Identifying and Categorising Profane Words in Hate Speech

Phoey Lee Teh
Sunway University
Bandar Sunway
47500, Malaysia
phoeyleet@sunway.edu.my

Chi-Bin Cheng
Tamkang Univerisity
P.O. Box 1212
Tamsui, Taiwan.
cbcheng@mail.tku.edu.tw

Weng Mun Chee
Sunway University
Bandar Sunway
Selangor, Malaysia
cheewmdan@gmail.com

ABSTRACT
This study attempts to explore the different types of Hate Speech appearing in social media by identifying profane words used in hate speech. This study also compares the profane words used in different generations to assist in identifying the user's profile. Five-hundred (500) comments posted on YouTube on the abusive topics were collected. Profane words are classified into eight different types of hate speech. The finding shows 35% of profane words found in our sample are words related to sexual orientation. Comparison of the terms between 1970 and 2017 also show a high percentage of profane words are sexual orientation. Though the results are found based on only 500 comments collected from YouTube link in the current study, they are useful in establishing the list of profane words which will serve as the base for automatic hate speech identification in our future study. The originality of this research is the development of a training list of profane words for each category and comparison of the type of the words used in 1970 century with today's social media platform.

CCS Concepts
• Information systems → Information Systems Applications • Human-centered Computer → Collaborative and Social Computing • World Wide Web → Web Application Social Networks • Web Application

Keywords
Hate Speech, Comments, Social Media, Classification, Profane Words

1. INTRODUCTION
Hate Speech or hate expression is commonly referred to as speech that contains abusive, insulting, intimidating, harassing or which incites violence, hatred or discrimination statement [11]. The internet has opened up the opportunity for freedom of speech, and many users today have no hesitation to express their view in the cyber world. Message that have been posted online, either intentionally or unintentionally in expressing hate towards others may cause potential harm to the victim. The effects of hate speech towards the victim are that the victim will develop psychological and pathophysiological symptoms similar to post-traumatic stress disorder (PTSD) which are panic, fear, anxiety, nightmares, intrusive thoughts of intimidation and denigration [9].

The popularity of online social media, such as Facebook, Twitter, Instagram and Youtube, boost the communication and information sharing between strangers; however, at the same time, they also become a hotbed for hate speech to breed. These hate speeches can not only harm individual victims but also create impacts to society, e.g. raising hostility between ethnic groups, or even leading to terrorist attacks, etc. To prevent the undesired impacts from hate speech, lately, in the year 2017, Germany has set the law to enforce hate speech on social media. The fine can even go up to 57 million dollars in the case if the social media fails to remove 70 percent of online hate speech within 24 hours [6]. However, the challenges in identifying and detecting which statement contain a hatred component in the speech in an online platform is not an easy task. The tremendous amount of messages generated continuously every moment on social media make it impossible to identify the hate speech manually, and thus make the automatic detection technique an ideal solution. Nevertheless, it is still difficult for the machine to detect a hate speech due to the intrinsic nature of ambiguity, incomplete and polysemy of natural language. Lexicons of negative words are necessary resources in extracting features of hate speech based on the assumption that hateful messages usually contain specifically harmful or profane words. The identification of the list of profane words contained in hate speech is thus helpful for the automatic detection of hate speech.

Earliest studies on profanity in communication disorders had focused on the usage of profane words in conversational speech [2,12]. The study by [2] initially intended to discover what college students talk about in their normal conversation, and found that 8.06 percent of the words used in conversations by college students related to sexual, and excretory profanities. However, the profane words consisting of 8.06 percent of Cameron's [2] vocabulary formed only 0.14 percent of the vocabulary in an earlier study by [12]. [2] argued that such discrepancy is a result of biased sampling or less representative vocabulary. To justify the argument of [2,12] investigated the use of profanity in conversational speech based on a sample from a college student population. Their result shows 7.44% of the collected words are profane. This ratio is close to the one (i.e. 8.06%) reported by [2]; however, some of the profane words listed [2] did not appear in the list of [12] and vice versa. This result implies that the profane words frequently used in conversation may vary in a different population or user groups. Thus, it is necessary to perform an analysis on the usage of profanity for the target user group if the lexicon approach is adopted for hate speech detection. Meanwhile, previous studies of profane words in the conversational speech were performed more than four decades ago, and the words used by people in the conversations must have evolved over the time. It is worthwhile to investigate the changes of profane words used by people in conversational speech. Such information can assist identifying the user's profile since

