development of robust two-stage architectures that utilize advanced neural networks like BERT to accurately identify various forms of abusive content in online texts. Future research could explore the combination of multiple models and the incorporation of additional semantic features to further enhance detection performance.

4.3. Rewriting and evaluation

In the hate speech rewriting experiment, ChatGPT is employed to revise identified samples of hate speech generated by the two-stage model. We input original hate speech into the ChatGPT and obtained revised expressions. This form of rephrasing contributes to the reduction of offensive and aggressive language, fostering more harmonious dialogue and discussions. To prevent the input of hate speech from being flagged by ChatGPT as a violation and only outputting warning messages, prompting methods using suitable cues are necessary. This study first inputs a specified task into the ChatGPT dialogue box: "rewriting hate speech". The task is described as follows: "There is a lot of hate speech related to the COVID-19 pandemic and Taiwanese politics circulating online. We want to rewrite this hate speech to make it less harmful to others." After specifying the task, ChatGPT is prompted: "Next, we will provide sentences that need to be rewritten. Please retain the original meaning as much as possible while modifying any hateful or offensive words or phrases to make them non-hateful. If there are issues related to the timeliness of the information or new terms, explanations will be provided." Once this process is complete, the hate speech identified by the study's two-stage model classification can be pasted for ChatGPT to rewrite. In Table 9, we selectively present some original texts and their corresponding rephrased outcomes, along with an analysis of the methods and linguistic features employed in the rephrasing process.

The categorization of hate speech rephrasing strategies can be consolidated into three primary categories to facilitate a more systematic analysis and understanding of these methods. The first category, "Semantic Substitution", includes strategies such as metaphorical substitution and the use of indefinite pronouns. This category involves altering or concealing the original hateful language's direct meaning by employing metaphors or vague pronouns to reduce its offensiveness. The second category, "Structural Adjustment", encompasses generalization terms and sentence restructuring (removing hatred terms). This approach focuses on modifying the linguistic structure or using broader, more general terms to soften or eliminate hateful elements, thereby making the language more neutral or ambiguous. The third category, Perspective Shift, involves changing the narrator's or observer's viewpoint to transform the overall context and diminish the hatefulness of the speech.

Since we provide only one prompt for all instances of hate speech, different styles of writing strategies will emerge to reduce hate speech. From Table 9, the selected results show that LLM employs various methods to rewrite hate speech, encompassing semantic substitution (e.g., replace "五毛"(China) to "中國"(China)), utilizing indefinite pronoun or noumenon references for commentary modification, and sentence restructure (removing hatred terms) as part of the rewriting process. After the process of rewritten statements, results show that there is an absence of terms associated with hatred within the revised sentences. The phrasing uses generalization terms instead of negative descriptors targeting specific groups. It rewrites user opinions in a more positive way by avoiding harmful language. Furthermore, initial comparative analyses suggest that the modified content retains the primary message intended by the user in their original comments and changes the perspective of user reviews from objective to subjective (perspective shift).

To conduct a comprehensive examination of linguistic variances between the initial and modified statements, we utilize LDA topic

Table 9
Selected rewritten statement.

