International Journal of Information and Management Sciences
**31** (2020), 227-246. DOI:10.6186/IJIMS.202009_31(3).0002

International Journal of
**Information**
and
**Management**
**Sciences**

# Profanity and Hate Speech Detection

*Phoey Lee Teh and Chi-Bin Cheng*

Sunway University and Tamkang University

**Abstract**

Profanity, often found in today's online social media, has been used to detect online hate speech. The aims of this study were to investigate the profanity usage on Twitter by different groups of users, and to quantify the effectiveness of using profanity in detecting hate speech. Tweets from three English-speaking countries, Australia, Malaysia, and the United States, were collected for data analysis. Statistical hypothesis tests were performed to justify the difference of profanity usage among the three countries, and a probability estimation procedure was formulated based on Bayes theorem to quantify the effectiveness of profanity-based methods in hate speech detection. Three deep learning methods, long short-term memory (LSTM), bidirectional LSTM (BLSTM), and bidirectional encoder representations from transformers (BERT) are further used to evaluate the effect of profanity screening on building classification model. Our experimental results show that the effectiveness of using profanity in detecting hate speech is questionable. Nevertheless, the results also show that for Australia tweets, where profanity is more associated with hatred, profanity-based methods in hate speech detection could be effective and profanity screening can address the class imbalance issue in hate speech detection. This is evidenced by the performances of using deep learning methods on the profanity screened data of Australia data, which achieved a classification f1-score of 0.84.

*Keywords:* Profanity, hate speech, tweets, bayes theorem, deep learning.

## 1. Introduction

Profanity is a socially offensive language that has been in use vastly across countries. Also known as bad language, vulgar language, wrong choice of words, expletives, swear words, curse words, or foul language, profanity and its use reflect a behaviour that is offensive or lacking in respect to others.

The earliest studies on profanity in communication disorders had focused on the usage of profane words in conversational speech (see Cameron [6], Nerbonne and Hipskind [23]) The study by Cameron [6] examined what college students talk about in their normal conversations and found that 8.06% of the words used relate to sexual and excretory profanities. Nerbonne and Hipskind [23] performed a similar experiment on a different sample of participants and obtained different results. This difference in results implies

that profane words frequently used in conversation may vary across different groups of people.

In recent years, the popularity of online social media, such as Facebook, Twitter, Instagram, and YouTube, is a boost to communication and information-sharing among strangers. However, at the same time, online social media has become a hotbed for hate speech to breed. Profanity is considered to be closely, though not equivalently, related to hate speech (see Xiang et al. [41]). The term "hate speech" is defined as a form of attack with the intent to spread, incite, promote, or justify racial hatred towards a targeted category such as race, ethnicity, gender, sexuality, religion, etc. (see Feldman et al. [15]). Hate speech in the form of vulgar, offensive, insulting, and abusive languages often express hate and comprise profane words.

Hateful messages that are posted online, either intentionally or unintentionally, cause potential harm to victims (see Delgado and Stefancic [11] and Nemes [22]). Victims may develop psychological and pathophysiological symptoms similar to post-traumatic stress disorder (PTSD)—panic, fear, anxiety, nightmares, intrusive thoughts of intimidation, and denigration Jay [18]. Some countries have taken serious measures against hate speech. For example, Germany enforced an anti-hate speech law on social media companies in 2017. Social media companies that fail to remove 70% of hate speech found on their platforms within 24 hours could be fined up to USD 57 million (see Eddy and Scott [14]).

It is a challenge to identify and detect statements or messages that contain components of hatred. The tremendous amount of messages generated continuously every moment on social media makes it impossible to manually identify hate speech and thus, automatic detection of hate speech is ideal. However, using profanity as keyword for the automatic detection of hate speech is not fully feasible as sentences containing profanity are not always hate speeches. For example, "What the hell is wrong with this air conditioner," is more of an emotional expression than an intentional abuse of language despite containing the profane word "hell." Conversely, hate can also be conveyed through vague jokes which contain no profanity (see Parekh [25]). For example, "When they see your eyes you are going to be deported," is a sentence that intentionally makes fun of a person's ethnicity. Agrawal and Awekar [1] had also showed that profanity-based methods have both low precision and recall on hate speech detection based on datasets from FormSpring, Twitter and Wikipedia. The current study attempts to quantify the effectiveness of using profanity to detect hate speech. A quantitative measure based on the Bayes theorem is formulated.

Despite the limitation of using profanity as a means to detect hate speech, profanity could still serve as an initial filter to reduce the workload of hate speech detection. With the fact that profanity is neither necessary nor sufficient for hate speech, we consider that the presence of profanity may confuse a classification model in distinguishing hate from not-hate speech. Thus, this study suggests a hate speech detection method by building two classification models, where one model is trained by data with profanity screened from the original dataset while another model is trained by data without profanity.

## 2. Literature Review

The use of swearing in adolescents or youths has increased over the past 10 years (see Jay [18]), averaging approximately 80 to 90 swear words per day (see Deseret [12]). Much of this increase has been attributed to mass media such as music, film, and television (see Sapolsky and Kaye [30]). Arnett [2] commented that the media serve as an important socialising function to the young and impressionable audience, while Bushman and Cantor [5] reported that parents are concerned with their children adopting coarse language as a result of media exposure. The cultivation theory supports the notion that heavy exposure to media messages could shape one's view of reality (see Cressman et al. [8]).

Cressman et al. [8] examined the type, frequency, and usage of profanity in films between 1980 and 2006 that featured and targeted teenagers. Based on the regulations of the Federal Communications Commission (FCC) and previous research conducted by Kaye and Sapolsky [20], Burnap and Williams [4], Cressman et al. [8] categorised profanity into five groups: (1) the seven dirty words that are considered unspeakable for broadcast by the FCC, (2) sexual words that describe sexual body parts or sexual behaviour in coarse ways, (3) excretory words referring to human waste products and processes, (4) mild words that are offensive in nature but not included in the above categories, and (5) strong words considered more offensive than mild words that trigger strong emotions and reactions. Using content analysis, the study found no significant change in preference for type of profanity depicted in the films over the decades. Teen and adult characters in the films both use similar profanity types, but the former is more likely to use the seven dirty words than the latter. In terms of gender, male characters use more profanity than female characters. Although this analysis was performed on film characters rather than real-world persons, the results imply that a difference exists in profanity usage among different groups of people.

In terms of online social media, Thelwall [35] investigated the use of curse words on MySpace profiles and found gender and age to influence profanity usage. Sood et al. [33] studied profanity usage in Yahoo! Buzz communities and reported differences in the frequency of profanity usage among different communities. Bak et al. [3] studied self-disclosure behaviour on Twitter, while Wang et al. [39] analysed 51 million tweets (involving about 14 million Twitter users) to examine the characteristics of cursing activity on Twitter.

