How Different Genders Use Profanity on Twitter?

Shang Cheong Wong
Department of Computing and Information Systems
School of Science and Technology
Sunway University, Malaysia.
+63 7491 8622
Shang.w@imail.sunway.edu.my

Phoey Lee Teh
Department of Computing and Information Systems
School of Science and Technology
Sunway University, Malaysia
+63 7491 8622 Ext:7151
phoeyleet@sunway.edu.my

Chi-Bin Cheng
Department of Information Management
Tamkang University,
Tamsui, Taiwan.
+886 2 2621 5656
cbcheng@mail.tku.edu.tw

ABSTRACT
Social media, is often the go-to place where people discuss their opinions and share their feelings. As some platforms provide more anonymity than others, users have taken advantage of that privilege, by sitting behind the screen, the use of profanity has been able to create a toxic environment. Although not all profanities are used to offend people, it is undeniable that the anonymity has allowed social media users to express themselves more freely, increasing the likelihood of swearing. In this study, the use of profanity by different gender classes is compiled, and the findings showed that different genders often employ swear words from different hate categories, e.g. males tend to use more terms from the "disability" hate group. Classification models have been developed to predict the gender of tweet authors, and results showed that profanity could be used to uncover the gender of anonymous users. This shows the possibility that profiling of cyberbullies can be done from the aspect of gender based on profanity usage.

CCS Concepts
• **Information systems→Information Systems→Information systems applications→ Data Mining→Association rules**.

Keywords
Profanity; Tweets; Data Analysis; Gender

1. INTRODUCTION
Social media. People use it often for various reasons, sharing thoughts, feelings, ideas, and opinions. There are 400 million tweets sent per day, and 1,000 comments are sent every second on Instagram [1]. Social media has not only enabled us to connect with people who speak different tongues but has also bred common tongues, which are cross-continent. For instance, "brb", "omg", or even "wtf". If these examples seem unfamiliar, social media now has "urban dictionary" to help the "illiterate" understand "social media language". If the reader has seen or used some of these terms before, it is then self-evident that social media has changed the way we communicate with each other.Following the example of social media language i.e. "wtf", some who read this term may feel offended while others may deem its use necessary to express their overwhelming surprise towards an experience. In worse cases, profanity has been used to hurt others emotionally behind the screens [2] [3]. This cause serious problems due to social media's openness to public participation [4]. With so many users on different platforms of social media, negatively motivated content can spread instantly and its effect multiplied at unthinkable rates [5][6]. Furthermore, although the example of hate crime used was carried out by a woman, females are two times more likely to become victims of cyberbullying [7] , and it is impossible that all those acts were carried out females alone. The lack of evidence pointing towards a certain gender is due to the argument of gender discrimination. Despite that, it no less means that males can also be perpetrators of hate or cyberbullying on social media. The next question: with the same capability to do evil, do male and females do it differently? To understand if there is a difference in the way males or females write, from previous studies, differences in language styles have been identified [8-11]. It is well understood that there are many dimensions to look at when analysing gender-linked language but unfortunately, few to none have been found looking in the direction where the use of profanity is studied [12]. With that said, the end goal of this study is to investigate the possible relationship between gender and profanity use.

ICCDA 2020, March 9–12, 2020, Silicon Valley, CA, USA

ACM ISBN 978-1-4503-7644-0/20/03…$15.00
https://doi.org/10.1145/3388142.3388145

2. LITERATURE REVIEW
People use social media to seek social interaction [13], to gain information [14], to pass the time, to watch others and even to seek employment [15]. Social media is largely motivated by social interactions [16-18]. People are more likely to participate in a content generation [19], given that they can gain more reaction from the others. This explains why people enjoy social interactions by methods such as "liking" or "commenting" to draw connections with other users. However, online users participate in cyberbullying, believing that they cannot be identified [20]. With the freedom and anonymity, some has taken the advantage to express freely without thinking of consequences, and being offensive online. Cyberbully on Facebook had the highest rate [21]. Cyberbullying is a repetitive, intentional, and targeted action which creates an imbalance in power [22], and 20% - 40% of all youths have experienced once in their life time [23]. Users employ hate speech to gain popularity with minimal effort [24], common hate speech revolves around the themes of gender, religion, and disability [12].

Technology is widely accessible, and interactions happen instantaneously with some thinking that the act was a form of "having fun" [25]. Though the use of profanity is not directly linked to cyberbullying, however anonymity and swearing were

correlated [26]. Whittaker& Kowalski [27] noted that new means of cyberbullying has shifted from basic text communication to social networking sites because they allow more indirect and public attacks. This raises the question of how "public" attacks are carried out by perpetrators. Does it refer to stating something hateful publicly in social media, or openly attacking another user? With this in mind, it further motivates the study to include the consideration of "public" and "targeted" elements when analysing the differences in gender-linked profanity use. Then again, is there a link between gender and profanity?

The use of profanity seems to be the blurred line between hate speech and non-hate speech as studied by past hate detection researches [28-30]. These studies have found that social media text can be categorised into >2 groups (hate and non-hate) of speech. The categorisation of hate speech is also supported by the point of including targets of hate speech into the context. This means that people who use offensive words may not be targeting anyone, or people may not feel attacked upon witnessing the act. It was proven when Malmasi & Zampieri [31] attempted to distinguish profanity from hate speech, stressing that the use of swear words may not necessarily be targeted at another individual but used only for emphasis. Results of their study showed that statements could be used to express hate even without the use of profanity. In this sub-section, the article will go through past related studies to first understand profanity use, followed by its possible relation to gender.