Rephrasing Strategies	#	Hate Speech	Rewritten Statement
Semantic Substitution	1	不要說三大廠 連**五毛**自家疫苗都要過	不要說三大廠 連**中國**自家疫苗都要過
		Translation: Despite avoiding mention of top three manufacturers, Chinese-made vaccines still need to face scrutiny. Rewrite Differences: 1 Metaphorical substitution: Replace "五毛" to "中國". replace derogatory terms related to China.	
Semantic Substitution	2	反正台派**黑道賤畜**會一直造謠國外的都沒三期	不論立場如何,**有些人**持續在傳播關於國外疫情控制措施的誤解與謠言。
		Translation: Regardless of their stance, some people continue to spread misunderstandings and rumors about foreign epidemic control measures. Rewrite Differences: 1 indefinite pronouns: Replace "黑道賤畜" to "有些人", use indefinite pronouns and ignore derogatory term "黑道."	
Semantic Substitution	5	推彙整 **綠共吱蟑**真的都以為沒人會查?	我認為我們應該推動彙整和檢查,以確保**政黨或團體**的行為符合法律要求。
		Translation: Advocating for audits to ensure political entities adhere to legal standards. Rewrite Differences: 1 metaphorical substitution: Replace "綠共吱蟑" to "政黨或團體", replace with noumenon, the political party.	
Structural Adjustment	3	台灣**自私垃圾老人**真的很多	台灣的一些**年長人士**可能存在一些**自私的行為**,這讓人感到困擾。
		Translation: Elderly in Taiwan exhibit troubling, possibly selfish behavior, causing concern. Rewrite Differences: 1 Generalization terms: Replace "老人" to "年長人士.". 2 Restructuring sentences (removing hatred terms): Revised the structure of opinions while disregarding the derogatory term "垃圾."	
Structural Adjustment	6	**犯婦**!還不出來跟萬華人磕頭道歉!	我認為**這位立委**應該為其推動的政策向萬華人道歉
		Translation: A legislator should apologize to the Wanhua district for advocating policies. Rewrite Differences: 1 Generalization terms: Replace "犯婦" to "這位立委", replace with pronouns. 2 Sentence restructuring (removing hatred terms): Replace "磕頭道歉" to "道歉".	
Structural Adjustment	7	可惡又是**機師控制不住小頭**,**空姐控制不了海鮮** … …	有些機師和空姐似乎**無法妥善遵守防疫規定** … …
		Translation: Some pilots and flight attendants seem non-compliant with pandemic protocols. Rewrite Differences: 1 Restructuring sentences (removing hatred terms): Revised the opinions while disregarding the derogatory terms "小頭" and "海鮮."	
Structural Adjustment	9	**拉機蟑螂**電視台	**某些**電視台傳遞的訊息可能**具有偏頗性**。
		Translation: Some TV stations possibly deliver biased information to viewers. Rewrite Differences: 1 Generalization terms: Replace "拉機蟑螂" to "某些", use indefinite pronouns. 2 Restructuring sentences (removing hatred terms): Replace "拉機蟑螂" to "具有偏頗性", revised opinion with direct statements.	
Mix Strategy(Structural Adjustment, Perspective Shift)	8	**側翼蟑螂**拜託都去打國產,**不要佔用別人**	**建議大家**都去接種國產疫苗,**這樣可以為他人釋放出更多的疫苗供應**。
		Translation: Encouraging widespread use of domestic vaccines to increase overall supply. Rewrite Terms: Replace 1 Sentence restructuring (removing hatred terms): "側翼蟑螂" to "建議大家", replace with general pronouns and revised in positive way. 2 Perspective shift: The rewritten sentence becomes more rational and objective, emphasizing the positive effect of getting vaccinated with domestic vaccines, which is to free up more vaccine supplies for others.	
Mix Strategy(Structural Adjustment, Perspective Shift)	10	**藍腦 跟 五毛** 好像**很怕**國產疫苗	**國民黨支持者**對國產疫苗**表現出一定程度的擔憂或疑慮**。
		Translation: Supporters of the opposition party express hesitancy toward domestically produced vaccines. Rewrite Differences: 1 Generalization terms: Replace "藍腦 跟 五毛" to "國民黨支持者", use noumenon and revised opinion with direct statements. 2 Perspective shift: The rewritten sentence becomes more rational and objective, replace "很怕" to "表現出一定程度的擔憂或疑慮".	
Mix Strategy(Semantic Substitution, Structural Adjustment, Perspective Shift)	4	幾隻**綠共**長官在上面啊 嘻嘻	有些**執政黨的**高級官員參與其中,這是一個事實。
		Translation: High-ranking officials from some ruling parties are involved, a confirmed fact. Rewrite Differences: 1 Indefinite pronouns: Replace "幾隻" to "有些", replace with noumenon. 2 Generalization terms: Replace "綠共" to "執政黨的", replace with noumenon, the ruling party. 3 Perspective shift: Replace "嘻嘻" with "這是一個事實" to change the condescending and aggressive tone to an objective stance of stating a fact.	

modeling to discern dissimilarities in thematic structures and present topic analysis in Table 10. The results clarify that the updated statements maintain the main goal of discussing topics related to COVID-19, although they demonstrate significant changes in the way the information is presented. The initial hate speech mainly focuses on themes centered on bias, fear, and hostility, while the revised statements lean towards a more rational and informative storytelling.