This is a preprint of an article to be published in 2018 Association for Computing Machinery. Phoey Lee Teh, Chi-Bin Cheng, Weng Mun Chee (2018) Identifying and Categorising Profane Words in Hate Speech, Proceedings of the 2nd International Conference on Compute and Data Analysis, Pages 65-69, https://dl.acm.org/citation.cfm?doid=3193077.3193078

people of different ages are likely to use different words in their conversations. The category of hate speech that a profane word belongs to is also useful information for hate speech detection. For example, the chance a profane word appears in the race hate category may be different from that of the word appearing in gender hate group. With such information, we can evaluate the probability that the speech is a hate or even the hate group it belongs to.

To achieve the goals of identifying profane words used in hate speech, this study collects user comments from Youtube and employs the corpus analysis and comparison tool, WMatrix, to parse the profane words from the comments. Wmatrix also identifies profane words that affect hate speech intensity. The current study serves as a pilot research to our future work of automatic hate speech detection by machine learning techniques. The results of this study will be used to perform an initial screen of profane words from Youtube comments. Findings of this study provide useful references for us to extract features from the comments for machine learning classification in our next stage study.

The remainder of this paper is organized as follows. Section 2 discusses methodologies for hate speech detection especially approaches related to the use of lexicons, and reviews previous studies on profanity in conversational speech. This followed by the analytic process presented in Section 3 which describes the use of WMatrix to identify and categorise profane words in hate speech. All the findings from the empirical study are presented and discussed in Section 4, and finally, the paper is concluded by Section 5.

2. LITERATURE REVIEW

In this 20 century, not only profane words are used in normal conversation, but it has also been used on Internet, social media particularly. Profanity is socially offensive language, which also calls bad language, vulgar language, or wrong choice of words or expletives. It may describe the behaviour of a person who is profoundly offensive or shows a lack of respect for others. Merriam-Webster has defined that intense hostility and aversion usually derive from fear, anger or sense of injury.

Researchers from [14] has described the status of hate speech in the different country. For instance, in Netherlands, it is a criminal offense to give expression insulting to groups or a person deliberately. Australia prohibits speech that offends, insults, humiliates or intimidates individual or groups. Britain bans abusive, offensive and threatening speech. Germany goes further in banning speech that violates the dignity of or maliciously degrades or defames a group. In recent year, they even set the rule to the company to delete hate speech from social media platform or else fine them with high cost [6]

Internet has opened up the opportunity for freedom of speech. Adolescent today has no hesitation to express their view in the cyber world. Messages that have been posted online, either intentionally or unintentionally in expressing hate towards others may cause potential harm to the victim [11] [5]. The study by [18] has also pinpointed that people use curse to show their strong emotion and cursing is so harmful to others when it is a form of insults such as name calling, harassment, hate speech and obscene telephone call. The effects of Hate Speech towards the victim is that the victim will develop psychological and pathophysiological symptoms similar to post-traumatic stress disorder (PTSD) which are panic, fear, anxiety, nightmares, intrusive thoughts of intimidation and denigration [9].

There are several approaches to identify hate targets, and it is always not an easy task. Based on three main thematic areas of race, nationality and religion, [7] create a model classifier that uses sentiment analysis techniques in particular subjectivity detection to not only detect that a given sentence is subjective but also to identify and rate the polarity. There are several examples of hate targets. As referring to [16], these categories of hate targets consists of Hate Speech with example of words that are classified of Race, Behavior, Physical, Sexual orientation, Class, Gender, Ethnicity, Disability, Religion, etc.

