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

Chih-Chien Wang¹, Min-Yuh Day¹, and Chun-Lian Wu¹
¹Graduate Institute of Information Management, National Taipei University, New Taipei City 237, Taiwan.
Corresponding author: Chih-Chien Wang (e-mail: wangson@mail.ntpu.edu.tw).

ABSTRACT There is minimal restriction to users' speech in cyberspace. The Internet provides a space in which 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, which makes the attacked targets uncomfortable. The proliferation of hate speech poses serious 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 which requires system developers to create a method to detect unacceptable hate speech while maintaining the online freedom of speech environment. No effective 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, creation of a Chinese hate speech lexicon, and the development of 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 Hate Speech, Lexicon, Deep Learning, Bidirectional Encoder Representations from Transformers, BERT, N-gram, TF-IDF, Natural Language Processing

I. INTRODUCTION

Using the Internet, people can easily exchange their points of view with others; thus facilitating effective communication. However, this also allows aggressive users to spread hate speech to people who have 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, owing to the nature of 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 tone, content, and targets of the speech [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, which makes governance policies difficult to regulate[1].

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: to determine which type of speech is hate speech and to detect the hate speech automatically. Before filtering out hate speech, people have to 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. Twitter [12] 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 Twitter [12] 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. Political speech, discussion, and argument should absolutely be protected by the principle of freedom of speech. Nevertheless, malicious criticism and attacks on a person, politician, or their supporters, just 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 as a type of hate speech.

To address one of the challenges of hate speech detection and determine which speech is discriminatory content that should be filter 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 not only limits the quantity by how much a human annotator can review, but also introduces subjective notions of what is considered as 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 for it. 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 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 focus 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.

There 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 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 for other languages and in other countries.

The remainder of this paper is organized 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. Previous literature has focused on hate speech detection and lexicon building, as presented in Table I.
TABLE I
SUMMARY OF RELATED LITERATURE ON HATE SPEECH DETECTION

| Literature          | Source | Language       | Detection Features | Method          | Best Performance         |
|---------------------|--------|----------------|--------------------|-----------------|--------------------------|
| Alfina, et al. [8]  | Twitter| Indonesian     | N-Gram, SVM, NB    | RFDT, BLR, SVM  | 93.5% (F1-measure)       |
| Basile, et al. [6]  | Twitter| Spanish and English | N-Gram, SVM, LSVM, LSTM, BiGRU, and BERT | (F1-score)       | 73%                        |
| Basile, et al. [6]  | Twitter| English and Spanish | N-Gram, SVM, LSVM, logistic regression, CNN, LSTM, Bi-GRU, and BiBERT | (F1-score)       | 73%                        |
| Burnap and Williams [9] | Twitter| English         | N-Gram, SVM        | BLR, LSVM, SVM  | 89% (Precision)          |
| ElSherief, et al. [13] | Twitter| English         | N-Gram, word2vec | SVM             | -                        |
| Gambäck and Sikdar [14] | Twitter| English         | CNN                | LSVM, SVM, BLR, RFDT, and Bi-GRU | 78.29% (F1-score)       |
| Gitari, et al. [15] | web forums and blogs | English        | Lexicon-based     | SVM, LSVM, SVM  | 70.83% (F1-measure)     |
| Malmasi and Zampieri [16] | Twitter| English         | Feature           | LSVm             | 78% (accuracy)           |
| Ousidhoum, et al. [7] | Twitter| French, and Arabic | Character n-grams | LR, Bi-LSTM      | 86% (Macro-F1)           |
| Warner and Hirschberg [17] | Yahoo and website| English        | SVM                | SVM, LSVM, 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 [17] collected hate speech (anti-Semitic speech) from Yahoo! groups that had been flagged by readers as offensive, and subsequently purged by administrators, and from the American Jewish Congress, 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. [15] created a classifier which can be used to detect hate speech in web forums and blogs. They used subjectivity and semantic features related to hate speech to generated 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 results of the classifier 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 that were used. They also did not develop a hate speech lexicon.

Gambäck and Sikdar [14] used the hate speech dataset created by Waseem and Hovy [4] and adopted convolutional neural network (CNN) models to detect hate speech. They attempted the use of 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 [14] (72.14%) did not improve on that of the LR model (77.75%) proposed by Waseem and Hovy [4].

Malmasi and Zampieri [16] collected 14,509 English speech samples on Twitter and classified them in 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. [13] 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 the use of 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 key 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, #internationaloffendafeministday, #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 towards 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 composed of 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 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 the other languages and use the hate speech detection model of one language 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 a dataset of 5,647 English tweets, 4,014 French tweets, and 3,353 Arabic tweets. Multiple language hate speech detection is useful 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 [18] and SentiWordNet [19], can be used to assign negative, neutral, and positive sentiments to all words. The use of lexicons is an essential and important approach in the natural language process (NLP) to determine the sentiment of speech. A lexicon-based hate speech detection approach requires a lexicon to tag words with a semantic label [15, 20]. For example, Lingardi, et al. [20] identified 76 derogatory terms and used them to recognize hate 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 special scenarios, such as movies or television drama shows. From this viewpoint, a lexicon is important in hate speech detection.