Upon comparative analysis of various thematic datasets, distinct keywords emerge in each dataset, delineating unique characteristics. For instance, the primary theme predominantly features discussions on pilot infection-related occurrences, while the secondary theme centers on discourse related to domestic vaccines. The tertiary theme, however, focuses on the ruling political party and its digital adherents. Notably, this phenomenon is more pronounced during the initial phase (section 1) but diminishes in significance during the subsequent stage. This discrepancy primarily arises due to the heightened emphasis on detecting abusive or hate speech during the second stage. Consequently, the commentary data prioritize information pertaining to filtered offensive opinions, resulting in a dearth of crucial vocabulary representative of that particular thematic domain.

Comparing the LDA keywords between abusive language and hate speech, it's evident that while abusive language includes negative terms like "靠北" (complaint), "屌" (vulgar term), "廢物" (trash), and "側翼" (sidekick), the negative terms in hate speech like "賤畜" (despicable animal), "綠蟑螂" (green cockroach), and "垃圾" (garbage) are more intense in section 1 and 2 in Table 10. The noticeable difference in both the strength and specificity of the keywords used in abusive language compared to hate speech is clear. Furthermore, the results obtained from the LDA clustering technique show that the topics identified in this study - namely, significant occurrences of epidemics, concerns related to vaccines, and matters concerning epidemic prevention policies - do not have distinct boundaries within the LDA topics. This suggests that when it comes to discussions among users, the use of abusive language and hate speech regarding epidemic-related issues does not have a clear separation based on different subjects. In general, the abusive language and hate speech related to the epidemic mainly revolve around criticizing the ruling party and its supporters, as well as specific groups affected by the epidemic.

The results subsequent to the rewriting conducted by ChatGPT (outlined in section 3) exhibit notable disparities from the preceding iterations. The rephrased content not only expunges derogatory expressions that possess the potential to offend or inflict harm upon the general populace but also effectively retains the primary purport and thematic elements encompassing the public's commentary on pandemic-related subjects. Overall, as per the findings derived from the LDA model, the modified discourse not only eliminates specific inflammatory language but also substantially preserves the essence of the public discourse and its discussions on pandemic-related themes.

5. Conclusion and future discussion

The rise of social media has provided an open space for everyone to voice their opinions; however, its openness has also given rise to

Table 10
Topic analysis results.

Section 1: Topic analysis from first stage (offensive detection)	
Theme	Keywords
Large-Scale Epidemic Events	"機師"(pilot)、"萬華"(location)、"隔離"(quarantine)、"病毒"(virus)、"台灣"(Taiwan, location)、"偷渡"(illegal border crossing)、"高端"(Medigen Vac)、"14"、"靠北"(complaint, pejorative words)、"防疫"(epidemic prevention)
Vaccine Related Issues	"病毒"(virus)、"台灣"(Taiwan, location)、"政府"(government)、"輝瑞"(Pfizer)、"美國"(USA)、"還好"(fine, *ambiguous*)、"扯"(nonsense)、"華航"(China airlines) 、"造謠"(spreading rumors) 、"14"
Pandemic Policy Related Issues	"傳染"(infection)、"中央"(government)、"媒體"(media)、"屌"(vulgar term)、"台北"(Taipei, *location*) 、"地方"(area)、"管理"(management)、"道歉"(apologize) 、"接觸"(contact)、"北市"(Taipei, *location*)
Section 2: Topic analysis from second stage (abusive/hate speech detection)	
Theme	Keywords
Large-Scale Epidemic Events	"垃圾"(garbage)、"中央"(government)、"病毒"(virus)、"造謠"(spreading rumors)、"側翼"(sidekick)、"隔壁"(China, metaphorical speaking)、"綠蟑螂"(green cockroach, pejorative references for the DPP)、"噁心"(disgusting)、"扯"(nonsense)、"對立"(opposite side)
Vaccine Related Issues	"綠共"(Pejorative references for the DPP)、"病毒"(virus)、"蟑螂"(cockroach, pejorative references for the DPP)、"三立"(media channel)、"綠蟑螂"(green cockroach, pejorative references for the DPP)、"萬華"(location)、"14"、"黨"(political party)、"支那"(China)、"噁心"(disgusting)
Pandemic Policy Related Issues	"媒體"(media)、"垃圾"(garbage)、"中央"(government)、"綠共"(green communists)、"預約"(reserved)、"萬華"(*location*)、"黨"(political party)、"賤畜"(derogatory terms referring to supporters)、"失智列車"(derogatory terms derived from current affairs)、"建置"(establish)
Section 3: Topic analysis from LLM rewritten opinions	
Topic	Keywords
Large-Scale Epidemic Events	"防疫"(against epidemics)、"似乎"(look like)、"感到"(feel like)、"規定"(regulation)、"支持者"(supporter)、**"機師"(pilot)**、"綠營"(the DPP)、"回應"(responded)、"這讓"(let, *ambiguous*)、"時"(when, *ambiguous*)
Vaccine Related Issues	"支持者"(supporters)、**"國產疫苗"(domestic vaccine)**、"感到"(feel)、"認為"(consider)、"綠營"(the DPP)、"政黨"(political party)、"言論"(opinions)、"疫情"(pandemic)、"中國"(China)、"特定"(specific)
Pandemic Policy Related Issues	"感到"(feel)、"網路"(network)、"綠營"(the DPP)、"令人"(make people feel)、"不滿"(not satisfied)、"存在"(exist)、"認為"(consider)、"台灣"(Taiwan)、"黨"(political party)、"表現"(performance)

