IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 25, 2021, accepted February 16, 2022, date of publication March 21, 2022, date of current version April 29, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160712

Political Hate Speech Detection and Lexicon Building: A Study in Taiwan

CHIH-CHIEN WANG, MIN-YUH DAY, (Member, IEEE), AND CHUN-LIAN WU
Graduate Institute of Information Management, National Taipei University, New Taipei City 23741, Taiwan

Corresponding author: Chih-Chien Wang (wangson@mail.ntpu.edu.tw)

This work was supported in part by the Ministry of Science and Technology (MOST), Taiwan, under Grant 109-2410-H-305-029.

ABSTRACT There is the minimal restriction to users' speech in cyberspace. The Internet provides a space where people can freely present their speech, which puts a Utopian sense of freedom of speech into practice. However, the appearance of hate speech is a significant side effect of online freedom of speech. Some users use hate speech to attack others, making the attacked targets uncomfortable. The proliferation of hate speech poses severe challenges to cyber society. Users may hope that social media platforms and online communities promote anti-hate speech. However, hate speech detection is still a developing technology that requires system developers to create a method to detect unacceptable hate speech while maintaining the online freedom of speech environment. No excellence detection approach has yet been proposed, although some literature has focused on it. The current study proposes an approach to build a political hate speech lexicon and train artificial intelligence classifiers to detect hate speech. Our academic and practical contributions include the collection of a Chinese hate speech dataset, creating a Chinese hate speech lexicon, and developing both a deep learning-based and a lexicon-based approach to detect Chinese hate speech. Although we focus on Chinese hate speech detection, our proposed hate speech detection system and hate speech lexicon development approach can also be used for other languages.

INDEX TERMS BERT, bidirectional encoder representations from transformers, deep learning, hate speech, lexicon, N-gram, natural language processing, TF-IDF.

I. INTRODUCTION

Using the Internet, people can easily exchange their points of view with others to facilitate effective communication. However, Internet and social media platform also allow aggressive users to spread hate speech to people with different opinions, especially when related to a political topic. Some users tend to post harsh words on social media to those who disagree with them and include hate speech when expressing negative opinions [1]. In addition to using rude language, users may also issue hate speech based on personal characteristics and attributes of an ethnic group or country, such as ''go back to your home country'' or ''people from that country are rapists.'' People are more likely to speak without restrictions on an online platform due to anonymity; therefore, hate speech appears more often in the cyber world than in the real world.

In Europe, many occurrences of hate speech are closely related to refugees. Social media sites are aware of the seriousness of the problem and have begun to address it. For example, a social media platform may advocate that if a message is reported as not conforming to the principles of platform use, it will be deleted within 24 hours [2]. The chief executive officer of Facebook has agreed to hand over the identification data of French users suspected of hate speech on the platform to judges on June 27th, 2019, and the deal is believed to be the first of its kind globally [3]. The seriousness of hate speech has involved the judicial level, and its influence has spread beyond previous perceptions.

Early identification of hate speech could prevent an escalation from speech to action [4]. Therefore, a method to prevent the spread of hate speech has become an important issue. The typical definition of hate speech, which may assist with its identification, refers to the speaker's tone, content, and targets [5]. However, there is often a contradiction between hate speech and free speech. Free speech is the symbol of a democratic system, which provides the citizens the right to hold their opinions and to challenge the opinions of others. Hate speech has a complicated connection with freedom of speech, making governance policies challenging to regulate [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Hualong Yu.

Previous literature has focused on hate speech to groups with particular attributes, such as immigrants, women [6], [7], religion [7], [8], and race [9]. However, with the popularity of online social media platforms, an increasing number of people are now aware of political issues. Followers of politicians can easily follow the whereabouts of politicians in real-time and understand new policies through online media platforms. However, people with polarized political standpoints may use social media to spread hate speech to criticize others with different political standpoints. Hate speech detection is essential to prevent the triggering of violence and prejudice, either from the offender or the victim of the action.

There are two typical challenges for a hate speech detection task: determining which type of speech is hate speech and detecting the hate speech automatically. Before filtering out hate speech, people must first decide which types of speech are categorized as hate speech. Most social media platforms have their own definitions of hate speech. For example, Facebook [10] defines hate speech as ''content targeting a person or group of people (including all subsets except those described as having carried out violent crimes or sexual offenses) on the basis of their aforementioned protected characteristic(s) or immigration status.'' However, Facebook allows content if it is ''in humorous or social commentary.'' YouTube [11] argued that it will ''remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, etc.'' However, when the primary purpose is educational, documentary, scientific, or artistic in nature, YouTube allows content that includes harassment. Basile, *et al.* [6] advocated that user ''may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.'' Additionally, Twitter does ''not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.''

The definitions by Facebook [10], YouTube [11], and Basile, *et al.* [6] argued that hate speech consists of discriminatory content targeting a person or a group of people based on their attributes, such as age, race, ethnicity, national origin, caste, religious affiliation, disability, gender, sexual orientation and gender identity, serious disease, and expression. Hate speech may target people with specific expressions, such as political standpoint expressions. However, to the best of our knowledge, no previous studies considered malicious criticism to supporters and politicians of an opposing stance as hate speech. The principle of freedom of speech should protect political speech, discussion, and argument. Nevertheless, malicious criticism and attacks on a person, politician, or their supporters, based on their political standpoint, may destroy the harmony of cyberspace, making it a polarized space or an echo chamber. Thus, attacks of malicious language toward people, based on their political standpoint, should also be considered a type of hate speech.

One challenge of hate speech detection is determining which speech is discriminatory content that should be filtered out. The speech audiences' reactionary feeling to the sentiment of words, phrases, sentences, and speech is essential to determine if it belongs to hate speech. We require the collective consensus of users to judge if content belongs to hate speech or not. The current human annotation of hate speech requires manual review, which limits the quantity by how much a human annotator can review and introduces subjective notions of what is considered hate speech [4].