We sometimes see people throwing insults at others with profanity, or even swear with no specific targets online. So, how did profanity come about? Why did people start using it? In 1901, Patrick [32] explored the psychology behind the use of profanity, focused on a type of profanity called "ejaculatory profanity". He first discussed how profanity originated from religious roots whereby people used to "exclaim oaths" that contained the names of religious figures or items. Later, he related the use of profanity to how animals would produce sounds that project alertness/security when they are threatened. To that, Patrick [32] concluded that in reaction to threatening situations, the use of profanity is a primitive and instinctive form of reaction that preserves the wellbeing of an individual, adding that emotion is not generated by using profanity but allayed by it. Being one of the earliest analyses, he submitted the impression that swearing is simply as a reflex when one's wellbeing is threatened, which a common observation is. For instance, some of us may recall a time a friend or family exclaiming "stupid (something)" when he/she accidentally bumped into "something" that has caused pain or in this context, stress. Note how the injured did not take time to formulate his/her response to the situation, but it occurred almost instantaneously afterwards. This is later proven by Stephens & Umland [33] where people use swear words to relieve or reduce experienced pain.

To further understand if swearing is an instinctive response and if our minds restrict ourselves to such exclamations, Turel & Qahri-Saremi [34] conducted a two-part study to comprehend (1) impulsive use of and (2) swearing on social networking sites. Their study showed that the use of profanity is not often intentional but instead, were results of preoccupation generated from the human cognitive-emotional system. To add, the authors claimed that swearing occurs when a person with weak cognitive system responds to stress. However, they also noted that finding only explained 16% of the variance in swearing on social networking sites. In a different article, Stephens & Zile [35] examined the relationship between emotional arousal, gender difference, and swearing fluency in a two-part study which in the first part, the identified that the more frequent people swear the more swear words they know while results showed that emotional arousal did increase swearing fluency.

Lastly, they found no relationship between gender and swearing fluency. To contrast, however, Jay [2] did find that the use of profanity costs the recipient of profanity emotional well-being, e.g. lowered self-confidence, and in some severe cases, lead to self-harm as described in Hinduja &Patchin's [36] findings. In another research, Feldman, Lian, Kosinski, & Stillwell [37] challenged the speculation that the use of profanity is associated with dishonesty and their results showed that profanity was found in honest language patterns instead. This re-emphasises the importance of context variables in the analysis of hate speech, which defines the implication of profanity use [38]. With regards to that, works by Silva et al. [39], and Teh, Cheng, & Chee [12] have put swear words into categories such as sexual orientation, and disabilities etc. based on the targets of hate identified. This, at least for the case of hate speech, has put profanity into different recognisable contexts.

Aside from knowing when and how swearing occurs, it is also important to recognise the perceptions people hold when they are in contact with profanity, whether intended for them or not. To answer that, DeFrank & Kahlbaugh [40] investigated the perception of profanity by observing paired conversations of different gender. They reported that the majority of participants use 11-15 profane words daily while being exposed to 6-10 profane words daily. It was also identified that "bitch" was considered the most offensive swear word. Moreover, they reported that profanity use did give less favourable impressions and reduced competence.

Interestingly, respondents were found to rate profane terms used as not profane but rated users of profanity with lower impression scores. The authors explained that commonly used swear words may not be considered offensive whereas rarely used profanity triggered "shock value" to observers, consequently appearing more offensive. Lastly, they noted that a combination of mixed genders triggered a bias where females appeared more offensive than males in conversations which they suspect is a result of expectations of gender roles in conversations.

About the gender behind of profanity, some (but limited) studies have found relationships between swear words and gender [8][41]. Thelwall's [41] results showed that the use of profanity was more prevalent among young American adults while there was no gender difference in profanity use in the UK whereas Bamman [8]did find that males are more frequent swearers than females. Note that this does not conflict with the findings (mentioned previously) of Stephens & Zile [35] as their focus was on "swearing fluency" and not "swearing frequency". It is also notable that what Bamman [8] and Thelwall [41] found were accompanied by the focus on gender-linked language rather than gender-linked use of profanity meaning the latter was only discovered as part of their analysis, not the main target of their assessment. In other words, there were no further explanations about why and how the observations surfaced. Hence, the following paragraphs will examine the relationship between gender and language instead. This also supports the aim of this study, which is to test if profanity-centred content aids in predicting gender as there is a lack of focus in this subject.

In 2001, Thomson & Murachver [10]researched language difference by gender by having participants predict the gender of

email authors. They found that females used more self-derogatory comments, compliments, apologies and subjective conjunctions while males were more likely to convey opinions and make insults. Also, results showed that people were sensitive to gender differences in language style and were able to accurately identify genders of the messages' authors even in the absence of gender-specific topics and physical indicators. In the following year, Colley & Todd [9] also studied language differences by gender through analysing emails sent out by their participants. Their results showed that females would often include warnings, used multiple question marks, and asked questions more than males would which they interpreted as markers of excitability. They also noted that both male and females disclosed about themselves to a recipient of the opposite sex than to the same-sex recipient. All in all, the study revealed that electronic discourse between opposite genders showed more intimacy or warmth as opposed to mails between same genders. Despite that, the authors did mention that results could be unnatural as participants were aware that the emails would be read for the analysis, leading to them have some sort of self-censorship (e.g. avoidance of impolite language).

Through a different spectacle, Park [11] used an open-vocabulary method to study gender-linked topics as well as assertiveness of different genders by analysing Facebook status updates. They found that female-linked topics often included intensive adverbs, and that these topics often related to social relationships and their associated emotions. As for males, they reported that the topics often involved sports and occupations and were more specific when referencing the topics such as by stating activities or objects involved. Furthermore, studies on gender-linked language have evolved from mapping gender schematics and gender identity salience to using predictive modelling to predict an author's gender [8] [42] [43]. This adaptation of predictive modelling in gender identification was driven by the anonymity of users. Gender-related information can be hidden through privacy settings. To overcome that, Burger, Henderson, Kim, & Zarrella [44]used profile descriptions and links to blog profiles to identify the genders of Twitters. Alternatively, Rao [43] developed a predictive model by combining several million n-gram features and found that females used more expressive phrases, whereas males used more affirmative words. In contrast, clustering method showed that gender identification of language is not limited to styles, stances, and personae as the context of language use can be generated from the language [8].