various issues. Among these, the propagation of abusive language and hate speech on social media has posed serious threats to society. During the global COVID-19 pandemic outbreak in 2020, discrimination and the subsequent spread of abusive language and hate speech related to the epidemic exacerbated the situation. Therefore, the establishment of a detection system for abusive language and hate speech using machine learning models has become a crucial objective for many researchers.

This study endeavors to compile and expand a lexicon that includes abusive and hate speech specifically associated with the COVID-19 pandemic. Once we gathered three themes of relevant opinions, three annotators were enlisted to independently label the data. The annotation consistency, as measured by the Fleiss Kappa value, reached 0.759, indicating that the labeling results were consistent and successfully establishing a dataset and lexicon. We developed a two-stage model and employed various machine learning models such as SVM, LSTM, Bi-LSTM, and BERT to identify instances of abusive and hate speech within opinions. The results revealed that the BERT model demonstrated the highest performance, with the accuracy of the first-stage model reaching 94.42 % and the accuracy of the second-stage model reaching 81.48 %. This demonstrates the effectiveness of this model architecture in efficiently identifying abusive and hate speech in opinions. This study presents a method that utilizes generative AI to rewrite hate speech, showcasing its effectiveness in removing offensive language while preserving the core essence of public discussions related to the pandemic. A LDA method was employed for conducting thematic analysis. The results showed that by employing ChatGPT to paraphrase comments with hate speech, different tactics such as metaphorical substitution, replacing generalization terms, restructuring sentences to remove hateful language, and shifting from objective to subjective perspective were used. The rephrased comments remained pertinent to the original subject matter. This demonstration highlights that rephrasing effectively reduces hateful language while preserving the original intent of communication among users. The findings have significant implications for society and online platforms, as they offer a potential solution to address abusive content. By utilizing appropriate lexicons in different languages, such as adopting existing lexicons from Italian [12] or following the lexicon generation process outlined in Fortuna and Nunes's [28] survey, we believe the results of the proposed two-stage approach can empower platform administrators to better manage hate speech. This, in turn, will ultimately foster a more constructive and positive community environment for users.

While the model proposed shows promising results in identifying and dealing with abusive language and hate speech linked to the COVID-19 outbreak, it does have its limitations. One key limitation is its reliance on existing lexicons or methods for generating lexicons, which may not keep up with the changing nature of abusive language and hate speech in online discussions. The model's success heavily relies on the quality and extent of the lexicon utilized, which can vary among different languages and cultural settings, restricting its usefulness in various online communities. Additionally, the rephrasing technique used, though effective in reducing offensive language while maintaining communication relevance, could unintentionally change the original meaning or emotional tone of the text, potentially causing misunderstandings or dissatisfaction among users. The model's performance may be influenced by the biases inherent in the training data, possibly resulting in inaccuracies or discrepancies in detecting abusive language among different demographic groups. These limitations emphasize the importance of continuous research and collaboration among stakeholders to consistently enhance and upgrade the model's capabilities, ensuring its efficacy in promoting a safer and more inclusive online space.