Wang et al. [39] aimed to answer a set of questions regarding the ubiquity, utility, and contextual dependency of cursing which have been recognised as crucial for understanding cursing in traditional offline communications. To create a lexicon of curse words for the study, Wang et al. [39] collected existing lists of curse words found on social media and extended them with curse words that had been used in previous studies. This lexicon-based method achieved a precision of 98.84%, a recall of 72.03%, and a F1 score of 83.33% for profanity detection. This method has high precision but lower recall due mainly to the variations in curse words (e.g., misspellings). According to the study's analysis, curse words occurred at a rate of 1.15% on Twitter with 7.73% of all tweets in the dataset

containing curse words. Male users were found to curse more often than women users, and both genders also used different types of curse words. High-ranked users (i.e., those with more followers) were also found to curse less than most low-ranked users. Wang et al. [39] also presented the top 20 most frequently used curse words found in their analysis, of which were employed in the current study.

Profanity has been used as the means or part of the means to identify hate speech on social media, based on the assumption that hateful messages usually contain specific negative words. Lexical resources are required to obtain such specific negative words. For example, Hatebase [17] is currently the largest online repository of structured, multilingual, and usage-based hate speech. The repository builds its lexical collection through crowdsourcing, comprising more than 1,000 hate-related words grouped into eight categories: archaic, class, disability, ethnicity, gender, nationality, religion, and sexual orientation. Each word in the database is given an offensivity score which ranges from 0 to 100, with 100 indicating most offensive hate-related word. Apart from that, Razavi et al. [28] manually compiled an Insulting and Abusing Language Dictionary, which contains words and phrases of different weights to represent the degree of their potential impact for hate speech detection.

Past studies that have employed lists of hate-related words to identify hate speech include Xiang et al. [41], Razavi et al. [28], Burnap and Williams [4], Wong et al. [40] and Nobata et al. [24]. Unlike previous studies that generally used hate-related words as features for supervised learning, Silva et al. [32] used the words to discover hate targets and identify hate speech in an unsupervised manner. Silva et al. [32] gathered data from Whisper and Twitter, and captured hate-related information through a sentence structure expressed as such,

$$\text{I} < \text{intensity} > < \text{user intent} > < \text{hate target} >,$$

where the component < user intent > is the verb that specifies the user's intent (e.g., hate), the component < intensity > is a qualifier to amplify the user's emotion in expressing his/her intent, and the component < hate target > is the person or group the intent is directed at.

To discover the hate target, Silva et al. [32] used two templates where one searches for terms containing the word "people" and the other employs hate-related (profane) words listed on Hatebase. The aforementioned sentence structure could only capture a portion of the hate speech that fits the structure. However, capturing hate speech was not the study's main purpose as its primary aim was to build a dataset to identify the targets of online hate speech.

As previously discussed, the use of profane words does not necessarily reflect hateful intent in a message and an actual hate speech may not contain any profanity at all (see Malmasi and Zampieri [21]). To discriminate general profanity from hate speech, Davidson et al. [10] applied supervised classification methods on a labelled dataset which distinguishes hate speech from offensive (but not hateful) language. Davidson et al. [10] used a lexicon that contains keywords of hate speech compiled by Hatebase to collect

tweets, and employed crowdsourcing to classify a sample of these tweets into three categories: (1) hate speech, (2) offensive language but not hate speech, and (3) neither of the above.

Davidson et al. [10] captured the syntactic structure of tweets to construct Penn Part-of-Speech (POS) tag unigrams, bigrams, and trigrams, as well as the quality of tweets with modified Flesch-Kincaid Grade Level and Flesch Reading Ease scores. Additionally, the study used a sentiment lexicon designed for social media to assign sentiment scores to each tweet, and also included binary and count indicators for hashtags, mentions, retweets, and Uniform Resource Locators (URLs) as well as features for the number of characters, words, and syllables in each tweet. In terms of data analysis, Davidson et al. [10] applied logistic regression, naive Bayes, decision trees, random forests, and linear support vector machines (SVMs) as classification models. The best performing model reported an overall precision of 0.91, a recall of 0.90, and a F1 score of 0.90. However, almost 40% of hate speech was misclassified and the precision and recall scores for this category were 0.44 and 0.61 respectively, which suggest that the classifier could not clearly distinguish between the first two categories.

Malmasi and Zampieri [21] used the dataset created by Davidson et al. [10] and employed a linear SVM to classify the data into the same three categories. Two groups of features were used in the classification: (1) character n-grams and word n-grams, and (2) word skip-grams. The resulting accuracy was 78% in identifying posts across the three categories. To further extend this study, Malmasi and Zampieri [21] applied different classification techniques, such as approaches based on single classifiers and more advanced ensemble classifiers, on the same dataset. The highest level of accuracy reported was 80% but similar to Davidson et al. [10], it was difficult to distinguish hate speech from general profanity. To detect cyberbullying on social media platform, Agrawal and Awekar [1] performed experiments using three real-world datasets: Formspring, Twitter, and Wikipedia, and employed deep learning methods, including convolutional neural network (CNN), LSTM, bidirectional LSTM (BLSTM), and BLSTM with attention, to build classification models. Recently, Salminen et al. [29] compared the performances by different models for online hate detection using multi-platform data. They collected 197,566 comments from four platforms: YouTube, Reddit, Wikipedia, and Twitter, and adopted several classification algorithms, including logistic regression, naive Bayes, support vector machines, XGBoost, and a simple feed-forward neural network. Their experiments showed the above models all outperformed the keyword-based baseline classifier.

## 3. Effectiveness of Profanity in Hate Speech Detection

This study formulated a probability estimation procedure based on the Bayes theorem to quantitatively measure the effectiveness of using profanity in hate speech detection. This quantitative measure is further illustrated with real data found on social media platform Twitter. This study particularly chose Twitter among all social media for its high popularity and retrievability. As of January 2020, users of Twitter had reached a total of 59.35 million users in the United States, and 45.75 and 16.7 million in Japan

and United Kingdom (see Clement [7]). The convenience of Twitter Archiver [36], which saves tweets that match the search keywords into a Google spreadsheet, also encouraged us to use tweets as our research subject.

This study aims to analyze tweets in English with different cultural backgrounds. Among the English-speaking countries, we divided them into three groups: 1) British, including United Kingdom, Australia and New Zealand; 2) North America, including the United States and Canada; and Asia, including Malaysia and Singapore. In each group we selected one country, where tweets with geographical locations in Australia, Malaysia, or the United States are selected for the amount of profane tweets in the rest countries are relatively scarce.

### 3.1. Bayesian probability estimation

There were two concerns that had to be considered when using profanity as a means to detect hate speech: (1) the probability that a tweet containing profane words is not a hate speech (false positive rate), and (2) the probability that a tweet not containing any profane word is a hate speech (false negative rate). As the sample data was retrieved using profane words as keywords, it was "incomplete" as only the false positive rate could be directly estimated from the collected tweets. Hence, the false negative rate was obtained using the Bayes theorem.

Let $f$ denote the event that a tweet contains profanity, and $h$ the event that a tweet is hateful. The first concern as discussed above was in fact the complement of the conditional probability $\text{prob}\,(h \mid f)$ (i.e., $1 - \text{prob}\,(h \mid f)$), defined as

$$\text{prob}\,(h \mid f) = \frac{\text{prob}\,(h \cap f)}{\text{prob}\,(f)}. \tag{3.1}$$

It was possible to compute $\text{prob}\,(h \cap f)$ from $\text{prob}\,(h \mid f)$ and $\text{prob}\,(f)$, where $\text{prob}\,(h \mid f)$ can be directly estimated from the sample data and $\text{prob}\,(f)$ obtained from literature (e.g., [31]).