With the understanding gained from all past research, said situations had inspired this study by targeting the aspect of profanity use, linked together with the suspected effect of gender as an attempt to answer the problem of anonymity that often comes with social media. Ideally, the use of profanity, be it hateful or not, should help predict the gender of anonymous users of social media. Nevertheless, should the use of profanity support gender prediction, identification of actual cyberbullies may be improved in future research when combined with other language detection tools.

3. METHODOLOGY

3.1 Data Collection

A total number of 106,024 tweets was scrape using Twitter Archiver, with the references of list of keywords from various hate categories [9]. The "retweet" results were filtered out as they are results of duplication.Two thousand top used male and female English names were scraped from the US Social Security Administration website[45]. Using Python programming language, the lists of names were filtered. After removing of duplicates (i.e. same names used across different decades), both male (368 names) and female (446 names) name were finalized. These names were used as checklists for names that are present in the "screen name" and "full name" columns. Notably, there were 14 names shared by both name lists, thus considered as unisexual names.

3.2 Data Preprocessing

Duplicated tweets were removed. This ensures that the same tweet with the same terms will not be counted twice or more. The dataset was left with 102,464 unique tweets and is ready for the name extraction and gender labelling. The users' first names were used in name extraction. The task was carried out using Regular Expressions (regex) in Python to capture the first word of the name compare them against both female and male lists. If a match were found, the matched name would be labelled. 24,213 rows of tweets were finalized.

Tweets that contain regex, with "@" sign followed by characters, numerals, and underscores (as per Twitter username standards) (e.g. @sample_name99) was identified and labelled as a "targeted" tweet and set to 1 or 0 if otherwise. Finalized with 10,487 "public" tweets: 5,702 male tweets and 4,785 female tweets. And 10,437 "targeted" tweets: 6,984 male tweets and 3,453 female tweets. Since all datasets were balanced, the female and male counts are equal. Then, "count vectorizer" was used, with its "vocabulary" option set to all the keywords taken from [12]. By doing so, only relevant profanity in the tweets is counted; non-keyword terms in the tweets were ignored and not counted. As a result, two datasets (each representing a gender group) were produced, which contained information such as the most/lease-used term by each gender. These two datasets were then merged to form a new dataset called "keyword count" with "keyword", "count_m (male count)", "count_f (female count)" columns for comparison in the following step.

As this new dataset of keyword counts did not contain the hate category data, a function (written in Python) was created. It contained all the keywords used and acted as the "checker". Based on the origin of the keywords, the respective hate categories were reassigned into a new column called "group". Then, both the counts of male and female were added to form a new column called "total_usage" whereas the difference between the counts (male minus female; a positive number indicates higher male usage of term and vice versa) were recorded in "usage_difference". This information served as a first indicator of the size (frequency) of each profane term compared to other terms as well as the difference of the term's use between two genders.

In order to understand the differences between gender by hate category, the keywordCount dataset was grouped by the "group" column value for each term, and the frequencies were summed accordingly. With the "total_usage" column, the percentages for male and female counts were then calculated and stored in "perc_m (male percentage)" and "perc_f (female percentage)" as well as the difference between percentages in "perc_difference (percentage difference)". However, the difference between groups could not directly used for comparison due to their different sizes. Hence, each group's weight was calculated by taking its "total_usage" divided by the sum of all "total_usage" and that value was stored in the "groupWeight" column. After that, the "weighted_difference" column was formed by multiplying "usage_difference" and "groupWeight". This way, the comparison is fairer as it is relevant to the group's proportion in the dataset. The same processes were repeated in the public and targeted

datasets, following the assumption difference in profanity use in the presence/absence of a receiver was assumed in this analysis.

4. RESULT

4.1 Hate Category and Gender

To ease understanding and prevent confusions, some terms/matters is clarified:

-Combined tweets- the tweets which includes both public (no other user tagged) and targeted (at least one other user tagged).

-Keyword- the profanity term from list of keywords from Teh, Cheng, & Chee [13]where "keyword", "profanity", "term" and "profane term/word" are the same and were used interchangeably.

-Group - hate category/group where "category", "group", "hate category" and "hate group" are the same and were used interchangeably.

The full list of keywords counted were 77 but only the top 10 terms used by either gender will be analysed. Results observed from the combined tweets dataset was used as base comparison against public/targeted because it is a mix of both datasets which provides a more wholesome picture.

keyword	count	group
fuck	562	sexual orientation
hell	543	religion
shit	532	other
god	359	religion
fucking	339	sexual orientation
incompetent	306	disability
idiot	303	disability
racist	300	behaviour
crap	283	other
stupid	282	disability

Figure 1. Top 10 profanity used by males in combined tweets.

From figure 1, "fuck" was the most used term and "stupid" was the 10thmost used term where the frequency of "fuck" is twice of that of "stupid". In this figure alone, the hate group with the most terms is "disability" while the others have 2 occurrences except for the "behaviour" hate category. This could indicate that males are more likely to use profanity from the hate group "disability".

keyword	count	group
hell	697	religion
fuck	614	sexual orientation
shit	453	other
god	418	religion
queer	366	sexual orientation
fucked	358	other
fucking	319	sexual orientation
bitch	315	gender
crap	312	other
stupid	310	disability

Figure 2. Top 10 profanity used by females in combined tweets

Figure 2, the terms "hell" and "fuck" have switched positions as compared with figure 1 but both frequencies are higher than those in figure 1. It is also observed that "sexual orientation" and "other" hate categories have the highest frequencies of appearance where "gender" and "disability" each appearing only once. Also, there is another variation of "fuck" in this list, e.g.:"fucked". Other words that were not in the male list include "queer" and "bitch" while figure 1 contained words such as "incompetent", "idiot" and "racist".