The future development of rewriting models presents opportunities for enhancing capabilities and effectiveness through the integration of advanced text generation methods and semantic understanding techniques. A focus on improving accuracy and contextual relevance of rewriting suggestions is crucial. Understanding user interaction with the rewriting process, including acceptance of suggestions while preserving original emotions and intentions in hate speech, is imperative. In addition to explore mechanisms to reduce misclassifications of normal speech as hate speech will be vital, as well as techniques such as ongoing model retraining and user feedback loops can enhance accuracy. The model should adapt to evolving expressions of hate speech over time through methods like continual learning and domain adaptation. These efforts will ensure the model's effectiveness in fostering a safer online environment. Progress in artificial intelligence and natural language processing could involve advanced large-scale language models to improve detection of abusive language. Integrating multimodal data and sentiment analysis may enhance the model's ability to comprehend content and emotions conveyed. Efficient processing of large-scale data using techniques like distributed computing and incremental updates is essential. Cross-lingual detection of abusive language is critical in a globally connected social media landscape, demanding models with strong generalization abilities to overcome language and cultural barriers. Collaboration among various stakeholders such as governments, educational institutions, social media platforms, and user communities is essential for fostering a safer, inclusive online environment and promoting healthy discourse and societal development through joint efforts in researching abusive and hate speech.

Data availability

Data in this article is detailed in the appendix. Original datasets and supplementary materials are available by email request.

Funding

The research received funding from the National Science and Technology Council, Taiwan under Grant Agreement No NSTC112-2221-E-027-059-.

CRediT authorship contribution statement

August F.Y. Chao: Writing – review & editing, Project administration, Methodology, Conceptualization. **Chen-Shu Wang:** Supervision, Funding acquisition, Conceptualization. **Bo-Yi Li:** Writing – original draft, Validation, Software, Resources, Methodology,

Conceptualization. **Hong-Yan Chen:** Visualization, Validation, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Chen-Shu Wang reports financial support was provided by National Science and Technology Council, R.O.C (Taiwan). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the National Science and Technology Council, Taiwan, for financially supporting this research under Contract No. NSTC112-2221-E-027-059-.

Appendix

In the research conducted by Yang and Lin (2020), they employed detection and rephrasing principles derived from the work of Su et al. (2017) to detect profanity expressions within statements. Nevertheless, certain profanity patterns, such as "歸覽趴會" (interpreted as 'fire full of my scrotum; pissed off'), pose considerable difficulty in discernment when examining opinions in pandemic context. Furthermore, certain patterns lack the presence of extreme sentiments within their contextual framework, as exemplified by "睡懶覺". In this study, specific groups were selectively employed for the acquisition of datasets, as detailed in Table 1.

Table 1
Selected Chinese Profanity Lexicon in TOCP (NTOU Chinese Profanity) (Yang and Lin, 2020)

Group Number	Profanity Lexicon	Description/Translation
Group1	幹您 幹拎 幹恁 幹你 幹妳 幹林 幹哩 幹淋 幹他 幹她 操您 操拎 操恁 操你 操妳 操林 操哩 操淋 操他 操她 肏您 肏拎 肏恁 肏你 肏妳 肏林 肏哩 肏淋 肏他 肏她 賽您 賽拎 賽恁 賽你 賽妳 賽林 賽哩 賽淋 賽他 賽她	Contains a variety of offensive and derogatory terms used in explicit or disrespectful contexts.
Group2	幹 操 肏 賽	Consists of explicit and vulgar verbs often used in a derogatory or offensive manner.
Group3	賤婊 婊子 破麻 賤婊子 淫蕩 淫娃 賤貨 賤女人	Derogatory terms targeting women, implying disrespect, promiscuity, or low moral standards.
Group4	賤人 賤	
Group5	機掰 雞掰 機八 雞八 機歪 雞歪 機機歪歪 雞雞歪歪	
Group6	牛逼 傻逼 臭逼	
Group7	尻	
Group8	三小	Derogatory terms targeting man, implying disrespect, annoying or low moral standards.
Group9	豪洨 唬洨 虎洨 毫洨	
Group10	甲洨 假洨 呷洨	
Group11	三洨 撒洨 殺洨 啥洨 沙洨	
Group12	魯洨 盧洨 嚕洨	
Group13	洨	
Group14	覽趴 懶趴 攬趴	
Group15	懶覺 覽覺 攬覺 懶較 覽較 攬較 懶叫 覽叫 攬叫 懶鳥 覽鳥 攬鳥	
Group16	屌	