Although there are some sentimental dictionaries and lexicons, to the best of our knowledge, there is no comprehensive Chinese hate speech lexicon for Taiwan. Hatebase.org claims to provide a multi-language hate speech lexicon; however, the Chinese hate speech terms included in their lexicon is 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 [17]. 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 [21]. N-gram is the most widely used tool in previous hate
speech studies [9, 14, 16, 17]. It is simply a sequence of n words and assists in deciding which n-grams can be grouped together to form single entities. N-gram is useful because online users may develop new terms or new 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 a lower precision, compared with the previous studies using machine learning or deep learning, because they classify the text containing specific terms as hate speech [15].

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, 17]. The SVM classifier is a classic machine learning model, which can be used for hate speech classification [6, 8, 9, 16, 17]. 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 a better choice for training the model with big data. Previous studies also used neural networks to predict stock price trends using financial news. Gambäck and Sikdar [14] adopted CNN models to detect hate speech, whereas Ousidhoum, et al. [7] adopted bidirectional LSTM (biLSTM) models for the same purpose. In the SemEval 2019 HatEval task, CNN, LSTM, Bi-GRU, and BERT were used to detect hate speech [6]. BERT is a pre-trained language model based on the transformer model framework proposed by Google [22], 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, provide 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 [13]. Thus, finding new hateful terms is necessary for the hate speech detection task. In the current study, we 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 use it 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 a large data set of hate speech to train the models.

---

Fig. 1. Experiment flow chart for political hate speech detection and lexicon building.
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, as follows.

**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, which is a news aggregator that integrates news from a variety of 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 politic news.

2) **Annotation**
We recruited three annotators to categorize these comments to politic 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. [23], they concluded that raters required more detailed instructions for the annotation. Thus, in the current study, we provided the annotators with the definition, guidelines, and examples of hate speech and normal speech. Previous literature revealed that hate speech is different from offensive speech [16]. We asked annotators to divide speech as hate speech, offensive speech, and normal speech, as Malmasi and Zampieri [16] 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 as hate speech by at least two annotators, it was considered as 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 as normal speech. Table II illustrates examples of hate speech and normal speech, categorized by annotators.

To assess inter-annotator agreement, we adopt the Fleiss’ kappa statistic [24], which provides an overall agreement measure of more than two annotators for a categorical rating (unlike Cohen’s Kappa [25], which only provides a measure of pairwise agreement). The result provided a Fleiss’ kappa of 0.267, which is a fair agreement.

| LABEL | DATA |
|-------|------|
| hate speech | 絲渣畜生黨的雜種網軍們加油啦 Fighting for the bastard cyber army of 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, and a bunch of people bury their own consciences. |

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 was categorized as hate speech by at least two annotators, the term would be considered as a hate speech term. In this study, we obtained 113 terms and included them in lexicon A. The inter-annotator agreement of the Fleiss’ kappa statistic [24] was 0.793, which was a significant agreement.

4) **Example of hate speech terms**
According to the study by ElSherief, et al. [13], 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.

**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. In study 1, we only obtained 1,069 hate speech samples, which is not enough for any hate speech deep learning analysis. 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 as potential hate speech, while the remained 91,227 comments were considered as normal speech. We recruited three annotators to categorize these comments to politic news as normal speech or hate speech. Only comments that were categorized as hate speech by at least two annotators were considered as hate speech. The inter-annotator agreement of the Fleiss’ kappa statistic [18] 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 as normal speech, although they included some terms that were considered as hate speech terms. 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 as hate speech. Some satire or humorous comments included hate speech terms, but should not be considered as hate speech.

In study 1, we found that only 8.93% political news comments should be considered as hate speech. Most annotator resources spent on normal speech. Using 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. Terms with relatively higher TF-IDF results were checked by three annotators. If a term was categorized as hate speech by at least two annotators, the term would be considered as hate speech term. The inter-annotator agreement of the Fleiss’ kappa statistic [18] was 0.731, which was a significant agreement. We obtained 19 new hate speech terms after annotation, named Lexicon B; thus, identifying a total of 132 hate speech terms (lexicon A and lexicon B).

**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 [22]. 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 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 III.

| TABLE III                              |
|----------------------------------------|
| **Confusion Matrix for Classification Evaluation** |
| Predict | Actual | Positive | Negative |
|---------|--------|----------|----------|
| Positive | True Positive (TP) | False Positive (FP) |
| Negative | False Negative (FN) | True Negative (TN) |

TN: Represents the number of negative examples if the results of the classification are correct.
FP: Represents a 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.

\[
\text{F1 Score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}
\]

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-turning using Python, run on Google Colab Notebooks, which is a Jupyter notebook environment that requires no setup and runs entirely in the cloud. We used Pytorch as our training framework, which is 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.

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 IV 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 is 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.