Table 1. Distribution of profanity in combined tweets by hate category and gender.

Hate Category	Male	Female
Behaviour	829	705
	54%	46%
Class	386	408
	49%	51%
Disability	1842	1394
	57%	43%
Gender	954	1037
	48%	52%
Others	1646	1776
	48%	52%
Physical	1095	1147
	49%	51%
Race	70	55
	56%	44%
Religion	566	579
	49%	51%
Sexual Orientation	850	1137
	43%	57%

From table 1, shown not large differences between the two genders. However, it also noticeable that "disability" hate group has a 14% difference with males holding the higher percentage. A similar case is observed in the "sexual orientation" category but this time, females have the bigger percentage. Secondly, males have 13% more counts in the "race" category than females. The first two observations seem to be consistent with figure 1 and 2 where males favour the "disability" and females favour "sexual orientation".

Table 2. Distribution of profanity in public/targeted tweets by hate category and gender.

Hate Category	Male		Female	
	Public	Targeted	Public	Targeted
Behaviour	317	512	370	335
	21%	33%	24%	22%
Class	175	211	234	174
	22%	27%	29%	22%
Disability	672	1170	662	732
	21%	36%	20%	23%
Gender	509	445	648	389
	26%	22%	33%	20%
Others	752	894	1058	718
	22%	26%	31%	21%
Physical	515	580	742	405
	23%	26%	33%	18%
Race	49	21	45	10
	39%	17%	36%	8%
Religion	260	306	318	261
	23%	27%	28%	23%
Sexual Orientation	466	384	708	429
	23%	19%	36%	22%

From table 2, we can see that males are still scoring high in the "disability" category and this time, it is noticed that 36% of the profanity in that hate group is used when the tweet involves at least one receiver, observing from figure 1, the terms seem to be used for name-calling. In the "race" category, most of their tweets do not involve a receiver. As for "sexual orientation", females have the highest percentage across the distribution and in public tweets. If we refer to figure 2, it could imply that females are generally exclaiming publicly using the terms, not involving others in tweets.

	group	count_m	count_f	total_usage	usage_difference	perc_m	perc_f	perc_diff	groupWeight	weighted_diff
2	disability	2021	1542	3563	479	0.567	0.433	0.134	0.168	80.472
4	other	1995	1903	3898	92	0.512	0.488	0.024	0.184	16.928
0	behaviour	927	764	1691	163	0.548	0.452	0.096	0.080	13.040
5	physical	1006	987	1993	19	0.505	0.495	0.010	0.094	1.786
6	race	115	95	210	20	0.548	0.452	0.096	0.010	0.200
1	class	395	419	814	-24	0.485	0.515	-0.030	0.038	-0.912
7	religion	1215	1378	2593	-163	0.469	0.531	-0.062	0.122	-19.886
3	gender	1178	1387	2565	-209	0.459	0.541	-0.082	0.121	-25.289
8	sexual orientation	1843	2041	3884	-198	0.475	0.525	-0.050	0.183	-36.234

Figure 3. Male and female profanity count in combined tweets sorted by weighted difference in descending order.

In figure 3, at the weighted difference column, "disability" and "sexual orientation" hate categories are at 2 extreme ends, which is still consistent with previous observation. To clarify, the "weighted difference" value does not represent the actual difference between gender classes but is instead, a normalised value for better comparison across the different hate categories, taking into account the different sizes of "total_usage" of profanity in each category. For example, although there is an 1% difference in "physical" category, its weighted difference is higher than the 9.6% difference in "race". Hence, relative to category's size in the distribution of hate categories, the difference observed in the "physical" class is more meaningful despite its low percentile difference. Additionally, it is observed that "disability" has the highest absolute value across all weighted differences (more than double of the other extreme – "sexual orientation") which shows that males have more often used terms from that category than females have.

Notice that the "weighted_diff" and "usage_difference" values contain negatives. It is because they were obtained by subtracting the female counts from the male counts. Thus, when subtracting from a lower male count, a negative value is produced. Following that logic, the greater the weighted difference value, the more profanity in that hate category was used by males than females, and vice versa. On top of that, it is observed that "physical", "race", and "class" hate category have weight difference values that are close to 0. This means that there are no major differences in profanity use from these hate categories between the two gender classes. With that said, a classification model might have difficulty classifying an author's gender if the tweet contained profanity from any of these three classes.

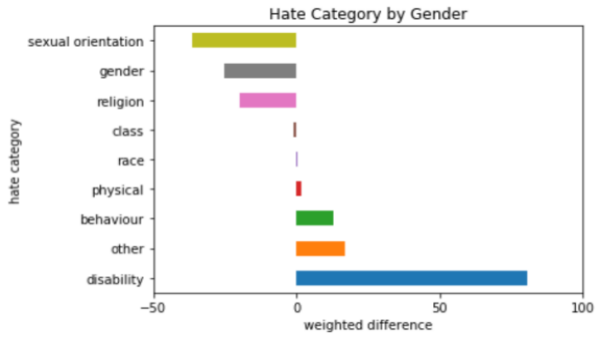

Figure 4. Horizontal bar chart of weighted difference by hate category.

Figure 4 is a visual representation of figure 3. Essentially, bars left of 0 are represented by higher female use whereas bars to the right of 0 represents more dominant male use. From the bar chart, we observe that the "disability" has the longest bar, largely greater than other values. This great difference means that males use profanity from the hate category much more than females do in tweets. Also, we can see in figure 4 that females are likely to switch between "sexual orientation", "gender", and "religion" hate categories when using profanity because there are no major differences between the distributions. On the other hand, males are more in favour of "disability" group of profanity over "behaviour" and "other".

keyword	count	group
hell	246	religion
crap	167	other
racist	165	behaviour
delusional	153	disability
fuck	149	sexual orientation
incompetent	145	disability
stupid	141	disability
shit	139	other
idiot	126	disability
god	116	religion

Figure 5. Top 10 profanity used by male in targeted tweets.