To catch bad language and remove a post is not an easy task. The study of [13] pointed out that to catch and remove bad language is a profoundly difficult task. They explained that abusive language might be very grammatical fluent, abusiveness can be cross sentence boundaries, and also appear in sarcastic comments. According to their study, most of the relevant works mainly focused on tackling the specific aspects of abusive language, and failed to detect those boundaries.

Besides [7] on classifying the Hate target types, there are also other detection methods have been proposed by different researchers. For instance, [13] included the use of annotation instructions, which detects whether a statement contains Hate Speech, Derogatory Language or Profanity words. As for [8] using the labelled image and the correlate between the features and cyberbullying and cyber aggression, on it liking behavior, the frequency of comments, following behavior as its features. Other than this, [4] pinpointed that lexical detection methods tend to have low prevision because they classify all messages containing particular terms as hate speech. There is also some other methods that have been proposed based on determining the kind of hate target. For instance, [17] proposed a variety of hate categories to distinguish the type of hate features using five distinct human annotators and defined a taxonomy on Italian public pages. Besides that, the study of [10] has identified another way of detecting certain categories of hate targets that are harder to identify. Whereby the community will use a particular set of code words to represent actual words. For example, Google to refer the Black, Yahoo to refer to Mexican, Skype to Jew, Bing to Chinese, Skittle to Muslim and Butterfly to Gay. Such word which at last able to detect those hate speech on racism category has been successfully able to project the hate content problem on Twitter into a classification problem which has never discussed before.

Table 1: Code Words to represent Actual Words

Code Words	Actual Words
Google	Black
Yahoo	Mexican
Skype	Jew
Bing	Chinese
Butterfly	Gay

This is a preprint of an article to be published in 2018 Association for Computing Machinery. Phoey Lee Teh, Chi-Bin Cheng, Weng Mun Chee (2018) Identifying and Categorising Profane Words in Hate Speech, Proceedings of the 2nd International Conference on Compute and Data Analysis, Pages 65-69, https://dl.acm.org/citation.cfm?doid=3193077.3193078