Upon embracing the political lexicon of hate speech, it becomes evident that certain terms remain absent within the framework of discussions concerning the pandemic, lacking pertinent contextual relevance. Additionally, there exists a subset of terms, such as "菜包賜死," for which discerning meaning proves challenging due to the absence of original elucidations within this political lexicon. To facilitate the aggregation of perspectives, a comprehensive categorization of the political hate speech lexicon was undertaken, resulting in the identification and segregation of its constituents into 21 distinct groups, intended for the purpose of opinion collection.

Table 2
Selected political hate speech lexicon from Wang et al., in 2022

Group	Terms	Description/Translation
1	共匪, 支那, 支那人, 支那狗, 支那豬, 滾回中國, 滾回大陸, 滾回中國吧, 滯台支那, 大陸低等人, 中國狗	Derogatory terms related to China
2	舔共, 舔中, 舔共仔, 舔共狗, 舔共韓狗, 舔中賤畜, 挺爛共匪, 死共匪, 共狗, 支那狗, 中共狗, 五毛狗, 五毛狗畜生, 舔共狗, 五毛狗, 五毛狗畜生, 中共的走狗, 支那賤種, 中共狗, 走狗, 走狗賣國賊	Insults directed at those perceived as pro-China
3	綠蛆, 綠畜, 綠狗, 綠霉, 綠毛雜種, 民進黨賤婦, 賤貨	Pejorative references for supporters of the DPP (Democratic Progressive Party)

(*continued on next page*)

Table 2 (*continued*)

Group	Terms	Description/Translation
4	綠蛆網軍, 綠肥貓, 綠畜打手們, 綠色畜生, 垃圾綠蛆, 垃圾綠蛆沒品, 綠色恐怖再現	Negative terms targeting online supporters of the DPP
5	綠蛆議員, 綠蛆媒體, 綠蛆集團, 冥視, 無恥綠媒	Derogatory references related to DPP politicians, media, or groups
6	藍蛆, 藍畜生, 狗民黨	Derogatory terms referring to supporters of the KMT (Kuomintang)
7	滾回香港, 港畜, 滾回去香港	Insults directed at individuals associated with Hong Kong or perceived as pro-Hong Kong
8	舔日, 日本走狗	Derogatory terms aimed at those seen as pro-Japanese
9	越南雜種	Derogatory term referencing people from Vietnam
10	黑韓, 喜韓狗, 舔共韓, 韓國愚, 韓狗屎, 草包, 韓草包, 喜韓狗, 韓狗屎	Insults targeting individuals supportive
11	台獨狗, 台獨狗畜牲	Derogatory terms for supporters of Taiwan independence
12	冥盡黨, 冥進黨, 冥禁黨, 冥燼黨	Derogatory references to the DPP in various forms
13	下架冥進黨	Calls for the removal or disbandment of the DPP
15	五毛, 舔共的, 走狗, 豢養, 漢奸, 共諜, 匪諜	Pejorative references for individuals seen as pro-China propagandists or government supporters
16	賣國賊, 政治淫婦, 邪惡菜妖	Insults directed at perceived traitors or morally corrupt politicians
17	菜渣, 菜陰魂, 菜陰文, 其邁賤畜生, 錢菊	Derogatory terms aimed at politician deemed contemptible or of low quality
18	小人, 豢養, 畜生, 賤畜, 賤貨, 賤種, 雜種, 垃圾沒品, 禍害, 無腦, 低等人, 垃圾, 狗黨, 淫婦, 政治淫婦, 菜渣	Various insults implying inferiority or moral deficiency
19	畜生, 畜生黨, 狗黨, 綠畜打手, 狗畜, 狗畜牲, 台獨狗畜牲, 蛆蟲, 狗賊, 雜種, 雜種畜生	Derogatory terms targeting political groups with negative connotations
21	滅共, 滾出, 滾出台灣, 快滾,	Calls for the removal or elimination of pro-China elements or individuals