Compare figure 5 to figure 1, "fuck" has fell from 1st place to 5th place while "hell" remained in the one of the top 2 positions. "Crap" was previously ranked 9th most used term but is now ranked 2nd for the case of targeted tweets. "God" has fell from 4th to last, and "shit" from 3rd to 8th as most used profanity by males.

In addition, a new term "delusional" was introduced in the top 10 list which is also categorised under the "disability" category. In figure 5, the "disability" hate group still holds more positions as most used profanity by males and has increased from 3 to 4. On the other hand, both "religion" and "other" hate group each still hold 2 of the same terms in the list. As a result, figure 5 supports the observation from figure 1 and figure 4 regarding the use of profanity about disability.

keyword	count	group
hell	304	religion
crap	179	other
fuck	176	sexual orientation
god	169	religion
delusional	142	disability
idiot	134	disability
shit	134	other
stupid	130	disability
incompetent	125	disability
queer	119	sexual orientation

Figure 6. Top 10 profanity used by female in targeted tweets.

Compare figure 6 and 2, "fuck" has fallen to 3rd spot while "crap" has moved up to 2nd from 9th most used profanity in figure 2. "Hell" remains the top term used for females even in targeted tweets. In figure 6, the term "delusional" was also newly included to the "top 10" list. Interestingly, we see that there are more keywords from the "disability" group in figure 6, all of which can be found in figure 5. Also, other variations of "fuck" are no longer found in figure 6. On the other hand, "queer" was previously in 5th place in figure 2 but now has fallen to last in list. Overall, it is observed that females have used more terms about disability when their tweets included other users.

group	count_m	count_f	total_usage	usage_difference	perc_m	perc_f	perc_diff	groupWeight	weighted_difference
disability	930	773	1703	157	0.546	0.454	0.092	0.207	32.499
behaviour	454	363	817	91	0.556	0.444	0.112	0.099	9.009
other	754	709	1463	45	0.515	0.485	0.030	0.178	8.010
physical	432	360	792	72	0.545	0.455	0.090	0.096	6.912
race	28	11	39	17	0.718	0.282	0.436	0.005	0.085
class	167	176	343	-9	0.487	0.513	-0.026	0.042	-0.378
gender	371	443	814	-72	0.456	0.544	-0.088	0.099	-7.128
religion	503	587	1090	-84	0.461	0.539	-0.078	0.132	-11.088
sexual orientation	524	646	1170	-122	0.448	0.552	-0.104	0.142	-17.324

Figure 7. Male and female profanity count in targeted tweets sorted by weighted difference in descending order.

Compare figure 7 and 3, "behaviour" has taken the 2nd spot while "other" is in the 3rd. Also, "gender" has moved to 7th while "religion" moved to 8th. An upward movement of hate category in this figure means 2 possible observations: (1) males were found to use more profanity from that category or (2) female were found to be using lesser profanity from that category. A downward movement mean the reverse of the case of an upward movement. Take the downward movement of religion for example, it means that females have used more terms from the "religion" group, or males have used less terms from the "religion" group. Besides that, we see that "weighted_difference" value of the "physical" category has shifted further from zero in favour of male usage of terms in that category.

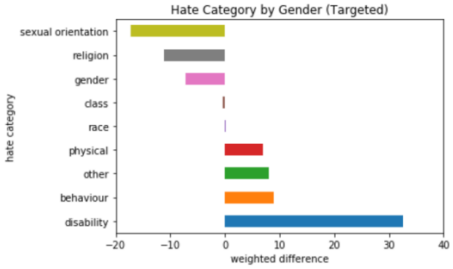

Figure 8. Horizontal bar chart of weighted difference by hate category in targeted tweets.

Consistent with observation from figure 7, "physical" now has a bar with a size (length) that is almost close to those of "behaviour" and "other" hate groups. Also, comparing with figure 4, notice the change of positions in "religion", "gender", "other" and "behaviour" groups as mentioned in the last paragraph. On top of that, no major changes are observed in "class" and "race" categories but since "physical" is in favour of male use, that means one less gender-neutral category of profanity terms for a possible profanity-based gender classifier. Lastly, the "disability" hate group still held the highest weighted difference (in favour of male usage), more than double of most hate categories.

keyword	count	group
fuck	417	sexual orientation
shit	398	other
hell	298	religion
fucking	267	sexual orientation
god	256	religion
ass	199	physical
fucked	198	other
gay	173	sexual orientation
idiot	168	disability
bull	162	gender

Figure 9. Top 10 profanity used by males in public tweets.

Compared to figure 1, with "idiot" being the remaining term from "disability" group, male use of profanity from that category seems less likely to appear in public tweets. In turn, new terms such as "bull", "ass", "gay" and "fucked" have made the top 10 list, introducing other hate categories (i.e. "physical" and "gender") into the mix. Thus, in male public tweets, their top 10 choice of profanity covers 6 hate groups instead 4 in combined tweets (figure 1). On top of that, "Fuck" remains the top term in the list as in Figure 9.