Another challenge is to filter out hate speech without mistakenly removing normal speech. Hate speech detection is a typical classification task; however, it is difficult to determine a simple classification rule. Some words may have a discriminatory meaning, which should be prohibited. However, even when no discriminatory words are used, the sentiment of the speech may still be malicious or discriminatory. For example, there is no ''prohibited word'' in the sentence ''All people from that country are bad guys and they should go back to their country.'' However, the sentence may include some hate sentiment. To maintain cyberspace as a harmonious and friendly environment, we need to eliminate speech with discriminatory sentiment instead of removing speech that includes specific words.

Some literature has focused on the challenge of hate speech detection; however, none have achieved overwhelming results. The current study aims to develop an approach to detect political hate speech and develop a hate speech lexicon. We propose a framework to collect users' comments on political news, annotate political hate speech, build a hate speech lexicon, and develop a detection model to filter hate speech.

The current study presents three main contributions: First, we collected a hate speech corpus, which can be used for hate speech detection research. Second, we built a hate speech lexicon based on our annotated corpus. Third, we compared the detection performance of a deep learning-based method and a lexicon-based method for hate speech detection. To the best of our knowledge, few, if any, previous studies focused on hate speech relative to political standpoints. Few previous studies, if any, focused on political hate speech detection in the Traditional Chinese language. The research outcomes include a hate speech corpus and lexicon, which may be used for practical proposes. Although we focus on Chinese hate speech detection in Taiwan, our proposed framework of hate speech corpus collection, hate speech lexicon development, and hate speech detection model training can also be used in other languages and countries.

We organized the remainder of this paper as follows: in the next section, we review the related works on hate speech detection; the dataset and methodology are explained in section 3; our experimental results and recommendations are discussed in section 4; and finally, section 5 presents the conclusion and description of future work.

II. RELATED WORK

Owing to the popularity of social media, researchers have recently noticed the problem of hate speech detection. Table 1 presented the previous literature focused on hate speech detection and lexicon building.

TABLE 1. Summary of related literature on hate speech detection.

Literature	Data Source	Language	Detection Method	Best Performance
Alfina, et al. [8]	Twitter	Indonesian	RFDT, BLR, SVM, NB	93.5% (F1-measure)
Basile, et al. [6]	Twitter	English and Spanish	SVM, LSVM, logistic regression, CNN, LSTM, Bi-GRU, and BERT.	73% (F1-score)
Burnap and Williams [9]	Twitter	English	BLR, RFDT, SVM	89% (Precision)
ElSherief, et al. [12]	Twitter	English	Lexicon Building	-
Gambäck and Sikdar [13]	Twitter	English	CNN N-Gram word2vec	78.29% (F1-score)
Gitari, et al. [14]	web forums and blogs	English	Lexicon-based Theme-based	70.83% (F1-measure)
Malmasi and Zampieri [15]	Twitter	English	LSVM	78% (accuracy)
Ousidhoum , et al. [7]	Twitter	English, French, and Arabic	LR, Bi-LSTM	86% (Macro-F1)
Warner and Hirschberg [16]	Yahoo and website	English	SVM	63.75% (F1-score)
Waseem and Hovy [4]	Twitter	English	N-Gram logistic regression	

A. PREVIOUS LITERATURE ON HATE SPEECH DETECTION

Warner and Hirschberg [16] collected hate speech (anti-Semitic speech) from Yahoo! groups that readers had flagged as offensive and subsequently purged by administrators, and from the American Jewish Congress that originally collected to classify websites that advertisers may find unsuitable. They used parts-of-speech as features and used a support vector machine (SVM) to detect hate speech. Their model achieved an accuracy, precision, and recall of 94%, 68%, and 60%, respectively, for an F1 measure of 63.75%. The baseline accuracy was 91% because 91% of the collected speech was not anti-Semitic.

Gitari, *et al.* [14] created a classifier that can detect hate speech in web forums and blogs. They used subjectivity and semantic features related to hate speech to generate a lexicon, which was employed to build a classifier for hate speech detection. The study determined that text with semantic, hate and theme-based features achieved the best performance in 70.83% of the F-score.

Burnap and Williams [9] collected 1901 tweets, of which 11.68% were human-annotated as hate speech. The topics they detected were race, nationality, and religion. The study used Bayesian logistic regression (BLR), random forest decision trees (RFDTs), SVM, and an n-gram model to make predictions. They advocated that the classifier results were optimal using a combination of classifiers with a voted ensemble meta-classifier.

Waseem and Hovy [4] collected 136,052 tweets and performed a manual search of common slur terms and hashtags pertaining to religious, sexual, gender, and ethnic minorities. They hired one expert annotator and three amateur annotators to annotate 16,000 tweets (16% hate speech and 84% non-hate-speech). They adopted logistic regression (LR) and used n-gram to detect hate speech. However, they did not provide a detailed list of the slur terms and hashtags used. They also did not develop a hate speech lexicon.

Gambäck and Sikdar [13] used the hate speech dataset created by Waseem and Hovy [4] and adopted convolutional neural network (CNN) models to detect hate speech. They attempted to use different features of random vectors, character 4-grams, word vectors, and word vectors with character n-grams. Their results showed that the model based on Word2vec embeddings and a random vector performed best in the F1-score (78.29%) and precision (86.68%), respectively. However, the best recall performance of the models proposed by Gambäck and Sikdar [13] (72.14%) did not improve on that of the LR model (77.75%) proposed by Waseem and Hovy [4].

Malmasi and Zampieri [15] collected 14,509 English speech samples on Twitter and classified them into three categories: hate speech, offensive speech, and normal speech. They extracted features using character n-gram, word n-gram, and word skip grams, and determined that 4-gram feature extraction with a linear support vector machine (LSVM) achieved a maximum accuracy of 78%.