The second concern was the conditional probability $\text{prob}\,(h \mid \neg f)$. Again, this can be derived by:

$$\text{prob}\,(h \mid \neg f) = \frac{\text{prob}\,(h \cap \neg f)}{\text{prob}\,(\neg f)} = \frac{\text{prob}\,(h) - \text{prob}\,(h \cap f)}{1 - \text{prob}\,(f)}, \tag{3.2}$$

where $\text{prob}\,(h)$ can be obtained from literature.

### 3.2. Data collection and data preprocessing

Tweets were retrieved using Twitter Archiver with keywords comprising the top 20 most frequently used curse words on Twitter identified by Wang et al. [39] and the profane terms grouped in categories by Teh et al. [34]. Wang et al. [39] analyzed about 51 million tweets and about 14 million users and found that seven most frequently used curse words accounted for more than 90% of all the cursing occurrences. Teh et al. [34] manually analyzed 500 posts from social media and use a corpus analysis tool, Wmatrix

Table 1: Profanity categories.

| Category | Description |
|---|---|
| Behaviour | Words that point to acts or conduct, especially towards others. |
| Disability | Words that attack a person's disability. |
| Ethnicity | Words that attack a person's social group in relation to national or cultural traditions. |
| Gender | Profane words that refer to gender or body parts. |
| Physical | Words that attack a person's physical appearance. |
| Race | Words that contain prejudice, discrimination, or antagonism directed at someone of a different race based on the belief that one's own race is superior. |
| Religion | Profane words related to religion. |
| Sexual orientation | Words that attack a person's sexual identity (e.g., gender, heterosexuality, homosexuality, bisexuality, etc.). |
| Social class | Words that discriminate or divide a society with respect to social or economic status. |
| Others | Profane words not classified in any of the above categories. |

[26, 27], to process the collected posts to extract keywords that are relevant to hate speech.

The top 20 most popular curse words as identified by Wang et al. [39] were fuck (covers 34.73% of all the curse word occurrences), shit (15.04%), ass (14.48%), bitch (10.34%), nigga (9.68%), hell (4.46%), whore (1.82%), dick (1.67%), piss (1.53%), and pussy (1.16%). The profane terms identified by Teh et al. [34] were categorised as Sexual Orientation (35.10%), Disability (20.14%), Gender (9.65%), Religion (4.76%), Race (7.82%), Behaviour (1.4%), Class (0.42%), and Others (15.57%).

By consulting the hate word categories of Hatebase and the categories suggested by Silva et al. [31], the collected profane words were manually assigned into 10 categories as illustrated in Table 1. However, as manual assignment was a time-consuming task, only top profane words found in most of the sample data were selected to be assigned a category. The profane keywords we used to retrieve tweets and their associated categories are presented in Table 2.

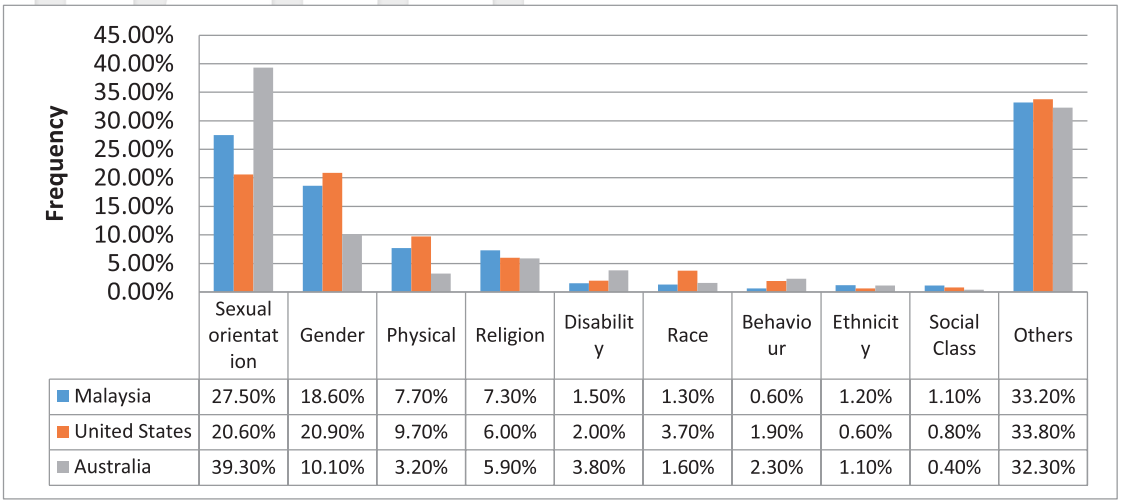
Between September 2017 and May 2018, 26250 tweets were collected from Twitter, where 17661 were from users in Australia, 4435 from the United States, and 4154 tweets from Malaysia. It is noted that the amount of retrieved tweets from Australia is much greater than that of the other two countries. This reflects that the Australian tend to use profanity in tweets more often than others.

Table 2: Keywords used to retrieve tweets.

| Category | Keywords |
|---|---|
| Behaviour | racist, racists, islamphobia, rapist, pissedr, pedos |
| Disability | retard, idiot, moron, dumbass, stupid, incompetent, delusional, douchebag, fucktard, dumbfuck, stupid trump, bigots, dumb asses, idoits |
| Ethnicity | chinese people, indian people, paki, chinese, malay |
| Gender | cunt, cunts, bitch, bitching, bitches, pussy, dick, dicks, cock, dogs, dog, bull, dickheads, dick face, misogynistic |
| Physical | asshole, assholes, ass, rape, raped, raping, suck, sucks, sucking, fuck up, fuck off, fucked up, piss, arseclown, arsehole, ass hole, fatass, piece of shit, pompousAhole, arsewipe |
| Race | nigger, nigga, niggas, niggers, sandnigger |
| Religion | islam, islamic, jesus, god, devil, hell, god king |
| Sexual orientation | gay, gays, lesbian, fag, faggot, faggots, faggot club, queer, fuck, fucking, fuckin, cocksucker, fagget, fucken, fcking, fking |
| Social class | bastard, bastards, sucker, hoe, hoes, slut, whore |
| Others | crap, bullcrap, piece, shithead, shit, damn, damnit, fucker, motherfucker, motherfucking, fucked, goatfuckers, fuckhead, fuckass, go to hell, like hell, hole, worthless, mfer, mfs, useless |

To remove dialectal variations, the retrieved tweets were geographically constrained to metropolitan areas that are within 10,000 miles from the country's capital state (i.e., Canberra, Australia; Washington DC, the United States; and Kuala Lumpur, Malaysia). Some of the tweets collected in the current study were written in a mixture of English and other languages. This was especially so for tweets written by users from Malaysia, a country which has three major languages—Malay, Chinese, and English. Hence, such adulterated tweets were manually removed from the collected data to maintain language consistency across the samples of tweets.