Figure 10 is largely similar to figure 2 in terms of keywords involved except for "lesbian" and "ass". The inclusion of "lesbian" increased the count of "sexual orientation" keywords to 4 instead of 3 in figure 2 which seems to suggest that profanities in "sexual orientation" is prevalent or possibly preferred in female public tweets. "Ass" has replaced "crap", and hence added "physical" to the hate categories involved in female public tweets.

keyword	count	group
fuck	438	sexual orientation
hell	393	religion
shit	319	other
fucked	279	other
fucking	256	sexual orientation
god	249	religion
queer	247	sexual orientation
bitch	235	gender
ass	208	physical
lesbian	188	sexual orientation

Figure 10. Top 10 profanity use by females in public tweets.

group	count_m	count_f	total_usage	usage_difference	perc_m	perc_f	perc_diff	groupWeight	weighted_difference
disability	957	769	1726	188	0.554	0.446	0.108	0.132	24.816
other	1255	1194	2449	61	0.512	0.488	0.024	0.187	11.407
behaviour	487	401	888	86	0.548	0.452	0.096	0.068	5.848
physical	632	627	1259	5	0.502	0.498	0.004	0.096	0.480
class	254	243	497	11	0.511	0.489	0.022	0.038	0.418
race	111	84	195	27	0.569	0.431	0.138	0.015	0.405
gender	896	944	1840	-48	0.487	0.513	-0.026	0.140	-6.720
religion	733	791	1524	-58	0.481	0.519	-0.038	0.116	-6.728
sexual orientation	1332	1395	2727	-63	0.488	0.512	-0.024	0.208	-13.104

Figure 11. Male and female profanity count in public tweets sorted by weighted difference in descending order.

Results in figure 11 mostly resemble those of figure 3 except for the actual counts and weighted difference, both of which are affected by the size of the dataset. One difference, however, is that weighted differences of "gender" and "religion" are the same while there was approximately a 6% difference between the 2 in figure 3. Regardless, profanity in both hate categories remain strongly employed groups of hate in female public tweets.

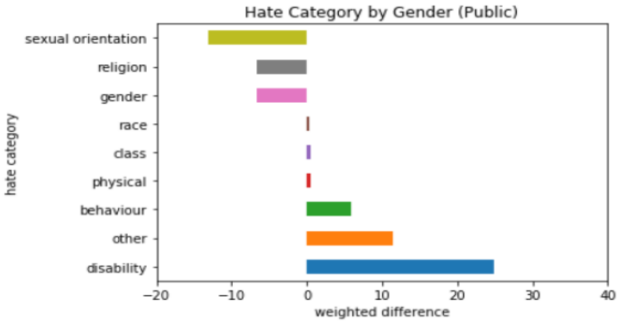

Figure 13. Horizontal bar chart of weighted difference by hate category in public tweets.

As mentioned under figure 13, "gender" and "religion" groups now share the same bar size, showing that when in public tweets, females often to used profane terms from either category when not using keywords from the "sexual orientation" category. In contrast to figure 8, the bar size of "physical" group is identical to that of figure 4, implying three gender-neutral hate categories in the case of public tweets.

4.2 Gender Classification using Profanity

Table 3.Classification results on combined tweets dataset

Combined

	gen	prec	recall	f1	accuracy	mean cv	auc
logReg	F	0.58	0.64	0.61	0.59	0.58	0.61
	M	0.6	0.53	0.56			
multiNB	F	0.58	0.63	0.61	0.59	0.58	0.62
	M	0.6	0.55	0.57			
SVM	F	0.59	0.67	0.63	0.6	0.57	0.59
	M	0.62	0.54	0.58			

Table 4. Classification results on public tweets dataset

Public							
	gen	prec	recall	f1	accuracy	mean cv	auc
logReg	F	0.57	0.6	0.58	0.57	0.58	0.61
	M	0.58	0.54	0.56			
multiNB	F	0.57	0.6	0.58	0.57	0.58	0.61
	M	0.57	0.54	0.56			
SVM	F	0.57	0.59	0.58	0.57	0.58	0.58
	M	0.58	0.55	0.56			

Table 5. Classification results on targeted tweets dataset

Targeted							
	gen	prec	recall	f1	accuracy	mean cv	auc
logReg	F	0.54	0.46	0.5	0.53	0.56	0.58
	M	0.53	0.6	0.56			
multiNB	F	0.54	0.46	0.5	0.54	0.55	0.58
	M	0.53	0.62	0.57			
SVM	F	0.68	0.1	0.18	0.53	0.53	0.53
	M	0.51	0.95	0.67			

5. CONCLUSION

Overall, it is observed that classification models for the combined and public datasets performed slightly better than a blind guess, with accuracy and mean cross validation (CV) scores averaged at 0.58, and AUC score averaged at 0.6. In the case of targeted tweets dataset, the accuracy averaged at 0.53, mean CV averaged at 0.55, and AUC averaged at 0.56, all of which are within 4-5% difference from those of combination and public tweets. Despite that, it is also observed that recall rates for male targeted tweets were generally higher than those in public or combined datasets. Unfortunately, we also notice that SVM performed especially bad in the targeted tweets dataset by looking at the accuracy, mean CV and AUC, along with a highly disproportionate recall rate. Though recall or precision rates are not distributed between classes, the observation of SVM's recall rate in table 5 indicates that the model favored the male class when learning profanity use patterns.

Precision is the rate of getting a correct classification out of all classifications made. Recall refers to the rate of getting a correct classification out of all the times a classification result was made to a specific target. Here, out of 10 times of classifying a tweet that is labelled "male", 6 tweets were correctly classified as male, the male recall rate would be 0.6. In table 3 and table 4, precision rate generally balanced and averaged at 0.58 whereas for recall, female scores are higher than those of males. In table 4, both LR and MNB perform equally well in terms of accuracy (0.57) and AUC (0.61) in classifying gender while in table 3, SVM may have 1% advantage in accuracy compared to LR and MNB but lose out on mean CV and AUC to the two. By comparing the other two models, MNB has a slight 1% advantage in AUC over LR which makes it the better model.