ElSherief, *et al.* [12] advocated that there are two types of targets for hate speech: a specific person (directed hate speech) and a group sharing a common protected characteristic (generalized hate speech). They identified that directed hate speech is more personal, directed, informal, and angrier, and often explicitly attacks the target (via name-calling) with fewer analytic words and more words suggesting authority and influence. Generalized hate speech is dominated by religious hate and is characterized by lethal words, such as murder, exterminate, and kill, and quantity words, such as million and many. In their study, they used multiple approaches to collect hate speech. In the critical phrase-based approach, they used Twitter's streaming application programming interface to obtain tweets, and they used the lexicon of hatebase.org, the world's largest online hate speech repository, as a lexical resource to search for hate speech. They obtained 28,318 hate speech occurrences using this approach. They also obtained 290 hate speech occurrences using a hashtag-based approach, which examined a set of 13 hashtags, such as #killallniggers, #internationalofffendafeministday, #getbackinkitchen, that are typically used

in the context of hate speech. They recruited annotators to identify whether or not the tweet contained hate speech, and whether the hate speech was directed towards a group of people (generalized hate speech) or an individual (directed hate speech). They used a word cloud to present the collected hate speech terms in their paper.

The International Workshop on Semantic Evaluation 2019 (SemEval 2019) included a task named HatEval to detect hate speech against immigrants and women [6]. A total of 74 teams participated in the task, using SVM, LSVM, logistic regression, CNN, long short-term memory (LSTM), bidirectional gated recurrent unit (Bi-GRU), and bidirectional encoder representations from transformers (BERT). The dataset comprised 13,000 English (39.76 hate speech) and 6,600 Spanish (41.93%) tweets. The task included two subtasks: detecting the presence of hate speech (subtask A), and distinguishing if the incitement is against an individual or a group (subtask B). The best performance of the F1 score for subtask A was 65.1% and 73%, and for subtask B, 57% and 70.5%, for English and Spanish, respectively.

Hate speech detection is a language-dependent issue. The detection model used for one language cannot be easily transferred to another language. Alfina, *et al.* [8] focused on detecting hate speech for the Indonesian language. They created a new dataset that encapsulated hate speech in general, including hatred for religion, race, ethnicity, and gender. Word n-unigram, word bigram, character trigram, and character 4-gram were used in their study. Their research results showed that the word n-gram feature outperformed the character n-gram. They also compared the performance of machine learning algorithms, including Naïve Bayes, SVM, BLR, and RFDT, for hate speech detection. They reported that the RFDT algorithm achieved the best performance, with an F1-score of 93.5% when using the word n-gram feature.

Most previous studies were typically oriented towards monolingual and single classification tasks. However, for multilingual social media platforms, it would be beneficial to translate one language to other languages and use one language's hate speech detection model for the others. Ousidhoum, *et al.* [7] presented a multilingual multi-aspect hate speech analysis dataset and used it to test the multilingual multitask learning approaches. They collect 5,647 English tweets, 4,014 French tweets, and 3,353 Arabic tweets. Multiple languages hate speech detection is helpful in a bilingual society. They compared both traditional baselines, using bag-of-words (BOW) as features on LR, and the deep learning-based method of bidirectional LSTM (BiLSTM) models with one hidden layer on each of the classification tasks. They revealed that deeper BiLSTM models performed poorly, owing to the size of the tweets, and identified that BiLSTM outperformed BOW-based models.

B. LEXICON-BASED AND SENTIMENT ANALYSIS

Lexicons, such as WordNet [17] and SentiWordNet[18], can assign negative, neutral, and positive sentiments to all words. The use of lexicons is an essential and vital approach in the natural language process (NLP) to determine the sentiment of speech.

The lexicon-based method is intuitive because a term should not appear in the public space of the cyber world if people feel that a term is uncomfortable. For example, ''nigga'' is a hateful term that should not appear in any normal speech, except for particular scenarios, such as movies or television drama shows. From this viewpoint, a lexicon is necessary for hate speech detection.

Although there are some sentimental dictionaries and lexicons, there is no comprehensive Chinese hate speech lexicon for Taiwan to the best of our knowledge. Hatebase.org claims to provide a multi-language hate speech lexicon; however, the Chinese hate speech terms included in their lexicon are limited. After querying the lexicon of hatebase.org (queried in June 2020), we were only able to obtain one hate speech term, ''台 巴 子'' (redneck from Taiwan), for Taiwan. When the scope was extended to all Chinese-speaking countries (in addition to Taiwan, Chinese is also used in China, Hong Kong, Macau, Singapore, and Malaysia), we only obtained 38 hate speech terms. Among these 38 hate speech terms, only three terms, ''臭婊子'' (Stinky bitch), ''土包子'' (redneck), and ''台巴子'' (redneck from Taiwan), are frequently used in Taiwan.

Some speech may be regarded as hateful, even though no single word contained in the speech is hateful on its own [16]. Lexicon-based methods have an innate weakness—they cannot filter out hate speech without pre-defined hate terms. Thus, previous studies also used other NLP tools, such as n-gram, term frequency-inverse document frequency (TF-IDF), and part-of-speech, as the text feature for hate speech detection [19]. N-gram is the most widely used tool in previous hate speech studies [9], [13], [15], [16]. It is simply a sequence of n words and assists in deciding which n-grams can be grouped to form single entities. N-gram is useful because online users may develop new terms or phrases to attack others; there are always new buzzwords appearing in cyberspace.

C. STATISTICAL ANALYSIS, MACHINE LEARNING, AND DEEP LEARNING

Lexicon-based detection methods tend to have lower precision than previous studies using machine learning or deep learning because they classify the text containing specific terms as hate speech [14].

A lexicon-based approach typically uses a simple yes-or-no classification or statistical analysis to calculate the probability of a speech sample being hate speech. For example, LR [4] and BLR [8], [9] are frequently used statistical techniques for hate speech detection.