As aforementioned, tweets that were written in a mixture of English and other languages were removed from the dataset to maintain language consistency. Nearly half of the tweets (1908 tweets) by users from Malaysia were discarded, while only three tweets and one tweet by users from the United States and Australia respectively were removed. Only 24340 tweets remained after this cleaning process. The distribution of tweets in accordance to their profanity category across the three countries are presented in Figure 1. The sexual orientation category seems to be the dominating category (not considering the others category) of all countries, particularly Australia.




| | Sexual orientation | Gender | Physical | Religion | Disability | Race | Behaviour | Ethnicity | Social Class | Others |
|---|---|---|---|---|---|---|---|---|---|---|
| Malaysia | 27.50% | 18.60% | 7.70% | 7.30% | 1.50% | 1.30% | 0.60% | 1.20% | 1.10% | 33.20% |
| United States | 20.60% | 20.90% | 9.70% | 6.00% | 2.00% | 3.70% | 1.90% | 0.60% | 0.80% | 33.80% |
| Australia | 39.30% | 10.10% | 3.20% | 5.90% | 3.80% | 1.60% | 2.30% | 1.10% | 0.40% | 32.30% |

Figure 1: The distribution of tweets according to their profanity category across Australia, the United States, and Malaysia.

### 3.3. Empirical study

Figure 1 shows that the usage of profanity in different categories varies among different countries. To confirm the significance of such differences, a Chi-square test was performed on the distribution of tweets according to their profanity category across the three countries. The testing result indicates a statistically significant ($p < .05$) difference among the three countries. This result also supported the observation that Twitter users in Australia utilised more sexual-oriented terms in their tweets than those in the other two countries.

Prior to employing Equations (3.1) and (3.2) to assess the effectiveness of using profane words to detect hate speech, the presence or absence of hatred in the tweets was first manually reviewed and identified by human coders. Each tweet was read by three human coders and the presence or absence of hatred was determined by a majority vote. However, due to the costly nature of such manual annotation, only a sub-sample of 3000 tweets from the original sample (1000 tweets from each country) was used for this purpose.

Equations (3.1) and (3.2) require the estimates of prob $(f)$ and prob $(h)$, which were obtained from past research. The probability of profane words being present in a tweet, prob(f), was estimated to be 7.73% (of 51 million tweets) according to Wang et al. [38].

The probability that a tweet contained hatred, prob $(h)$, was more difficult to estimate due to the overwhelming number of posts constantly being generated. Davidson et al. [10] searched for tweets containing terms from the Hatebase lexicon and identified a sample of tweets from 33458 Twitter users. They then extracted the timeline for each user, resulting in a set of 85.4 million tweets of which 24802 tweets containing terms from the Hatebase lexicon was randomly sampled and manually coded by CrowdFlower workers. The study found that only 5% of the tweets were labeled as hate speech.

Table 3: The probability of hate speech occurring with and without profane words.

| Country | Profanity | prob $(h \mid f)$ | prob $(h \cap f)$ | prob $(h \cap \neg f)$ | prob $(h \mid \neg f)$ |
|---|---|---|---|---|---|
| Australia | All | 45.00% | 3.48% | 1.52% | 1.65% |
| | Sexual orientation | 27.80% | 0.84% | 4.16% | 4.29% |
| | Gender | 6.90% | 0.05% | 4.95% | 4.98% |
| United States | All | 19.40% | 1.50% | 3.5% | 3.79% |
| | Sexual orientation | 4.90% | 0.10% | 0.40% | 5.00% |
| | Gender | 5.10% | 0.07% | 4.93% | 5.00% |
| Malaysia | All | 23.90% | 1.85% | 3.15% | 3.42% |
| | Sexual orientation | 5.80% | 0.12% | 4.88% | 4.98% |
| | Gender | 9.30% | 0.13% | 4.87% | 4.94% |

It could be biased to use the resulting hate speech percentage (i.e., 5%) obtained by Davidson et al. [10] as the estimate of prob(h), since the sample of Davidson et al. [10] was based on the Hatebase lexicon. In other words, the sample was from Twitter users who were already likely to use profane terms. In another study, Van Hee et al. [37] collected 113698 posts in English and 78387 posts in Dutch from social networking site ASK.fm, where the ratio of posts in English involving bullying was 4.73%. This ratio was very close to the one by Davidson et al. [10]. Thus, the current study adopted 5% as the estimate of prob $(h)$.

The computational results of the probabilities based on Equations (3.1) and (3.2) are presented in Table 3, where the column of "Profanity" indicates the computation of probabilities is performed with all categories in Table 2 as a whole, or based solely on sexual orientation or gender categories. Sexual orientation and gender categories were particularly highlighted as they were the major profanity categories based on the statistics shown in Figure 1. The computation of prob $(h \mid f)$ was directly obtained from the retrieved tweets of the current study as all tweets contained profanity. Thus, the ratio of hate instances in the sample can be used as an estimate of prob $(h \mid f)$.

Based on Table 3, it can be observed that the capability of using profanity to detect hate speech is limited. For Malaysia, only 23.9% of the tweets containing profane words can be considered as hate speech. Only 5.8% of tweets containing sexual-oriented profanity were hate speeches while 9.3% of tweets containing gender-related profanity were hate speeches. The effectiveness rate was only 19.4% for the United States. In comparison, the use of profanity in hate speech detection for tweets from Australia was more effective (45%).

Such differences may imply that Twitter users in both Malaysia and the United States tend to use profanity in their tweets without hate intent. On the contrary, Twitter

users in Australia tend to include profanity in their tweets with hate intent, especially sexual-oriented profanity (27.8%). The probability of missing a hate speech by profanity checking (i.e., $\text{prob}(h \mid \neg f)$ for the three countries were 1.65% for Australia, 3.79% for the United States, and 3.42% for Malaysia. This low result for Australia is associated with the higher $\text{prob}(h \mid f)$ estimate among tweets from Australia.

The above results imply that using profanity to detect hate speech is still a feasible method, but improvement is needed to distinguish actual hate speech from false detection. Despite the limited capacity of using profanity as an initial screen for hate speech detection, it could still be a useful method under certain conditions.

### 3.4. The effect of profanity ratio in the probability estimation

The probability estimation results in Table 3 question the effectiveness of using profanity to detect hate speech. However, the results also indicate that using profanity as an initial screen for hate speech detection could be more effective under certain conditions (i.e., tweets from Australia). Next, we explore the possible conditions that make profanity-based methods more feasible and discuss the advantage of using the profanity-based methods in dealing with the issue of class imbalance.

The reported results in table 3 imply that profanity can only detect a minor portion of tweets with hate intent from the United States and Malaysia, as indicated by the estimates of $\text{prob}(h \cap f)$ and $\text{prob}(h \cap \neg f)$. Considering the overall ratio of hate speech to be 5%, as assumed in the previous section, then only 30% (i.e., $\frac{\text{prob}(h \cap f)}{\text{prob}(h)} = \frac{1.50\%}{5\%}$) of hate tweets from the United States contained profanity while 37% (i.e., $\frac{\text{prob}(h \cap f)}{\text{prob}(h)} = \frac{1.85\%}{5\%}$) of hate tweets from Malaysia contained profanity. In other words, 70% and 63% of hate tweets from the United States and Malaysia respectively did not contain any profanity, which suggest the ineffectiveness of using profanity in detecting hate speech.