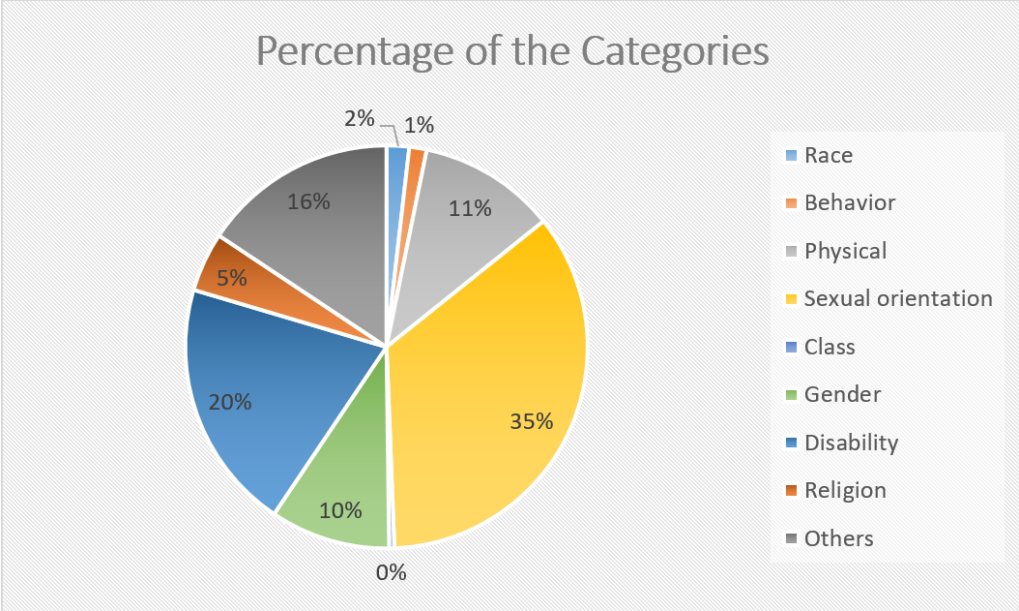

Figure 1: Percentage of the Hate Categories

Table 2: Categories of Profane Words used in Hate Speech

CATEGORY	PERCENTAGE	TERMS
Sexual Orientation	35.10%	gay, gays, lesbian, fag, faggot, faggots, faggot club, queer, fuck, fucking, fuckin, cocksucker
Disability	20.14%	retard, idiot, moron, dumbass, stupid, incompetent, delusional, douchebag, fucktard, dumbfuck, stupid trump
Gender	9.65%	cunt, cunts, bitch, bitching, bitches, bitching, pussy, dick, dicks, cock, dogs, dog, bull
Religion	4.76%	islam, islamic, jesus, god, devil, hell, god king
Race	7.82%	nigger, nigga, niggas, niggers, sandnigger
Behaviour	1.4%	racist, racists, islamophobia, rapist, pissedr
Class	0.42%	bastard, bastards, sucker
Ohers	15.67%	rap, bullcrap, piece, shithead, shit, shithead, damn, damnit, fucker, motherfucker, motherfucking

3. METHODOLOGY

To understand the categories of hate speech, further exploration to understand the classification of hate targets is vital. The reason for this step is to find the different categories so that to determine the profane words groups use in today's comment.

3.1 Data Collection

Firstly, comments that posted under Youtube clips and are considered relating to abusive topics are collected. In this study, 500 such comments are reviewed and identified manually. The corpus analysis tool, Wmatrix [15], is adopted to process the collected comments to extract key words that are relevant to hate speech. The analysis result contains 6890 emotional words and the emotional categories they belong to. For instance, keyword *frustrated* is classified into the Sad category, and *freak out* is classified into Shock category, etc. There are a total of 6 categories of Emotional action, according to the analysis result by Wmatrix, including, General, Liking, Calm/Violent/Angry, Happy/Sad: Happy, Happy/Sad: Contentment, Fear/bravery/Shock, and Worry/Concern/Confident.

3.2 Categorisation

To discover which types of profane words are used in which types of hate discussions. The 500 abusive comments are further classified into eight different hate categories. The eight hate categories are: Race, Behavior, Physical, Sexual Orientation, Class, Gender, Disability, Religion and Others. The rationale behind this categorisation is to obtain the information that which profane word

This is a preprint of an article to be published in 2018 Association for Computing Machinery. Phoey Lee Teh, Chi-Bin Cheng, Weng Mun Chee (2018) Identifying and Categorising Profane Words in Hate Speech, Proceedings of the 2nd International Conference on Compute and Data Analysis, Pages 65-69, https://dl.acm.org/citation.cfm?doid=3193077.3193078

is more likely to appear in a certain hate category, thus providing a mean of probability estimate when assessing hate speech by automatic methods.

3.3 Benchmarking

Reviewing all the profane words that appear in the comments, there are 736 profane words in total being used today as in the year 2017, which take around 11 percent of the total keywords extracted (i.e. 6890). We also compares the list of profane words identified in this study with the lists of words reported in the year 1970 by (Cameron, 1970) as well as in the year 1972 by [12]. This comparison not only shows the evolution of frequently-used profane words from generation to generation, it can also assist in describing the profile of an abuser. Through identifying the type of profane words an abuser uses in the social media towards their young victims, the percentage of identifying the abuser's generation can be estimated to a certain extent.

4. RESULTS

Figure 1 shows the distribution of profane words used in each category. Apparently, profane terms that are related to Sexual Orientation scored the highest percentage of 35% of the entire population. Following with 20% of abusive terms relating to Social Class, or status and 16% of them toward physical abuse. Words in behavior list scored hardly appear, with only 1%.

Terms that appear in the comments commonly repeated with different types of spelling. Table 1 exhibits the distribution of hate categories in hate comments and the profane words that appear in each category. Profane words in the category of sexual orientation take 35.1%, and the category of disability contributes 20.14% and the category of physical 11.05%. These three categories together account for more than 66% of the abusive comments.