However, machine learning models and deep learning models outperform lexicon-based statistical analysis approaches [7], [9], [16]. The SVM classifier is a classic machine learning model, which can be used for hate speech

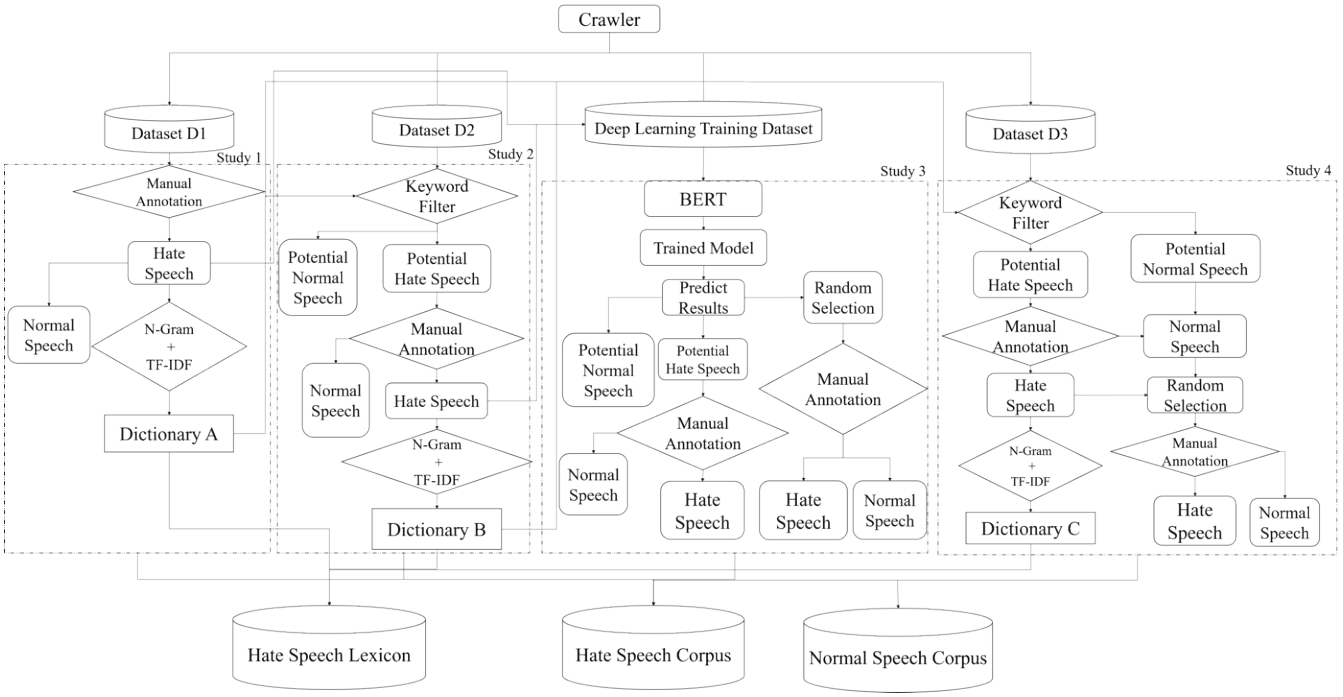


FIGURE 1. Experiment flow chart for political hate speech detection and lexicon building.

classification [6], [8], [9], [15], [16]. Previous studies also adopted RFDTs [8, 9] to detect hate speech.

Traditional machine learning cannot process large-scale data training with more complex detection; therefore, deep learning becomes better for training the model with big data. Previous studies also used neural networks to predict stock price trends using financial news. Gambäck and Sikdar [13] adopted CNN models to detect hate speech, whereas Ousidhoum, *et al.* [7] adopted bidirectional LSTM (biLSTM) models for the same purpose. CNN, LSTM, Bi-GRU, and BERT had been used in the SemEval 2019 HatEval task to detect hate speech [6]. BERT is a pre-trained language model based on the transformer model framework proposed by Google [20], which can also be used in hate speech detection.

Based on the above discussion, we identified several research gaps that can be met in the current study. Firstly, only a few Chinese hate speech studies have been conducted, even though Chinese hate speech is common in cyberspace. Previous studies have paid little attention to building a Chinese hate speech lexicon, even though lexicon building is a fundamental task for hate speech research. Third, the BERT model, proposed by Google, provides a new alternative for the NLP task; few studies have used the BERT model to detect hate speech. The current study can contribute to the research gaps mentioned above.

III. METHODS

A lexicon is useful in hate speech detection. Malicious users often use misspellings and abbreviations to avoid filters and classifiers [12]. Thus, finding new hateful terms is necessary for the hate speech detection task. The current study used n-Gram and TF-IDF to extract the essential and high-frequency terms to develop a Chinese hate speech lexicon. After building the lexicon, we attempted to detect hate speech.

Deep learning models outperform lexicon-based models in hate speech detection; thus, the current study uses deep learning to detect hate speech. However, deep learning requires an extensive data set of hate speech to train the models.

In the current study, we developed an approach to increase the hate speech dataset and manually annotated the dataset to verify the detection performance. Figure 1 presents our research framework. We conducted four studies to collect the datasets of hate speech and normal speech, build the hate speech lexicon, and develop the deep learning model to detect hate speech. We explain the details of all four studies, including how to construct the datasets, develop the hate speech lexicon, and train the model.

A. STUDY 1: INITIAL LEXICON BUILDING

1) DATA-CRAWLING

In study 1, we used a web crawler to extract the online user comments to Taiwanese political news from LINE Today, a news aggregator that integrates news from various news media. Unlike other news aggregators, LINE Today is also a social media platform on which people can comment on the news. LINE is a freeware application for instant text, voice, and video messages on mobile phones, tablets, and personal computers. It is the most popular instant communications application in Taiwan. In the first study, we crawled 11,917 comments to political news.

TABLE 2. Example of hate speech and normal speech.

Label	Data
hate speech	綠渣畜生黨的雜種網軍們加油啦 Fighting for the bastard cyber army of the green jerk and beast party. (Note: In Taiwan, green means the Democratic Progressive Party and other allied political parties with similar political standpoints.)
normal speech	現在的媒體、雜誌和名嘴，很容易被收買，因為拿好處就辦事，一堆人的良心不見了 The current media, magazines, and celebrities can easily be bribed. They only help with doing something to take advantage of, and many people bury their own consciences.