On the other hand, 69.6% (i.e., $\frac{\text{prob}(h \cap f)}{\text{prob}(h)} = \frac{3.84\%}{5\%}$) of hate tweets from Australia contained profanity, which greatly reduced the rate of false negative instances (i.e., $\text{prob}(h \mid \neg f)$) to 1.65%. The profanity-based method is seemingly more effective for tweets from Australian. Therefore, the feasibility of using profanity to detect hate speech could be dependent on the profanity ratio (i.e., $p(f)$) and $\text{prob}(h \cap f)$.

However, no evidence exists so far regarding the relation between the profanity ratio and $\text{prob}(h \mid f)$. In the current study's probability estimation, $p(f)$ was assumed the same (7.73%) across all countries. To examine the effect of the profanity ratio, the current study varied the value of $p(f)$ and observed the changes of the probability estimates under the assumptions of different $\text{prob}(h \mid f)$ estimates (see Figure 2).

Since the estimate of $\text{prob}(h \mid f)$ was assumed to be fixed, $\text{prob}(h \cap f)$ was found to increase linearly when $p(f)$ increases. When $\text{prob}(h \mid f)$ was low (e.g., $\text{prob}(h \mid f) = 0.055$ or 0.1), the $p(f)$ had to be very high to make the false negative rate ($\text{prob}(h \mid \neg f)$) acceptable. When $\text{prob}(h \mid f)$ was moderate (e.g., $\text{prob}(h \mid f) = 0.2$ or 0.3), a $\text{prob}(h \mid f)$ value

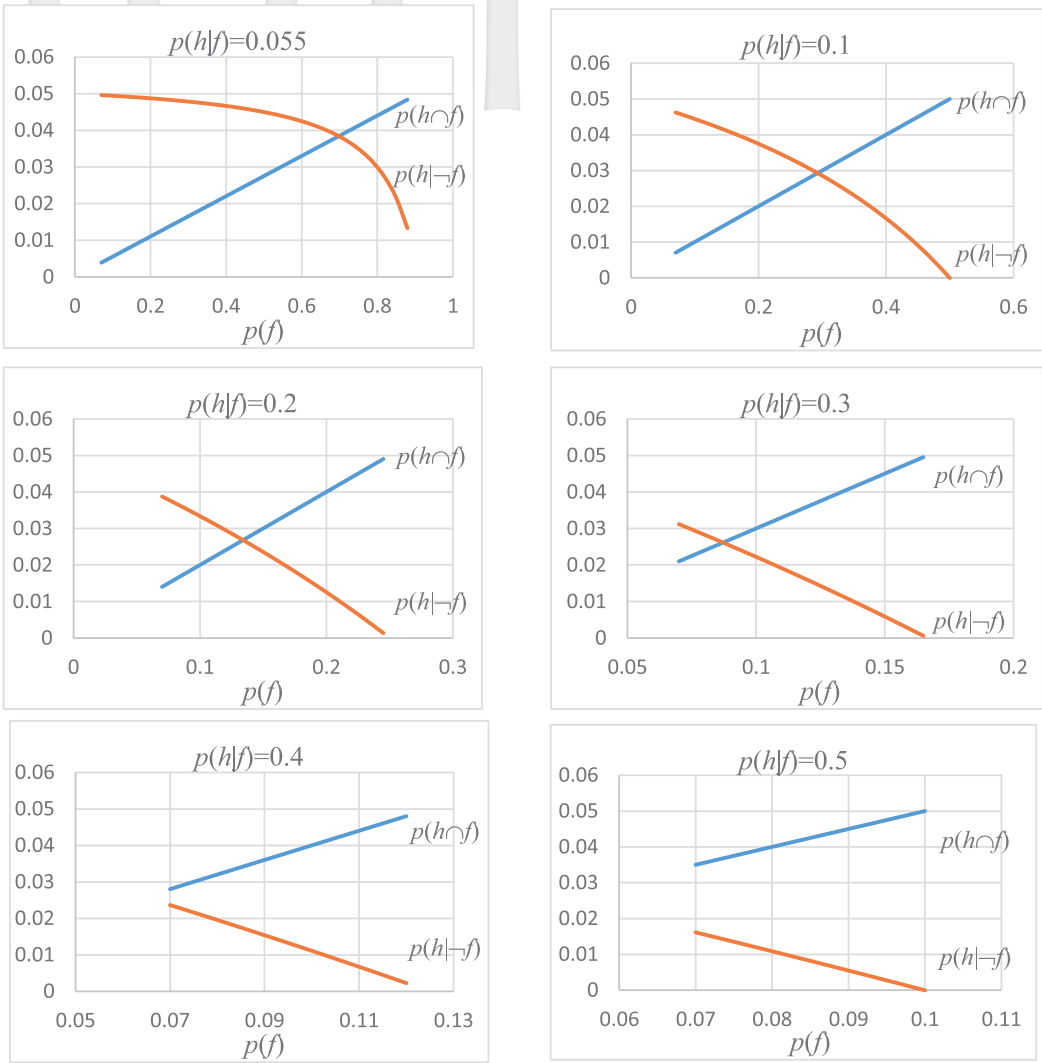



Figure 2: Probability estimates under diferrent profanity ratios.

greater than 0.15 yielded a satisfactory false negative rate. Finally, when prob $(h \mid f)$ was high (e.g., prob $(h \mid f)$=0.4 or 0.5), the profanity-based methods were found suitable for hate speech detection as the false negative rate dropped fast when prob $(h \mid f)$ increased slightly. Thus, it could be concluded that the use of profanity in detecting hate speech would be more effective when prob $(h \mid f)$ and $p(f)$ are higher.

Although the estimate of $p(f)$ by Wang et al. [37] was based on a very large set of tweets, such a probability would likely be varied in different subdomains, such as geographical areas or political subjects. Therefore, the estimates of such probability with respect to individual subdomains are meaningful when determining if profanity should be used in detecting hate speech in certain subdomains. Such estimates will be carried out in a future study.

## 4. Hate Speech Detection by Deep Learning based on Profanity Screening

Three deep learning methods, long short-term memory (LSTM), bidirectional LSTM (BLSTM), and bidirectional encoder representations from transformers (BERT) are adopted for hate spc detection. LSTM is a special kind of recurrent neural network (RNN) capable of processing arbitrary sequences of inputs with its internal memory, and is effective for text classification (see Johnson and Zhang [19]). With the capabilities of sequential inputs processing and memory of precedent inputs, LSTM is ideal for time series pattern reorganization. The words in a sentence form a sequence and the relations between preceding and succeeding words are strong. Thus, LSTM has been applied to and succeed in domains of speech recognition, language modeling, and translation. Considering the outstanding performance of LSTM in natural language processing, this study adopts LSTM in the classification of hate speech. BLSTM is a variation of LSTM, aiming to improve the performance of LSTM. BERT Devlin et. al. [13] is a multi-layer bidirectional Transformer encoder based on the original implementation described in Vaswani et al. [38] and is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.

