

This preprint article has been accepted for publication in International Journal of Information and Management Sciences. <http://ijims.ms.fku.edu.tw/IJIMS/contents/recruit.php?class=106>

Phoey Lee Teh, Chi-Bin Cheng. (2020) Profanity and Hate Speech Detection. International Journal of Information and Management Sciences. Vol 31, Issue 3, Pages 227-246.

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 greater than 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,

I < intensity > < user intent > < 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 $\text{prob}(h|f)$ (i.e., $1 - \text{prob}(h|f)$), defined as

$$\text{prob}(h|f) = \frac{\text{prob}(h \cap f)}{\text{prob}(f)} \quad (3.1)$$

It was possible to compute $\text{prob}(h \cap f)$ from $\text{prob}(h|f)$ and $\text{prob}(f)$, where $\text{prob}(h|f)$ 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 $\text{prob}(h|\neg f)$. Again, this can be derived by:

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

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 Piao [26] and Rayson [27], to process the collected posts to extract key words 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.

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.

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.

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.

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

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 Fig. 1. The sexual orientation category seems to be the dominating category (not considering the others category) of all countries, particularly Australia.

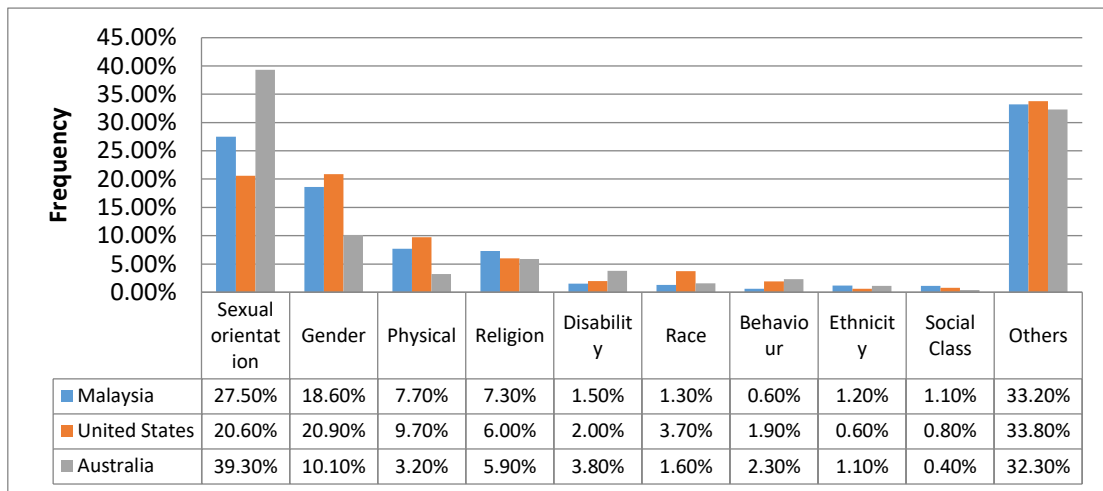


Figure 1: The distribution of tweets according to their profanity category across Australia, the United States, and Malaysia.

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.

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 Fig. 1. The computation of $prob(h|f)$ 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 $prob(h|f)$.

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

Country	Profanity	$prob(h f)$	$prob(h \cap f)$	$prob(h \cap \neg f)$	$prob(h \neg f)$
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%

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. $prob(h|\neg f)$) 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 $prob(h|f)$ 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 $prob(h \cap f)$ and $prob(h \cap \neg f)$. Considering the overall ratio of hate speech to be 5%, as assumed in the previous section, then only 30% (i.e., $\frac{prob(h \cap f)}{prob(h)} = \frac{1.50\%}{5\%}$) of hate tweets from the United States contained profanity while

37% (i.e., $\frac{prob(h \cap f)}{prob(h)} = \frac{1.85\%}{5\%}$) 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., $\frac{prob(h \cap f)}{prob(h)} = \frac{3.84\%}{5\%}$) of hate tweets from Australia contained profanity, which greatly reduced the rate of false negative instances (i.e., $prob(h|\neg f)$) 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 $prob(h \cap f)$.

However, no evidence exists so far regarding the relation between the profanity ratio and $prob(h|f)$. 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 $prob(h|f)$ estimates (see Fig. 2).

Since the estimate of $prob(h|f)$ was assumed to be fixed, $prob(h \cap f)$ was found to increase linearly when $p(f)$ increases. When $prob(h|f)$ was low (e.g., $prob(h|f) = 0.055$ or 0.1), the $p(f)$ had to be very high to make the false negative rate ($prob(h|\neg f)$) acceptable. When $prob(h|f)$ was moderate (e.g., $prob(h|f) = 0.2$ or 0.3), a $prob(h|f)$ value greater than 0.15 yielded a satisfactory false negative rate. Finally, when $prob(h|f)$ was high (e.g., $prob(h|f) = 0.4$ or 0.5), the profanity-based methods were found suitable for hate speech detection as the false negative rate dropped fast when $prob(h|f)$ increased slightly. Thus, it could be concluded that the use of profanity in detecting hate speech would be more effective when $prob(h|f)$ and $p(f)$ are higher.

This is a preprint and brief. Please refer to the original source of the journal at:

<http://ijims.ms.tku.edu.tw/IJIMS/contents/recruit.php?class=106>

Phoey Lee Teh, Chi-Bin Cheng. (2020) Profanity and Hate Speech Detection. International Journal of Information and Management Sciences. Vol 31, Issue 3, Pages 227-246.