2) ANNOTATION

We recruited three annotators to categorize these comments to political news as normal speech or hate speech. To assist the annotators with the categorization of comments, we developed an annotation assistance system, which allowed annotators to view the news headline, news reports, and users' comments.

The reliability of the annotations is essential for a hate speech detection system. In the study by Ross, *et al.* [21], they concluded that raters required more detailed instructions for the annotation. Thus, in the current study, we provided the annotators with definitions, guidelines, and examples of hate speech and normal speech. Previous literature revealed that hate speech differs from offensive speech [15]. We asked annotators to divide speech as hate speech, offensive speech, and normal speech, as Malmasi and Zampieri [15] did in their study. Each comment to political news was annotated manually into the following three categories:

(1) Hate Speech: A sentence with an abusive intention on specific attributes of a group or individual, such as political beliefs, party membership, race, gender, age, sexual orientation, or gender identity, but not including satire or humorous comments.
(2) Offensive Speech: A sentence with irrational expression or the creation of opposing comments.
(3) Normal Speech: A sentence with neutral, positive, constructive, and non-offensive expression.

In this study, we only consider hate speech. If a speech sample was considered hate speech by at least two annotators, it was considered hate speech in the study. After annotation, we obtained 1,069 (8.93%) hate speech, and the other 10,848 (91.07%) comments were considered normal speech. Table 2 illustrates hate speech and normal speech examples, categorized by annotators.

To assess inter-annotator agreement, we adopt the Fleiss' kappa statistic [22], which provides an overall agreement measure of more than two annotators for a categorical rating (unlike Cohen's Kappa [23], which only provides a measure of pairwise agreement). The result provided a Fleiss' kappa of 0.267, a fair agreement.

3) N-GRAM TO FIND HATRED TERM

After the annotation process, we cleaned up the data by removing unnecessary symbols (>, ～, etc.) and emojis (😂, 🐶, etc.). These symbols and emojis are not considered hate speech, and we cannot filter out speech due to their use. Thus, we did not include them in the hate speech lexicon.

N-gram and TF-IDF were used to segment the speech samples labeled as hate speech. We used unigram (1-gram) to 6-gram for Chinese word segmentation to find the potential terms. Terms with a higher frequency (appeared at least three times) were checked by three annotators. If a term were categorized as hate speech by at least two annotators, the term would be considered as a hate speech term. This study obtained 113 terms and included them in lexicon A. The Fleiss' kappa statistic [22] was 0.793, which was a significant agreement.

4) EXAMPLE OF HATE SPEECH TERMS

According to the study by ElSherief, *et al.* [12], the keywords of hate speech in English are more similar to swear words or discriminatory words that the public uses to describe people with specific attributes, such as "queers," "Jihadi," or "bitches". However, we determined that the semantics of words in the Chinese language not only include the typical negative or positive opinion but also use a rich metaphor or homophone features as a description. In the following, we explain the features of the terms in the lexicon:

(1) Negative Polarity: Hate speech should involve a negative semantic orientation. The feature of extracted words in the lexicon are matched weakly or strongly to a negative meaning, such as "雜種" (Bastard) or "一群走狗" (A group of lackeys).
(2) Target Attributes: Some hate terms assail the targets with specific attributes, especially political beliefs, party membership, race, and gender. For example, the term "含糞" (mouth with shit) uses a homophonic style to draw an analogy of supporters of Kuo-yu Han (a Taiwanese politician) to shit (feces). Furthermore, the term "綠蛆們" abuses the supporters of the Democratic Progressive Party as maggots. We determined that it is more probable for these Chinese hate speech words to use metaphors to attack the targets with attributes.

B. STUDY 2: EXTENDING THE HATE SPEECH DATASET

The primary purpose of study 2 is to extend the hate speech dataset. In study 1, we only obtained 1,069 hate speech samples, which is not enough for any hate speech deep learning analysis. We also identified that the hate speech proportion was approximately 8.93%. If we hope to collect a dataset of 5,000 hate speech, we have to annotate approximately 55,000 speech samples, which is a resource-consuming task. We did not have sufficient resources to realize that; therefore, we developed an efficient approach to extend the hate speech dataset.

In study 2, we crawled 100,000 news comments from LINE Today. We used the 113 terms obtained by study 1 to filter the collected comments. Among the 100,000 news comments, 8,773 comments that included hate speech terms from lexicon A were considered potential hate speech, while the remaining 91,227 comments were considered normal speech.

We recruited three annotators to categorize these comments to political news as normal speech or hate speech. Only comments categorized as hate speech by at least two annotators were considered hate speech. The Fleiss' kappa statistic [22] was 0.986, which was an almost perfect agreement.

Among the 8,773 potential hate speech comments, only 3,427 comments were annotated manually as hate speech. The other 5,346 comments were considered normal speech, although they included some terms that were considered hate speech. Some of these comments included limited hateful sentiments; however, annotators did not think that the comments should be annotated as hate speech. Moreover, in our definition of hate speech, satire or humorous comments were not considered hate speech. Some satire or humorous comments included hate speech terms but should not be considered hate speech.

In study 1, we found that only 8.93% of political news comments should be considered hate speech. Most annotator resources are spent on normal speech. Using the lexicon approach to initial screen the speech, we can increase the hate speech proportion to 39.06%. The lexicon approach was useful to reduce the annotator resources for manually checking hate speech, although the lexicon approach cannot be directly used to detect hate speech.

Study 2 also adopted n-gram and TF-IDF to segment the labeled hate speech. We used unigram (1-gram) to 6-gram for Chinese word segmentation to find the potential hate speech terms. Three annotators checked terms with relatively higher TF-IDF results. If a term was categorized as hate speech by at least two annotators, the term would be considered a hate speech term. The Fleiss' kappa statistic [22] was 0.731, which was a significant agreement. We obtained 19 new hate speech terms after annotation, named Lexicon B, thus identifying 132 hate speech terms (lexicon A and lexicon B).