This section first introduces the three deep learning methods, and then performs the hate speech classification on the three-country dataset of the previous section, and then the dataset of Davidson et al. [10], respectively.

### 4.1. LSTM, BLSTM and BERT

The neuron of LSTM consists of three gates to control the learning of memory overtime: a forget gate, an input gate and an output gate as shown in Figure 3. The input gate controls the admission of an input $z$ to the memory by the product of the

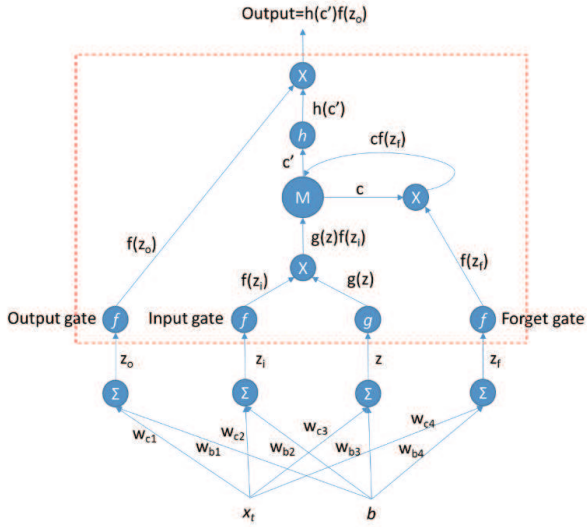



Figure 3: The cell of LSTM (see Shyur et al. [31]).

input activation, $g(z)$, and the input gate activation $f(z_i)$; the forget gate determines how much the memory is maintained with the product of the old memory $c$ and the forget gate activation $f(z_f)$, and the memory is updated by $c' = g(z)f(z_i) + cf(z_f)$; and the output gate controls the degree that the memory is passed to other cells. The above operations allow earlier inputs being kept in the memory cell until the forget gate is closed, and enable the network to learn and determine how long to hold the old memory, and how to associate the old memory to the new inputs.

The conceptual architecture of the LSTM model for classifying hate/not-hate speeches is presented in Figure 4. The input layer is a Bag-of-word (BOW) representation of the input text. Tokens of the dictionary are obtained by discovering the top 50000 most frequent words among all collected tweets. The embedding layer with 32 nodes transforms the tokenized text into word vectors. To prevent from overfitting, a dropout layer with a 0.5 dropout rate is added after the embedding layer. The LSTM layer consisting of 32 nodes receives the word vectors from the embedding layer as its inputs. The fully connected layer is a regular hidden layer of the traditional multi-layer perceptron neural network (MLP), and again another dropout layer with the same dropout rate is added after it. Finally, the output layer produces the classification result in a one-hot-encoding format. The model is implemented by Keras.

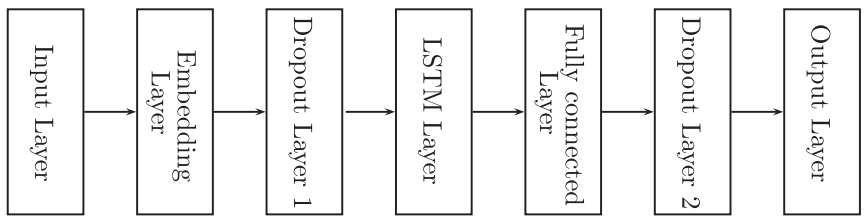



Figure 4: Architecture of the LSTM classification model.

Graves and Schmidhuber [16] proposed the bidirectional LSTM (BLSTM) that extends the standard LSTM network to enhance the prediction performance. The structure of a BLSTM network is presented in Figure 5, where, two hidden layers from the two opposite direction connect to the same output, one feeding forward and another one backwards in time. BLSTM learns the representation of data from past time steps and future time steps simultaneously to double the amount of input information. The architecture of the BLSTM used in our classification similar to the one presented in Figure 4, with the LSTM layer replaced by a BLSTM layer.

The construction of BERT involves two steps: 1) pre-training the model on unlabelled data over different tasks, and 2) fine-tuning the pre-trained parameters using labelled data from the downstream task. BERT is pre-trained using two unsupervised tasks: a Masked Language Model (MLM) to pre-train objective, and a Next Sentence Prediction (NSP) to understand sentence relationships. MLM randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. It enables the representation to fuse the left and the right context, which allows the pre-training of a deep bidirectional Transformer.

At the fine-tuning stage, the self-attention mechanism in the Transformer allows BERT to model any downstream task. BERT with self-attention encodes a concatenated text pair, which effectively includes bidirectional cross attention between two sentences. For each task, just simply plug in the task inputs and outputs into BERT and fine-tune all the parameters end to end.

### 4.2. Classification for the three-country dataset

The three-country dataset is divided into a training set and a testing set with the ratio of 80% to 20%. The performance metrics of the testing results by three different deep learning methods for the three countries, respectively, are presented in Table 4. It has to be noted that the true measures of recalls are slightly lower that those shown in Table 4 since those hate instances containing no profanity did not enter the dataset.

The three deep learning methods outperform each other in some cases. All methods perform well on the data of Australia comparing to that on the data of the other two countries. This performance difference is likely attributed to the class imbalance problem in the data of Malaysia and USA. The probabilities prob $(h \mid f)$ in Table 3 show that the two classes (hate/not-hate) are balanced after profanity screening for Australia tweets, but they are unbalanced for that of Malaysia and USA. To confirm our observation, we oversample the hate instances in the Malaysia and the USA tweets, and perform the training by LSTM. The results are shown in Table 5, where two different sizes of oversampling are done for each country, i.e. double and triple of the hate instances. The results in Table 5 show that the performance of the LSTM models are indeed improved, but still not comparable that by the Australia data for the overfitting tendency by oversampling.

### 4.3. Classification for the dataset of Davidson et al. [10]

We further justify the effect of prob $(h \mid f)$ and $p(f)$ toward the performance of the profanity screening method based on an annotated dataset of hate tweets provided by Davidson et al. [10]. The dataset contained 24783 instances, in which 1430 were annotated as hate speech. The profane words in Table 2 were used to match the dataset and as a result, 19952 instances containing profanity were identified, which accounted for more than 80% of the entire dataset. This extraordinarily high ratio of profanity was possibly due to the way these tweets were collected. Davidson et al. [10] retrieved the tweets by tracking the users with a record of using profanity words; hence, the tweets retrieved by this manner naturally have a high tendency of profanity.