In table 5, the recall rate for males seem to higher than those table 3 and table 4. This could possibly be related reduction on one gender-neutral hate category as mentioned under figure 8. Besides that, there does not seem to be any major differences among classification results of the different datasets, with targeted dataset results being slightly lesser with the other two. Regardless, we conclude that separating the datasets into "public" and "targeted" did not yield any useful results to show difference. In other words, when analysing profanity use by gender, the presence/absence of a target is not likely to affect the classification results. Furthermore, with an average accuracy, mean CV, and AUC at approximately 0.6, though profanity may not be a strong feature to classify, there is still a relationship between profanity use and gender.

6. REFERENCES

[1] Ping, C., 103 Crazy Social Media Statistics to Kick off 2014, 2013. http://thesocialskinny.com/103-crazy-social-media-statistics-to-kick-off-2014/ (accessed June 20, 2019).

[2] Jay, T., Do Offensive Words Harm People? 2009*Psychol. Public Policy*, Law. 15. 81–101. DOI= http://d037/a0015646oi:10.1.

[3] Kaya, T. and Bicen, H. 2016. The effects of social media on students' behaviors; Facebook as a case study, *Comp. Human Behavour* 59. 374–379. DOI = http://doi:10.1016/j.chb.2016.02.036.

[4] Gruzd, A., Wellman, B. and Takhteyev, Y. Imagining Twitter as an Imagined Community, *Am. Behav. Sci*. 55 (2011) 1294–1318. DOI = http://doi:10.1177/0002764211409378.

[5] Laub, Z. 2019. Hate Speech on Social Media: Global Comparisons, Counc. Foreign Relations. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons (accessed June 18, 2019).

[6] Ramstack,T. 2019.Social Media Giants Questioned on Encouraging Hate Crimes, Well News. (2019). https://www.thewellnews.com/social-media-giants-questioned-on-encouraging-hate-crimes/ (accessed June 18, 2019).

[7] Rawhide, T. 2018. Teen Cyberbullying and Social Media Use on the Rise [INFOGRAPHIC], Rawhide. https://www.rawhide.org/blog/infographics/teen-cyberbullying-and-social-media-use-on-the-rise/ (accessed June 18, 2019).

[8] Bamman, D., Eisenstein, J., and Schnoebelen, T. 2014 Gender identity and lexical variation in social media, *J. Socioling*. 18. 135–160. DOI = http://doi:10.1111/josl.12080.

[9] Colley, A. and Todd, Z. 2002Gender-linked differences in the style and content of e-mails to friends, *J. Lang. Soc. Psychol*. 21. 380-392 DOI = http://doi:10.1177/026192702237955.

[10] Thomson, R. and Murachver, T.2001. Predicting gender from electronic discourse.*Br. J. Soc. Psychol.* 40. 193–208. DOI = http://doi:10.1348/014466601164812.

[11] Park, G., Yaden,D.B., Schwartz, H.A., Kern,M.L., Eichstaedt, J.C., Kosinski, M., Stillwell,D., Ungar, L.H., and Seligman, M.E.P. 2016 Women are Warmer but No Less Assertive than Men: Gender and Language on Facebook, PLoS One. 11. 1–26. DOI = http:// doi:10.1371/journal.pone.0155885.

[12] Teh,P.L., Cheng,C.-B., and Chee,W.M. 2018. Identifying and Categorising Profane Words in Hate Speech, *Proc. 2nd*

Int. Conf. Comput. Data Anal. - ICCDA 2018. 65–69. DOI = http://doi:10.1145/3193077.3193078.

[13] Whiting, A. and Williams, D. 2013. Why people use social media: a uses and gratifications approach, *QualitativeMarketResearch.* 16. 362–369. DOI = http://doi:10.1108/QMR-06-2013-0041.

[14] Teh, P.L. Huah, L.,and Si, Y. 2014. The Intention to Share and Re-Shared among the Young Adults towards a Posting at Social Networking Sites, *New Perspect. Inf. Syst. Technol.* 1. 13–21. DOI = https://doi.org/10.1007/978-3-319-05951-8_2.

[15] Sander, T., Teh,P. L. and Sloka, B.2017. Your Social Network Profile Reveals You, *International Journal Web Information System*. 13 (2017) 14–24. DOI = https://doi.org/10.1108/IJWIS-06-2016-0029.

[16] Khan,M.L. 2017. Social media engagement: What motivates user participation and consumption on YouTube?, *Comput. Human Behav.* 66 (2017) 236–247. DOI = http://doi:10.1016/j.chb.2016.09.024.

[17] Elmore, K. 2010.What is Social Media?, *Bus. Prem. Collect.* (2010) 12–13. DOI = http://doi:10.1016/B978-1-59749-986-6.00001-1.

[18] Lipscomb, J.2010. What is social media?, Bus. Prem. Collect 74. https://www.talkwalker.com/blog/what-is-social-media-intelligence.(accessed June 20, 2019).

[19] Toriumi, F., Yamamoto, H. and Okada,I. 2012.Why do people use social media? Agent-based simulation and population dynamics analysis of the evolution of cooperation in social media, *Proc. - 2012 IEEE/WIC/ACM Int. Conf. Intell. Agent Technol. IAT 2012*. 2 (2012) 43–50. DOI = http://doi:10.1109/WI-IAT.2012.191.

[20] Carter, M.A. 2013.Protecting Oneself from Cyber Bullying on Social Media Sites – a Study of Undergraduate Students, *Procedia - Soc. Behav.* Sci. 93 (2013) 1229–1235. DOI = http://doi:10.1016/j.sbspro.2013.10.020.

[21] Carter,M.A. 2013.Third Party Observers Witnessing Cyber Bullying on Social Media Sites, *Procedia - Soc. Behav. Sci.* 84 (2013) 1296–1309. DOI = http://doi:10.1016/j.sbspro.2013.06.747.