C. STUDY 3: DEEP LEARNING MODEL

In study 3, we used a deep learning model, based on BERT, to detect political hate speech in the Traditional Chinese language. BERT is a pre-trained language model based on the transformer model framework, a popular and state-of-the-art attention model for a wide variety of NLP tasks. The Google team trained the general-purpose ''language understanding'' model on a huge text corpus, including Wikipedia with 2,500 million words and a book corpus with 800 million words, in the 12-layer to 24-layer transformer; the model was then used for downstream NLP tasks. BERT shows that a bi-directionally trained language model can have a deeper sense of language context and flow than single-direction language models [20]. The model has two main features during the pre-training section:

(1) Masked language model: The model randomly masks 15% of the words in the sentence so that the model uses the context features to predict the masked words.
(2) Next sentence prediction: The model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document.

BERT can be used for a variety of language tasks while only adding a classification layer to the core model for fine-tuning training and can be used for classification tasks. Therefore, we used the BERT-based model to train the hate speech detection model.

1) DATA PREPARATION

We used the 4,496 hate speech comments obtained from studies 1 and 2 (1,069 from study 1 and 3,427 from study 2). We randomly selected 4,478 normal speech samples (collected from studies 1 and 2) to construct a dataset of 8974 comments, composed of a balanced number of normal speech and hate speech samples. We divide the dataset, using 80% for training data and 20% for testing data. The maximum word length of the news comments was set to 40; only the first 40 Chinese words were included for news comments THAT WERE longer than 40 words.

2) DETECTION PERFORMANCE EVALUATION

We used the precision, recall, accuracy, and F1-score, calculated by the confusion matrix, to evaluate the prediction model performance. The four performance evaluation indicators are described in Table 3.

TABLE 3. Confusion matrix for classification evaluation.

		Actual	
		Positive	Negative
Predict	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

TN: Represents the number of negative examples if the classification results are correct.
FP: Represents the number of negative examples if the classification results are incorrect.
FN: Represents the number of positive examples if the classification results are incorrect.
TP: Represents the number of positive samples and the results of positive classification.

Precision: The ratio of correctly predicted positive labels to the total predicted positive labels.

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|}$$

Recall: The ratio of correctly predicted positive labels to all labels in the actual class.

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|}$$

Accuracy: The ratio of correctly predicted labels to the total labels.

$$\text{Accuracy} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

F1-score: The weighted average of precision and recall.

$$\text{F1 Score} = \frac{2 * (Recall * Precision)}{Recall + Precision}$$

3) BERT FINE-TUNING

BERT includes two phases: pre-training and fine-tuning. The pre-training phase models are well trained by Google researchers, using the huge text corpus of Wikipedia. Google offers pre-trained models of different sizes: base, large, Xlarge, XXlarge. We adopted the base model in this study to reduce computing time.

We implemented fine-tuning using Python, running on Google Colab Notebooks, a Jupyter notebook environment that requires no setup and runs entirely in the cloud. We used Pytorch as our training framework, an open-source machine learning library developed by Facebook's AI Research laboratory. We used Adam as our optimizer and implemented Compute Unified Device Architecture in the training section to enhance training efficiency.

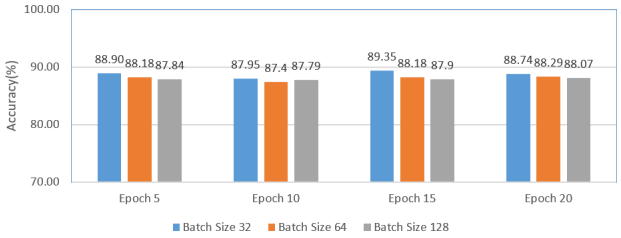


FIGURE 2. Performance comparison of different epochs and batch sizes for Chinese hate speech detection.

TABLE 4. Experimental results for bert model for chinese hate speech detection.

Model ID	Batch Size	Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
M01	32	5	88.90	83.83	96.43	**89.69**
M02	32	10	87.95	**88.48**	87.30	87.89
M03	32	15	**89.35**	87.24	92.20	89.65
M04	32	20	88.74	82.83	**97.77**	89.68
M05	64	10	87.40	80.99	**97.77**	88.59

In the epoch and batch size test, we used epochs of 5, 10, 15, and 20 and batch sizes of 32, 64, and 128 to construct different combinations. Figure 2 presents the performance comparison of different epochs and batch sizes for Chinese hate speech detection. Table 4 presents the experimental results using the BERT model for Chinese hate speech detection. The best classification results from the four types of evaluation using the test set are 97.77%. Using a batch size of 32 with an epoch of 20 and a batch size of 64 with an epoch of 10 can achieve the same score in recall evaluation.

D. STUDY 4: EVALUATION OF DETECTION PERFORMANCE OF BERT MODEL

In study 4, we verified the detection performance of the BERT model fine-tuning from study 3. We used a web crawler to extract another 100,939 online users' comments to Taiwanese political news from LINE Today. We test the detection efficiency using the developed BERT model.

1) BERT MODEL DETECTION

We used the fine-tuning BERT model in study 3 to tag the collected comments. Of the 100,939 comments, the BERT model tagged 11,331 comments as potential hate speech.

We recruited three annotators to categorize these comments to political news as normal speech or hate speech. Among the 11,331 potential hate speech samples, 7,927 were rated as hate speech by at least two annotators and were therefore considered as hate speech in the study. The other 3,404 were considered as normal speech. The precision was 69.7% (7927/11331), and the inter-annotator agreement of the Fleiss' kappa statistic [22] was 0.508, which indicated a moderate agreement.

The other performance indicators, such as accuracy, recall, and F1-score, were not available because we did not have sufficient resources to hire annotators to categorize all 100,393 comments. To estimate the accuracy, recall, and F1-score of the deep learning model, we randomly sampled 1000 comments and recruited three annotators to categorize them. We obtained the estimated detection performance of the BERT model with a precision of 73.2%, recall of 54.7%, and F1-score of 62.6%. The inter-annotator agreement of the Fleiss' kappa statistic [22] was 0.784, which was a significant agreement.