References

[1] H. He, S. Kim, A. Gustafsson, What can we learn from# StopHateForProfit boycott regarding corporate social irresponsibility and corporate social responsibility? J. Bus. Res. 131 (2021) 217–226.
[2] S. Subyantoro, S. Apriyanto, Impoliteness in Indonesian language hate speech on social media contained in the Instagram account, Journal of Advances in Linguistics 11 (2) (2020) 36–46.
[3] P. Alonso, R. Saini, G. Kovács, Hate speech detection using transformer ensembles on the hasoc dataset. Speech and Computer: 22nd International Conference, 2020, pp. 7–9.
[4] R. Langton, Speech acts and unspeakable acts. Philos Public Aff, 1993, pp. 293–330.
[5] M.J. Matsuda, Public response to racist speech: considering the victim's story, MICH. L. REV. 87 (1988) 2320–2322.
[6] K.F. Anderson, Diagnosing discrimination: stress from perceived racism and the mental and physical health effects, Socio. Inq. 83 (1) (2013 Feb) 55–81.
[7] I. Maitra, M.K. McGowan, Speech and Harm: Controversies over Free Speech, Oxford University Press, 2012.
[8] S. Sontag, H.H. Broun, Illness as Metaphor: Farrar. *Straus And Giroux*, 1978, p. 87.
[9] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of abusive language. Proceedings of the International AAAI Conference on Web and Social Media, 2017, pp. 512–515.
[10] M. Mozafari, R. Farahbakhsh, N. Crespi, A BERT-based transfer learning approach for hate speech detection in online social media. International Conference on Complex Networks and Their Applications, 2019, pp. 928–940.
[11] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on twitter using a convolution-gru based deep neural network. The Semantic Web: 15th International Conference, 2018, pp. 745–760.
[12] B. Vidgen, L. Derczynski, Directions in abusive language training data, a systematic review: garbage in, garbage out, PLoS One 15 (12) (2020).
[13] P. Saha, M. Das, B. Mathew, A. Mukherjee, Hate speech: detection, mitigation and beyond. Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, 2023, pp. 1232–1235.
[14] H.P. Su, Z.J. Huang, H.T. Chang, C.J. Lin, Rephrasing profanity in Chinese text. Proceedings of the First Workshop on Abusive Language Online, 2017, pp. 18–24.
[15] C. Clune, E. McDaid, Content moderation on social media: constructing accountability in the digital space, Account Audit. Account. J. 37 (1) (2024) 257–279, https://doi.org/10.1108/AAAJ-11-2022-6119.
[16] A. Boutyline, R. Willer, The social structure of political echo chambers: variation in ideological homophily in online networks, Polit. Psychol. 38 (3) (2017 Jun) 551–569.
[17] T. Tiţa, A. Zubiaga, Cross-lingual hate speech detection using transformer models. https://arxiv.org/pdf/2111.00981.pdf, 2021. (Accessed 6 June 2024).
[18] N. Chetty, S. Alathur, Hate speech review in the context of online social networks, Aggress. Violent Behav. 40 (2018) 108–118.
[19] A. Weber, Manual on Hate Speech, Council of Europe, 2009.
[20] C. Yong, Does freedom of speech include hate speech? Res. Publica 17 (4) (2011) 385–403.
[21] L. Fan, H. Yu, Z. Yin, Stigmatization in social media: documenting and analyzing hate speech for COVID-19 on Twitter, Proceedings of the Association for Information Science and Technology 57 (1) (2020), https://doi.org/10.1002/pra2.313.
[22] N. Vishwamitra, R.R. Hu, F. Luo, L. Cheng, M. Costello, Y. Yang, On analyzing covid-19-related hate speech using bert attention. 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 669–676.
[23] J. Wang, X. Zhang, W. Liu, P. Li, Spatiotemporal pattern evolution and influencing factors of online public opinion——evidence from the early-stage of COVID-19 in China, Heliyon 9 (9) (2023) e20080.
[24] M. Haman, The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic, Heliyon 6 (11) (2020) e05540, 1-9.
[25] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: In Proceedings of the NAACL Student Research Workshop, 2016, pp. 88–93.
[26] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content. Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 145–153.