![Fig. 2. Performance comparison of different epochs and batch sizes for Chinese hate speech detection.](image)

| TABLE IV |
|-----------|
| **Experimental Results for BERT Model for Chinese Hate Speech Detection** |
| Model ID | Batch Size | Epochs | 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       |

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 [24] 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 [24] 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 results of 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 were therefore considered as hate speech in the study. The other 3,596 samples were considered as normal speech. The precision was 55.4% (4,337/7,823). The inter-annotator agreement of the Fleiss’ kappa statistic [18] was 0.348, which was considered as a fair 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. As with the BERT deep learning model, we used 1,000 randomly sampled comments and recruited three annotators to categorize them to estimate the accuracy, recall, and F1-score of the deep learning model. 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 inter-annotator agreement of the Fleiss’ kappa statistic [18] was 0.639, which was a significant agreement.

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

We also adopted a procedure similar to that in studies 1 and 2 to identify new hate speech terms. We recruited three annotators to categorize the new terms that were found. The inter-annotator agreement of the Fleiss’ kappa statistic [18] was 0.828, which represented an almost perfect agreement. We obtained 21 new hate speech terms after annotation, called Lexicon C. We thus obtained 153 hate speech terms in total (lexicons A, B, and C).

3) Hate Speech Data Set and Lexicon

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. In the four studies detailed above, we obtained such a dataset, as summarized in Table VI.

Table VII 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 there is no hate speech dataset 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 assist with reducing the required effort for manual annotation. Moreover, a hate speech lexicon can help Internet users realize which terms may be considered as hate speech from the viewpoint of other users. 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 that users should not use in cyberspace.

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, the use of 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 used terms that are prohibited.

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.

In this study, we only consider the lexicon and BERT deep learning approaches. However, there are many deep learning approaches and NLP approaches that can be used to address the hate speech detection problem. Future studies can compare the detection performance of different deep leaning approaches. Different NLP approaches, including a sentimental analysis, can also be used to 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, the future study may attempt to 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 as hate speech terms. Future studies can determine the degree of hate and divide the hate speech terms into several levels of intensity.

### IV. CONCLUSIONS AND FUTURE RESEARCH SUGGESTIONS

Hate speech may be directed to either a specific person or a group sharing a common characteristic [13]. In the current study, we focused on politically-related hate speech, used to maliciously attack people with 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. In our study, we 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 extend 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 results of the study, 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.

| TABLE VII |
| HATE SPEECH LEXICON |
| Terms | URL |
|---|---|
| Lexicon A | 113 | https://docs.google.com/spreadsheets/d/137PuLywpxx3rRzqXXmP45bAjXxTHQ7n7S5EL&vaAg/w?usp=sharing |
| Lexicon B | 19 | https://docs.google.com/spreadsheets/d/1EdQyDxeFhHD87vlp5Pnol0kGqgDR3iF7_6PbQisible1k/edit?usp=sharing |
| Lexicon C | 21 | https://docs.google.com/spreadsheets/d/1mW41Upl4n9CkJYtR1LHSzBfaDTakLx105iAqTH7Mf3s/edit?usp=sharing |
| Total | 153 | https://docs.google.com/spreadsheets/d/1X6sckZ5BEw11rWpe8807PSTU_11maba9StehsmPqLykA/edit?usp=sharing |

### REFERENCES

[1] N. A. Setyadi, M. Nasrun, and C. Setianingsih, "Text Analysis for Hate Speech Detection Using Backpropagation Neural Network," in 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), 2018: IEEE, pp. 159-165.

[2] J. Titcomb. "Facebook and Twitter promise to crack down on internet hate speech." [https://www.telegraph.co.uk/technology/2016/05/31/facebook-and-twitter-promise-to-crack-down-on-internet-hate-spe/](https://www.telegraph.co.uk/technology/2016/05/31/facebook-and-twitter-promise-to-crack-down-on-internet-hate-spe/)

[3] M. Rosemain. "Exclusive: In a world first, Facebook to give data
on hate speech suspects to French courts.

https://www.reuters.com/article/us-france-tech-exclusive/exclusive-facebook-to-give-data-on-hate-speech-suspects-to-french-courts-minister-idUSKCN1TQ1Y (accessed.

[4] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in Proceedings of the NAACL student research 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," Journal of Applied Mathematics, vol. 2014, 2014.

[6] V. Basile et al., "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 54-63.

[7] N. Oussidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and Multi-Aspect Hate Speech Analysis," arXiv preprint arXiv, no. 11049, 2019.

[8] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2017: IEEE, pp. 233-238.

[9] P. Burnap and M. 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, ed, 2019.

[11] YouTube, ed, 2019. [Online]. Available: https://support.google.com/youtube/answer/2801939?hl=en.

[12] Twitter, ed, 2019.

[13] 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 Twelfth International AAAI Conference on Web and Social Media, 2018.

[14] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in Proceedings of the first workshop on abusive language online, 2017, pp. 85-90.

[15] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 4, pp. 215-230, 2015.

[16] S. Malmasi and M. Zampieri, "Challenges in discriminating profanity from hate speech," Journal of Experimental and Theoretical Artificial Intelligence, vol. 30, no. 2, pp. 187-202, 2018.

[17] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the second workshop on language in social media, 2012: Association for Computational Linguistics, pp. 19-26.

[18] J. L. J. P. b. Fleiss, "Measuring nominal scale agreement among many raters," vol. 76, no. 5, p. 378, 1971.