2) BASELINE MODEL: LEXICON APPROACH

We used the lexicon approach as a baseline model to compare the detection performance of the BERT model and the baseline lexicon approach. We used the hate terms included in lexicons A and B to detect hate speech from the collected 100,939 news comments. As a result of the lexicon detection approach, we tagged 7,823 comments as potential hate speech.

We recruited three annotators to categorize these comments to political news as normal speech or hate speech. Among the 7,823 potential hate speech samples, 4,337 were rated as hate speech by at least two annotators and therefore considered hate speech in the study. The other 3,596 samples were considered normal speech. The precision was 55.4% (4,337/7,823). The Fleiss' kappa statistic [22] was 0.348, which was considered a fair agreement.

The other performance indicators, such as accuracy, recall, and F1-score, were unavailable because we did not have sufficient resources to hire annotators to categorize all 100,393 comments. As with the BERT deep learning model, we used 1,000 randomly sampled comments and recruited three annotators to categorize them to estimate the deep

learning model's accuracy, recall, and F1-score. We obtained the estimated detection performance of the lexicon model with an accuracy of 54.1%, recall of 60.6%, and F1-score of 57.1%. The Fleiss' kappa statistic [22] was 0.639, which was a significant agreement.

Table 5 presents the performance comparison between the BERT and lexicon models. The results show that BERT significantly improved the detection performance achieved by the lexicon approach.

TABLE 5. Detection performance of the bert model and lexicon approach.

Model	Precision	Accuracy	Recall	F1-Score
BERT Model	69.7%	73.2%	54.7%	62.6%
Lexicon Approach	55.4%	54.1%	60.6%	57.1%

3) HATE SPEECH DATA SET AND LEXICON

We also adopted a procedure similar to studies 1 and 2 to identify new hate speech terms. We recruited three annotators to categorize the new terms that were found. The Fleiss' kappa statistic [22] was 0.828, which represented an almost perfect agreement. After annotation, we obtained 21 new hate speech terms called Lexicon C. We thus obtained 153 hate speech terms in total (lexicons A, B, and C).

Hate speech detection continues to be a developing issue, and the dataset is fundamental for hate speech research. To the best of our knowledge, there is no publicly available dataset lexicon for Traditional Chinese hate speech. In the current study, we collect a lexicon that can be used for detection and a dataset that can be made available for future hate speech research. We obtained such a dataset in the four studies detailed above, as summarized in Table 6.

TABLE 6. Hate speech dataset.

	Total	Annotated as hate speech	Annotated as normal speech	Unannotated
Corpus 1	11,917	1,069	10,848	0
Corpus 2	100,000	3,427	5,346	91,227
Corpus 3A	100,939	7,927	3,404	89,608
Corpus 3B	100,939	4,337	3,486	93,116
Corpus 3A-1	1,000	66	934	0
Corpus 3B-1	1,000	150	850	0

Table 7 presents the hate speech lexicon developed in this article. A hate speech lexicon is useful from many perspectives. Firstly, it is essential to extend the hate speech dataset when no hate speech dataset is available. Because the hate speech ratio is low (in study 1, we obtained a ratio of only 8.9%), it is not possible to manually annotate all speech samples. A hate speech lexicon is an effective starting point to detect potential hate speech. The potential hate speech list can reduce the required effort for manual annotation.

Moreover, a hate speech lexicon can help Internet users realize which terms may be considered hate speech from other users' viewpoints. If users do not intend to irritate others, they should not use the terms included in the hate speech lexicon. Thus, the hate speech lexicon provides a blacklist of terms users should not use in cyberspace.

TABLE 7. Hate speech lexicon.

	Terms	URL
Lexicon A	113	https://docs.google.com/spreadsheets/d/137PuLywpXu3Rr3qXXmP45bAlXrB-HQn7zSEL6kVaAcg/edit?usp=sharing
Lexicon B	19	https://docs.google.com/spreadsheets/d/1EdQtDXeFnHD87vIpSNpoI0kGqDR3iF7_6PrdQKBiP1k/edit?usp=sharing
Lexicon C	21	https://docs.google.com/spreadsheets/d/1mW41Up4m9CkJTyR1LHSzBfaDTakLxlO5iAqtJH7Ml6s/edit?usp=sharing
Total	153	https://docs.google.com/spreadsheets/d/1X6SckZ5BEw11rWpeR80JPSTU_muaba9StehsmPqLykA/edit?usp=sharing

In addition, a hate speech lexicon may provide a basis for social media platforms to filter out hate speech. Although deep learning would be more powerful for hate speech detection, using a hate speech term filter may be easier to convince users of why a comment is tagged as hate speech. People can easily understand if a social media platform informs them that their comments are filtered because they use prohibited terms.

IV. CONCLUSION AND FUTURE RESEARCH SUGGESTIONS

Hate speech may be directed to a specific person or a group sharing a common characteristic [12]. The current study focused on politically-related hate speech used to maliciously attack people from a different political standpoint. We demonstrated an approach to collect a hate speech dataset, build a hate speech lexicon, and develop a hate speech detection model. Our study started our experiment by initially collecting a hate speech dataset, building a hate speech lexicon, and extending the hate speech dataset using the hate speech lexicon. Moreover, we used the extended hate speech dataset to conduct a hate speech detection model based on the BERT deep learning model. Finally, we compared the performance of the BERT model and the lexicon-based approach.

Based on the study results, we identified that political hate speech terms in Taiwan, using the Traditional Chinese language, are more inclined to use metaphors to abuse people with specific attributes, especially political beliefs, party membership, and race attributes.

Moreover, the results showed that the lexicon-based hate speech detection model yielded a precision of 55.4%, while the precision of the BERT model was 69.7%. However, the BERT model can obtain a better detection performance than the lexicon approach. Thus, the BERT deep learning model has the potential to detect hate speech.

This study only considers the lexicon and BERT deep learning approaches. However, many deep learning approaches and NLP approaches can be used to address the hate speech detection problem. Future studies can compare the detection performance of different deep learning approaches. Different NLP approaches, including

a sentimental analysis, can also detect hate speech. Future studies can also consider the revised ALBERT (''A Lite'' version of BERT) and RoBERTa (Robustly optimized BERT approach) models for a detection performance comparison.