[22] Ong, R. 2015.Cyber-bullying and young people : How Hong Kong keeps the new playground safe, *Comput. Law Secur. Rev. Int. J. Technol. Law Pract.* 31 (2015) 668–678. DOI =http://doi:10.1016/j.clsr.2015.07.005.

[23] Tokunaga, R.S. 2010.Following you home from school: A critical review and synthesis of research on cyberbullying victimization, *Comput. Human Behav.* 26 277–287. DOI = http://doi:10.1016/J.CHB.2009.11.014.

[24] Chetty, N. and Alathur, S. 2018.Hate speech review in the context of online social networks, Aggress. *Violent Behav.* 40 (2018) 108–118. DOI = http:// doi:10.1016/j.avb.2018.05.003.

[25] Sabella, R.A., Patchin, J.W. and Hinduja,S. 2013. Cyberbullying myths and realities, *Comput. Human Behav.* 29 (2013) 2703–2711. DOI= http://doi:10.1016/j.chb.2013.06.040.

[26] Kwon, K.H. and Cho,D. 2017. Swearing Effects on Citizen-to-Citizen Commenting Online: A Large-Scale Exploration of Political Versus Nonpolitical Online News Sites, *Soc. Sci. Comput. Rev.* 35 (2017) 84–102. DOI = http://doi:10.1177/0894439315602664.

[27] Whittaker,E. and Kowalski, R.M. 2015. Cyberbullying Via Social Media, *J. Sch. Violence*. 14 (2015) 11–29. DOI = http://doi:10.1080/15388220.2014.949377.

[28] Magu,R., Joshi, K. and Luo, J.2017. Detecting the Hate Code on Social Media, (2017).*In the Preceeding of the Eleventh International AAAI Conference on Weblogs and Social Media.* 2017, 608-612.

[29] Gitari, N.D., Zuping, Z., Damien, H., and Long,J. A. 2015. lexicon-based approach for hate speech detection, Int. J. Multimed. *Ubiquitous Eng.* 10 (2015) 215–230. DOI = http://doi:10.14257/ijmue.2015.10.4.21.

[30] Malmasi, S. and Zampieri,M. 2017. Detecting Hate Speech in Social Media, (2017).*RANLP 2017*.DOI = http://doi:10.26615/978-954-452-049-6_062.

[31] Malmasi,S. and Zampieri,M. 2018.Challenges in discriminating profanity from hate speech, *J. Exp. Theor.* Artif. Intell. 30 (2018) 187–202. DOI = http://doi:10.1080/0952813X.2017.1409284.

[32] Patrick, G.T.W. 1901. The psychology of profanity, *Psychol. Rev*. 8 (1901) 113–127. DOI = http://doi:10.1037/h0074772.

[33] Stephens, R. and Umland, C. 2011.Swearing as a response to pain - Effect of daily swearing frequency, J. Pain. 12.1274-1281.DOI=http://doi:10.1016/j.jpain.2011.09.004.

[34] Turel, O. and Qahri-Saremi, H. 2018.Explaining unplanned online media behaviors: Dual system theory models of impulsive use and swearing on social networking sites, *New Media Soc*. 20 (2018) 3050–3067. DOI = http://doi:10.1177/1461444817740755.

[35] Stephens, R. and Zile,A. 2017.Does Emotional Arousal Influence Swearing Fluency?, *J. Psycholinguist. Res.* 46 983–995. DOI = http://doi:10.1007/s10936-016-9473-8.

[36] Hinduja, S., and Patchin,J.W. 2008.Offline Consequences of Online Victimization Offline Consequences of Online Victimization : School Violence and Delinquency, 8220. DOI = http://doi:10.1300/J202v06n03.

[37] Feldman, G., Lian, H., Kosinski, M., and Stillwell, D. 2017. Frankly, We Do Give a Damn: The Relationship Between Profanity and Honesty, *Soc. Psychol. Personal*. Sci. 8 (2017) 816–826. DOI = http://doi:10.1177/1948550616681055.

[38] Wang, W., Chen, L., and Thirunarayan, K., and Sheth,A.P. 2014. Cursing in English on twitter, *Proc. 17th ACM Conf. Comput. Support. Coop.* CSCW '14. (2014) 415–425. DOI = http://doi:10.1145/2531602.2531734.

[39] Silva, L., Mondal, M., Correa, D., Benevenuto, F., and Weber,I. 2016. Analyzing the Targets of Hate in Online Social Media, http://arxiv.org/abs/1603.07709 (accessed October 8, 2018).

[40] DeFrank, M. and Kahlbaugh, P. 2019. Language Choice Matters: When Profanity Affects How People Are Judged, *J. Lang. Soc. Psychol*. 38 (2019) 126–141. DOI = http://doi:10.1177/0261927X18758143.

[41] Thelwall,M. 2008.Fk yea I swear: cursing and gender in MySpace, *Corpora.* 3 (2008) 83–107. DOI = http://doi:10.3366/E1749503208000087.

[42] Palomares,N.A. 2004. Gender Schematicity, Gender Identity Salience, and Gender-Linked Language Use, 30 (2004) 556–

588. DOI=https://doi.org/10.1111/j.1468-2958.2004.tb00745.x

[43] Rao, D., Yarowsky, D., Shreevats, A., and Gupta,M. 2010. Classifying latent user attributes in twitter, Proc. 2nd Int. Work. Search Min. User-Generated Contents - SMUC '10. 37. DOI = http://doi:10.1145/1871985.1871993.

[44] Burger, J.D., Henderson,J., Kim, G., and Zarrella, G. 2011.Discriminating Gender on Twitter, Assoc. Comput. Linguist. 146 (2011) 1301–1309. DOI = http://doi:10.1007/s00256-005-0933-8.

[45] Social Security Administration, Popular Baby Names by Decade, U.S. Taxpay. Expens. (n.d.). https://www.ssa.gov/OACT/babynames/ (accessed June 20, 2019).