From the instances identified as profanity, the current study tried four different sampling sizes—250, 500, 750, and 1000 tweets—and three replications for each sample size. Again, it was assumed that the probability of hate speech in the population was 5%. However, the probability of profanity was set at 80% to comply with the dataset of Davidson et al. [10]. The resulting probability estimates are presented in Table 6. The probabilities of prob $(h \mid f)$ and prob $(h \cap f)$ were relatively steady across all sample sizes, but prob $(h \mid \neg f)$ was quite sensitive to the number of hate instances contained in the

Table 4: Classification performance by three deep learning methods.

| Country | Method | Precision | Recall | Accuracy | f1-score |
|---|---|---|---|---|---|
| Australia | LSTM | 0.93 | 0.77 | 0.87 | 0.84 |
| | BLSTM | 0.90 | 0.68 | 0.83 | 0.77 |
| | BERT | 0.87 | 0.76 | 0.85 | 0.81 |
| Malaysia | LSTM | 0.47 | 0.47 | 0.81 | 0.47 |
| | BLSTM | 0.55 | 0.50 | 0.75 | 0.52 |
| | BERT | 0.58 | 0.50 | 0.78 | 0.51 |
| USA | LSTM | 0.56 | 0.35 | 0.80 | 0.43 |
| | BLSTM | 0.41 | 0.34 | 0.74 | 0.37 |
| | BERT | 0.60 | 0.29 | 0.80 | 043 |
| Overall | LSTM | 0.60 | 0.61 | 0.81 | 0.60 |
| | BLSTM | 1.00 | 0.37 | 0.86 | 0.54 |
| | BERT | 0.76 | 0.61 | 0.87 | 0.68 |

Table 5: Testing results of LSTM models with oversampling.

| | | Precision | Recall | Accuracy | f1-score |
|---|---|---|---|---|---|
| Malaysia | Double | 0.55 | 0.61 | 0.84 | 0.58 |
| | Triple | 0.51 | 0.53 | 0.83 | 0.52 |
| USA | Double | 0.65 | 0.40 | 0.83 | 0.50 |
| | Triple | 0.52 | 0.39 | 0.79 | 0.45 |

sample. The probability of $\text{prob}\,(h \mid \neg f)$ became steady when the sample size became larger (i.e., 1000). The results in Table 6 generally comply with the first case (i.e., $\text{prob}\,(h \mid f) = 0.055$) in Figure 2, suggesting that the profanity-based method is not an ideal approach for this case. To confirm this expectation, the same three deep learning models used earlier are applied to the dataset of Davidson et al. [10] after profanity screening. As expected, their performances on hate speech prediction are not good, with precision=0.21, recall=0.44 and f1-score=0.28 for LSTM; precision=0.52, recall=0.15 and f1-score=0.23 for BLSM; and precision=0.5, recall=0.40 and f1-score=0.44 for BERT.

## 5. Concluding Remarks

The current study investigated the issue of profanity usage on Twitter across differ-

Table 6: Probability estimates based on the dataset of (see Davidson et al. [10]).

| Sample | # hate instances | prob $(h \mid f)$ | prob $(h \cap f)$ | prob $(h \cap \neg f)$ | prob $(h \mid \neg f)$ |
|---|---|---|---|---|---|
| Original | 1430 (5.78%) | 5.66% | 4.56% | 1.21% | 6.23% |
| Sample | (1)13 (5.20%) | 5.20% | 4.16% | 0.84% | 4.20% |
| size= | (2)14 (5.60%) | 5.60% | 4.48% | 0.52% | 2.60% |
| 250 | (3)11 (4.40%) | 4.40% | 3.52% | 1.48% | 7.40% |
| Sample | (1)28 (5.60%) | 5.60% | 4.48% | 0.52% | 2.60% |
| size= | (2)29 (5.80%) | 5.80% | 4.64% | 0.36% | 1.80% |
| 500 | (3)21 (4.20%) | 4.20% | 3.36% | 1.64% | 8.20% |
| Sample | (1)40 (5.33%) | 5.33% | 4.27% | 0.73% | 3.67% |
| size= | (2)38 (5.07%) | 5.07% | 4.05% | 0.95% | 4.73% |
| 750 | (3)34 (4.53%) | 4.53% | 3.63% | 1.37% | 6.87% |
| Sample | (1)50 (5.00%) | 5.00% | 4.00% | 1.00% | 5.00% |
| size= | (2)49 (4.90%) | 4.90% | 3.92% | 1.08% | 5.40% |
| 1000 | (3)51 (5.10%) | 5.10% | 4.08% | 0.92% | 4.60% |

ent user groups and formulated a probability estimation procedure based on the Bayes theorem to quantify the effectiveness of using profanity-based methods in hate speech detection. Tweets from Australia, the United States, and Malaysia, were collected based on a set of profane words used as keywords. These three countries use English as an official language but have different cultural backgrounds. The collected tweets were cleaned and annotated by human coders, and were used to support the proposed probability estimation procedure and three deep learning methods, LSTM, BLSTM and BERT, for hate speech detection.

The results from this study suggest that different user groups indeed use profanity in tweets in different manner, and that such a manner affects the effectiveness of using profanity-based methods in detecting hate speech. In particular, the results show that for Australia tweets, where profanity is more associated with hatred, profanity-based methods in hate speech detection could be effective and profanity screening can address the class imbalance issue in hate speech detection. This was evidenced by the performance of using LSTM deep learning model on the profanity screened data of Australia data, which achieved a classification f1-score of 0.84.

This study obtained the data sample by retrieving tweets with keywords suggested by Wang et al. [39] and Teh et al. [34]. Although they have shown the efficiency of using such keywords in obtaining hatred instances, it implies that there are hatred instances not included in our data sample since containing no keywords. The patterns resided in

these un-retrieved hate instances might be very different from those contain keywords, and thus the deep learning methods fail to detect hatred cases of this type. In our future study, we will extend the data sample to include hatred instances without profanity. The collection of such instances demands a great effort of human review, since they constitute an extremely minor share of the entire tweets. A more convenient way to collect such instances would be to utilize the abusive behavior report by Twitter. Also, our future study will explore more features in hate speech detection from the structure of a hatred sentence by employing graph computing. The skip n-gram of words in a sentence could provide the information regarding the pattern of a hatred sentence. We consider to model the relations between words by graphs and obtain features from such relations via graph computing.

Currently, most studies on hate speech online have focused on the three major platforms: Facebook, YouTube and Twitter. However, their predominance as the biggest international social networks is no longer uncontested. Other networks are on the rise and young users especially lose interest in the 'old' platforms, e.g., In April 2019, Instagram had more active accounts globally than Twitter. Since hate groups and extremists move their propaganda to the new social media where they can reach their target audience most easily, it is important to take those changes in the social media landscape into consideration Criley [9]. Thus, our future study will also extend the data sample to the emerging and growing social media.

## References

[1] Agrawal, S. and Awekar, A. (2018). *Deep learning for detecting cyberbullying across multiple social media platforms*, In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 10772 LNCS, pp. 141-153).

[2] Arnett, J. J. (1995). *Adolescents' uses of media for self-socialization*, Journal of Youth and Adolescence, Vol.24, No.5, 519-533.

[3] Bak, J. Y., Kim, S. and Oh, A. (2012). *Self-disclosure and relationship strength in twitter conversations*, 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference, 2(July), 60-64.

[4] Burnap, P. and Williams, M. L. (2015). *Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making*, Policy and Internet, Vol.7, No.2, 223-242.