The dataset is important for a hate speech detection task. We collected a dataset that can be used in a future study. However, a large dataset is essential for the classification task; therefore, future studies may extend the dataset of hate speech.

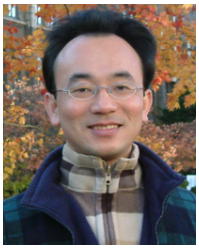
The current collected dataset focuses on comments to political news. Political hate speech is only one type of hate speech source. There exists a variety of hate speech types, such as hate speech focusing on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Future studies may use the approach developed by this study to collect hate speech datasets for different types of hate speech.

Moreover, a hate speech lexicon is useful for hate speech research. However, in the current study, the lexicon contained only 153 terms; thus, the size of the hate speech lexicon is still limited. Future studies may attempt to extend the hate speech lexicon. In the current study, we did not consider the degree of hate for the hate speech terms. We only classified the 153 terms like hate speech terms. Future studies can determine the degree of hate and divide the hate speech terms into several intensity levels.

REFERENCES

[1] N. A. Setyadi, M. Nasrun, and C. Setianingsih, ''Text analysis for hate speech detection using backpropagation neural network,'' in *Proc. Int. Conf. Control, Electron., Renew. Energy Commun. (ICCEREC)*, Dec. 2018, pp. 159–165.
[2] J. Titcomb, ''Facebook and Twitter promise to crack down on internet hate speech,'' The Telegraph. Accessed: Mar. 15, 2022. [Online]. Available: https://www.telegraph.co.U.K./technology/2016/05/31/facebook-and-twitter-promise-to-crack-down-on-internet-hate-spee/
[3] M. Rosemain. Exclusive: In a world first, Facebook to give data on hate speech suspects to French courts. Reuters. Accessed: Mar. 15, 2022. [Online]. Available: https://www.reuters.com/article/us-france-tech-exclusive/exclusive-facebook-to-give-data-on-hate-speech-suspects-to-french-courts-minister-idUSKCN1TQ1TJ
[4] Z. Waseem and D. Hovy, ''Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,'' in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93.
[5] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, ''Comparison of ARIMA and artificial neural networks models for stock price prediction,'' *J. Appl. Math.*, vol. 2014, pp. 1–7, Mar. 2014.
[6] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, ''SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter,'' in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 54–63.
[7] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, ''Multilingual and multi-aspect hate speech analysis,'' 2019, *arXiv:1908.11049*.
[8] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, ''Hate speech detection in the Indonesian language: A dataset and preliminary study,'' in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Oct. 2017, pp. 233–238.
[9] P. Burnap and M. L. Williams, ''Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making,'' *Policy Internet*, vol. 7, no. 2, pp. 223–242, 2015.
[10] Facebook. *Facebook Community Standards—Hate Speech*. Accessed: Mar. 15, 2022. [Online]. Available: https://www.facebook.com/communitystandards/hate_speech/
[11] YouTube. *Hate Speech Policy*. Accessed: Mar. 15, 2022. [Online]. Available: https://support.google.com/youtube/answer/2801939?hl=en
[12] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, ''Hate lingo: A target-based linguistic analysis of hate speech in social media,'' in *Proc. 12th Int. AAAI Conf. Web Soc. Media*, 2018, pp. 42–51.
[13] B. Gambäck and U. K. Sikdar, ''Using convolutional neural networks to classify hate-speech,'' in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 85–90.
[14] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, ''A lexicon-based approach for hate speech detection,'' *Int. J. Multimedia Ubiquitous Eng.*, vol. 10, no. 4, pp. 215–230, Apr. 2015.
[15] S. Malmasi and M. Zampieri, ''Challenges in discriminating profanity from hate speech,'' *J. Exp. Theor. Artif. Intell.*, vol. 30, no. 2, pp. 187–202, Mar. 2018.
[16] W. Warner and J. Hirschberg, ''Detecting hate speech on the world wide web,'' in *Proc. 2nd Workshop Lang. Soc. Media*, 2012, pp. 19–26.
[17] G. A. Miller, ''WordNet: A lexical database for English,'' *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[18] A. Esuli and F. Sebastiani, ''SentiWordNet: A publicly available lexical resource for opinion mining,'' in *Proc. 5th Int. Conf. Lang. Resour. Eval. (LREC)*, 2006, pp. 417–422.
[19] K. Dinakar, R. Reichart, and H. Lieberman, ''Modeling the detection of textual cyberbullying,'' in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 11–17.
[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' 2018, *arXiv:1810.04805*.
[21] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, ''Measuring the reliability of hate speech annotations: The case of the European refugee crisis,'' 2017, *arXiv:1701.08118*.
[22] J. L. Fleiss, ''Measuring nominal scale agreement among many raters,'' *Psychol. Bull.*, vol. 76, no. 5, p. 378, 1971.
[23] J. Cohen, ''A coefficient of agreement for nominal scales,'' *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

CHIH-CHIEN WANG is currently a Distinguished Professor with the Graduate Institute of Information Management, National Taipei University. He has published more than 100 academic articles. His current research interests include electronic commerce, online consumer behavior, internet marketing, and cyber psychology. He is also the Co-Editor-in-Chief of *International Journal of Electronic Commerce Studies* and the Editor-in-Chief of *Contemporary Management Research*.



MIN-YUH DAY (Member, IEEE) received the Ph.D. degree from the Department of Information Management, National Taiwan University. He is currently an Associate Professor with the Graduate Institute of Information Management, National Taipei University (NTPU), Taiwan. He has published more than 70 academic articles in international journals and a number of international conference proceedings. His current research interests include electronic commerce, financial technology, artificial intelligence, big data analytics, data mining and text mining, social media marketing, information systems evaluation, question answering systems, and biomedical informatics.

CHUN-LIAN WU received the master's degree from the Graduate Institute of Information Management, National Taipei University, Taiwan.