[27] M. Niemann, D.M. Riehle, J. Brunk, J. Becker, What is abusive language? Integrating different views on abusive language for machine learning, Multidisciplinary International Symposium on Disinformation in Open Online Media. (2019) 59–73.
[28] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Comput. Surv. 51 (4) (2018) 1–30.
[29] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, Proceedings of the 26th international conference on World Wide Web companion (2017) 759–760. Apr3.
[30] S.D. Swamy, A. Jamatia, B. Gambäck, Studying generalisability across abusive language detection datasets. Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), 2019, pp. 940–950.
[31] N. Nikhil, R. Pahwa, M.K. Nirala, R. Khilnani, Lstms with attention for aggression detection. Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 52–57.
[32] H. Liu, P. Burnap, W. Alorainy, M. Williams, Scmhl5 at TRAC-2 shared task on aggression identification: bert based ensemble learning approach, European Language Resources Association (ELRA) (2020) 62–68.
[33] C.C. Wang, M.Y. Day, C.L. Wu, Political hate speech detection and lexicon building: a study in taiwan, IEEE Access 10 (2022) 44337–44346.
[34] A. Chang, W. Jiao, Predicting health communication patterns in follower-influencer networks: the case of Taiwan amid COVID-19, Asian Journal for Public Opinion Research 8 (3) (2020) 246–264.
[35] H. Yang, C.J. Lin, Tocp: a dataset for Chinese profanity processing. Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 6–12.
[36] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Network. 18 (5–6) (2005) 602–610.
[37] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022, 2023.
[38] A. Fernández, V. López, M. Galar, M.J. Del Jesus, F. Herrera, Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches, Knowl. Base Syst. 42 (2013) 97–110.
[39] D.R. Faisal, R. Mahendra, Two-stage classifier for COVID-19 misinformation detection using BERT: a study on Indonesian tweets. https://arxiv.org/pdf/2206.15359.pdf, 2022. (Accessed 6 June 2024).
[40] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, M. Granitzer, I feel offended, don't be abusive! implicit/explicit messages in abusive and abusive language. Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 6193–6202.
[41] W. Tan, Q. Yao, J. Liu, Two-stage COVID19 classification using BERT features. European Conference on Computer Vision, 2022, pp. 517–525.
[42] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[43] C. Graves, Generative AI can help you tailor messaging to specific audiences, Harv. Bus. Rev. (2023). https://hbr.org/2023/02/generative-ai-can-help-you-tailor-messaging-to-specific-audiences. (Accessed 6 June 2024).
[44] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Chen Z. ChenY, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, J. Wen, A survey of large language models, 2023, https://arxiv.org/pdf/2303.18223.pdf, 2023. (Accessed 6 June 2024).
[45] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (5) (1971) 378–382.
[46] T.R. Nichols, P.M. Wisner, G. Cripe, L. Gulabchand, Putting the kappa statistic to use, Qual. Assur. J. 13 (3–4) (2010) 57–61.
[47] P. Baldi, P.J. Sadowski, Understanding dropout, Adv. Neural Inf. Process. Syst. 26 (2013) 2814–2822.
[48] S. Khan, M. Fazil, V.K. Sejwal, M.A. Alshara, R.M. Alotaibi, A. Kamal, A.R. Baig, BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection, Journal of King Saud University-Computer and Information Sciences 34 (7) (2022) 4335–4344.
[49] A. Bisht, A. Singh, H.S. Bhadauria, J. Virmani, Kriti, Detection of hate speech and offensive language in twitter data using lstm model. Recent Trends in Image and Signal Processing in Computer Vision, 2020, pp. 243–264.
[50] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.