[5] Bushman, B. J. and Cantor, J. (2003). *Media Ratings for Violence and Sex: Implications for Policymakers and Parents*, American Psychologist, Vol.58, No.2, 130-141.

[6] Cameron, P. (1970). *The words college students use and what they talk about*, Journal of Communication Disorders, Vol.3, No.1, 36-46.

[7] Clement, J. (2020). Twitter: most users by country | Statista. Retrieved June 29, 2020, from `https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/`

[8] Cressman, D. L., Callister, M., Robinson, T. and Near, C. (2009). *Swearing in the Cinema*, Journal of Children and Media, Vol.3, No.2, 117-135.

[9] Criley, R. A. (2001). *Beyond the "big three" Alternative platforms for online hate speech*, Acta Horticulturae, Vol.545, 79-85.

[10] Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017). *Automated hate speech detection and the problem of offensive language*, Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, 512-515.

[11] Delgado, R. and Stefancic, J. (2014). *Hate Speech in Cyberspace*, In Wake Forest Law Review, Vol.49, 319-344.

[12] Deseret, N. (n.d.). *What the heck? Casual cursing by teens is rising - Deseret News*, Retrieved June 29, 2020, from `https://www.deseret.com/2008/2/25/20072690/what-the-heck-casual-cursing-by-teens-is-rising`

[13] Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*, NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), 4171-4186.

[14] Eddy, M. and Scott, M. (n.d.). *Delete Hate Speech or Pay Up, Germany Tells Social Media Companies - The New York Times*, Retrieved June 29, 2020, from `https://www.nytimes.com/2017/06/30/business/germany-facebook-google-twitter.html`

[15] Feldman, G., Lian, H., Kosinski, M. and Stillwell, D. (2017). Frankly, We Do Give a Damn. Social Psychological and Personality Science, 194855061668105.

[16] Graves, A. and Schmidhuber, J. (2005). *Framewise phoneme classification with bidirectional LSTM and other neural network architectures*, Neural Networks, Vol.18, No.5-6, 602-610.

[17] Hatebase. (n.d.). Retrieved June 29, 2020, from `https://hatebase.org/`

[18] Jay, T. (2009). *Do Offensive Words Harm People*? Psychology, Public Policy and Law, Vol.15, No.2, 81-101.

[19] Johnson, R. and Zhang, T. (2016). *Supervised and semi-supervised text categorization using LSTM for region embeddings*, 33rd International Conference on Machine Learning, ICML 2016, 2, 794-802.

[20] Kaye, B. K. and Sapolsky, B. S. (2009). *Watch Your Mouth! An Analysis of Profanity Uttered by Children on Prime-Time Television*, Mass Communication and Society, 5436(November 2014), 37-41.

[21] Malmasi, S. and Zampieri, M. (2018). *Challenges in discriminating profanity from hate speech*, Journal of Experimental and Theoretical Artificial Intelligence, Vol.30, No.2, 187-202.

[22] Nemes. (2002). *Detecting hate speech on the world wide web*, Journal of Communication Disorders, Vol.1, No.2, 215-230.

[23] Nerbonne, G. P. and Hipskind, N. M. (1972). *The use of profanity in conversational speech*, Journal of Communication Disorders, Vol.5, No.1, 47-50.

[24] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. and Chang, Y. (2016). *Abusive language detection in online user content*, 25th International World Wide Web Conference, WWW 2016, 145-153.

[25] Parekh, B. (2006). *Hate Speech: Is there a case for banning*? Public Policy Research, Vol.12, No.4, 213-223.

[26] Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, Jimenez, R., Knight, D., Kren, M., Lofberg, L., Nawab, R., Shafi, J., Teh, P., Mudraya, O. (2016). *Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages*, Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, 2614-2619.

[27] Rayson, P. (2008). *From key words to key semantic domains*, International Journal of Corpus Linguistics, Vol.13, No.4, 519-549.

[28] Razavi, A. H., Inkpen, D., Sasha, U. and Matwin, S. (2010). *Offensive language detection using multi-level classification.pdf*.

[29] Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. gyo, Almerekhi, H. and Jansen, B. J. (2020). *Developing an online hate classifier for multiple social media platforms*, Human-Centric Computing and Information Sciences, Vol.10, No.1, 1-34.

[30] Sapolsky, B. S. and Kaye, B. K. (2005). *The Use of Offensive Language by Men and Women in Prime Time Television Entertainment*, Atlantic Journal of Communication, Vol.13, No.4, 292-303.

[31] Shyur, H., Cheng, C.-B. and Hsiao, Y. (n.d.). Using Deep Learning Approach in Flight Exceedance Event Analysis, 1-16.

[32] Silva, L., Mondal, M., Correa, D., Benevenuto, F. and Weber, I. (2016). *Analyzing the targets of hate in online social media. Proceedings of the* 10*th International Conference on Web and Social Media*, ICWSM 2016, (June), 687-690.

[33] Sood, S. O., Antin, J. and Churchill, E. F. (2012). *Profanity use in online communities*, Conference on Human Factors in Computing Systems - Proceedings, (April), 1481-1490.

[34] Teh, P. L., Cheng, C. Bin and Chee, W. M. (2018). *Identifying and categorising profane words in hate speech*, ACM International Conference Proceeding Series, 65-69.

[35] Thelwall, M. (2008). *Fk yea I swear: cursing and gender in MySpace*, Corpora, Vol.3, No.1, 83-107.

[36] Tweet Archiver - G Suite Marketplace. (n.d.). Retrieved June 29, 2020, from `https://gsuite.google.com/marketplace/app/tweet_archiver/976886281542`

[37] Van Hee, C., Jacobs, G., Emmery, C., DeSmet, B., Lefever, E., Verhoeven, B., Pauw, G., Daelemans, W., Hoste, V. (2018). *Automatic detection of cyberbullying in social media text*, PLoS ONE, Vol.13, No.10, 1-23.

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. (2017). *Attention is all you need*, Advances in Neural Information Processing Systems, 2017-Decem(Nips), 5999-6009.

[39] Wang, W., Chen, L., Thirunarayan, K. and Sheth, A. P. (2014). *Cursing in english on twitter*, Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 415-425.

[40] Wong, S. C., Teh, P. L. and Cheng, C.-B. (2020). *How Different Genders Use Profanity on Twitter?* In International Conference on Compute and Data Analysis (pp. 1-9).

[41] Xiang, G., Fan, B., Wang, L., Hong, J. and Rose, C. (2012). *Detecting offensive tweets via topical feature discovery over a large scale twitter corpus*, ACM International Conference Proceeding Series, 1980-1984.

Department of Computing and Information System, Faculty of Science and Technology, Sunway University, Malaysia.

E-mail: phoeyleet@sunway.edu.my

Major area(s): Text and sentiment analysis, information extraction and visualization.

Department of Information Management, Tamkang University, Taiwan, ROC.

E-mail: cbcheng@mail.tku.edu.tw (Corresponding author)

Major area(s): Soft computing, machine learning, decision analysis.

(Received September 2019; accepted September 2020)