With the information provided by Table 1, we can gain knowledge about which profane words are more critical in judging the category of hate speech. To detect hate speech, the above information can assist in evaluating the probability of a comment being hate speech with the presence of profane words listed in Table 1. For example, words that describe human body and anatomy, or represent human disability, may indicate a higher probability of identifying the comment as hate speech. This list of profane words used in hate speech could also provide a chance to be identify with discourse markers. And with this discourse markers, the information can assist in judging the type of hate speech.

Along our analysis, we have found some new hate/profane words used. After comparing today's profane words with the list produced in the past 30 years. The usage of profane words has been changed across the different time line. With this information, we can extend the base of hate/profane words.

In percentage, for each profane word, we identify the percentage of usage over the overall number of words, and we compare the percentage in figure 2. Based on the top 10 highly used profanity words in 2017 YouTube comments, comparing with the samples from the studies of [3] and of [12]. We verify that the word ``fuck'' scored the highest in year 2017 over the total number of words appearing in those comments. Profanity words such as ``fuck'' and ``ass'' have increased in usage in the year 2017 comparing to 1970 and 1972.

It has also evidenced that some of the usage has diminished after so many years. It is evidenced by the words ``hell'', ``god'' and ``damn'' which were more frequently used in the years 1970 and 1972 comparing to the year 2017. However, there are some words that are still used frequently after so many years, such as the words ``bitch'' ,``shit'' and ``cock''.

5. DISCUSSIONS

From our result, it is presented that there are cases that cannot be correctly classified by the lexicon-based approach. In particular, Type I error, where the cases is not a hate speech but judged as one, and Type II error, where the case is a hate speech but judged as not-hate speech. In the case of Type I error where profane words presented in the comment, such a comment is often an emotional expression rather than an intentional abuse of language, but the comment would be judged as a hate speech when lexicon-based approaches are applied. Type II error occurs in a few different cases.

One of the cases is that users substitutes one letter to another in a profane word, e.g. fxck, or deletes a single letter in the word, or the insertion of a single letter and the transposition of a single letter [1] in the profane word. In such a case, lexicon-based approaches fail because these intentionally misspelling words are not in the base. Another case of Type II error is the problem of typos. Typos are very common in comments, e.g. Niggar, Nigga, etc. One way to deal with these cases is to include all the possible misspelling or typos in the word base; however, the task would be very tedious and time-consuming to complete manually. Big data analysis can help identify the frequently used typos, by focusing on common typos and concluding a dictionary to simulate all the typos, a bigger corpus can be built.

6. CONCLUSIONS AND FUTURE RESEARCH

This study has analyzed the comments on YouTube and obtained a list of frequently-used profane words and their categorization into different types of hate speech. Comparisons of the usages of profane words in different generation were also carried out to understand the evolution of frequently-used terms over time. This information can assist in describing the profile the abuser. Our analysis result also demonstrated that errors of assessing hate speech can occur by using a lexicon-based approach. Comments with simply emotional expression or typos can lead to misjudgment of hate speech.

In our future study, machine learning techniques will be applied to the detection of hate speech on social media. The results found in the current report will be used to extract features from comments for machine learning techniques to be applied. The distribution of profane words in hate comments can also be used to estimate prior probabilities when Bayes theorem-based approaches are used. Besides profane words, features of comments can be extended based on the findings of this report. For example, word counts of comments, hate intensity, etc. Findings of the report can also be used to formulate rules to further refine the (Magu \& Kshitij, 2017) classification results by machine learning.

This is a preprint of an article to be published in 2018 Association for Computing Machinery. Phoey Lee Teh, Chi-Bin Cheng, Weng Mun Chee (2018) Identifying and Categorising Profane Words in Hate Speech, Proceedings of the 2nd International Conference on Compute and Data Analysis, Pages 65-69, https://dl.acm.org/citation.cfm?doid=3193077.3193078

7. ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.

8. REFERENCES

1. Cyril N. Alberga. 1967. String similarity and misspellings. *Communications of the ACM* 10, 5: 302–313. https://doi.org/10.1145/363282.363326
2. Paul Cameron. 1970. The words college students use and what they talk about. *Journal of Communication Disorders* 3, 1: 36–46. https://doi.org/10.1016/0021-9924(70)90030-4
3. Paul Cameron. 1970. The words college student use and what they talk about. *Journal of Communication Disorders* 3, 1: 36–46. https://doi.org/10.1016/0021-9924(70)90030-4
4. Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language.
5. Richard Delgado and Jean Stefancic. 2015. Hate Speech In Cyberspace. *Wake Forest Law Review* 1, 1: 8–23. https://doi.org/10.3868/s050-004-015-0003-8
6. Melissa Eddy and Mark Scott. Delete Hate Speech or Pay Up, Germany Tells Social Media Companies - The New York Times. *The New Yorks Times*. Retrieved July 25, 2017 from https://www.nytimes.com/2017/06/30/business/germany-facebook-google-twitter.html
7. Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering* 10, 4: 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21
8. Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Poster : Detection of Cyberbullying in a Mobile Social Network : Systems Issues. In *The Inernational Conference on Mobile Systems, Applications, and Services, MobiSys*.
9. Timothy Jay. 2009. Do offensive words harm people? *Psychology, Public Policy, and Law* 15, 2: 81–101. https://doi.org/10.1037/a0015646
10. Rijul Magu, Kshitij Joshi, and Jiebo Luo. 2017. Detecting the Hate Code on Social Media. Icwsm: 608–611. Retrieved from http://arxiv.org/abs/1703.05443
11. Irene Nemes. 2010. Information & Communications Technology Regulating Hate Speech in Cyberspace : Issues of Desirability and Efficacy Regulating Hate Speech in Cyberspace : Issues of Desirability and Efficacy. 1, July 2012: 37–41. https://doi.org/10.1080/136008302200003190
12. G. Patrick Nerbonne and Nicholas M. Hipskind. 1972. The use of profanity in conversational speech. *Journal of Communication Disorders* 5, 1: 47–50. https://doi.org/10.1016/0021-9924(72)90029-9
13. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web*: 145--153. https://doi.org/10.1145/2872427.2883062
14. Bhikhu Parekh. 2006. Hate speech. *Public Policy Research* 12: 11. https://doi.org/10.1111/j.1070-3535.2005.00405.x
15. Paul Rayson. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13, 4: 519–549. Retrieved August 13, 2015 from http://www.jbe-platform.com/content/journals/10.1075/ijcl.13.4.06ray
16. Leandro Araújo Silva. 2017. A Measurement Study of Hate Speech in Social Media. 85–94.
17. Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. *CEUR Workshop Proceedings* 1816: 86–95.
18. Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. Cursing in English on twitter. In *The 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing,* 415–425. https://doi.org/10.1145/2531602.2531734

Columns on Last Page Should Be Made As Close As Possible to Equal Length

Authors' background

Your Name	Title*	Research Field	Personal website

This is a preprint of an article to be published in 2018 Association for Computing Machinery. Phoey Lee Teh, Chi-Bin Cheng, Weng Mun Chee (2018) Identifying and Categorising Profane Words in Hate Speech, Proceedings of the 2nd International Conference on Compute and Data Analysis, Pages 65-69, https://dl.acm.org/citation.cfm?doid=3193077.3193078

Phoey Lee Teh	Associate Professor	Text Analysis, Sentiment Analysis, Data Analysis	https://scholar.google.com.my/citations?user=m1gANC8AAAAJ&hl=en
Chi-Bin Cheng	Professor	Soft Computing, Machine Learning, Decision Analysis, Supply Chain Management, e-Commerce	http://www.im.tku.edu.tw/en/teacher/cheng-chi-bin/
Weng Mun Chee	Student	Sentiment Analysis	Nil

***This form helps us to understand your paper better, the form itself will not be published.**

***Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor**

This is a preprint of an article to be published in 2018 Association for Computing Machinery. Phoey Lee Teh, Chi-Bin Cheng, Weng Mun Chee (2018) Identifying and Categorising Profane Words in Hate Speech, Proceedings of the 2nd International Conference on Compute and Data Analysis, Pages 65-69, https://dl.acm.org/citation.cfm?doid=3193077.3193078