Rephrasing Profanity in Chinese Text

Hui-Po Su, Zhen-Jie Huang, Hao-Tsung Chang[1] and Chuan-Jie Lin[2]
Department of Computer Science and Engineering
National Taiwan Ocean University
{[1]htchang.cse, [2]cjlin}@mail.ntou.edu.tw

Abstract

This paper proposes a system that can detect and rephrase profanity in Chinese text. Rather than just masking detected profanity, we want to revise the input sentence by using inoffensive words while keeping their original meanings. 29 of such rephrasing rules were invented after observing sentences on real-word social websites. The overall accuracy of the proposed system is 85.56%

1 Introduction

Profanity, or offensive language, is often seen in social media, especially when users can write things anonymously. In nowadays, many social media or chatrooms have policies to detect and mask offensive words in order to reduce web abuse.

Most of the English profanity processing systems maintain a list of offensive words and their substitutions. A simple match-and-replace method can handle most of the cases (Razavi *et al*., 2010; Vandersmissen, 2012; Xiang *et al*., 2012; Bretschneider *et al*., 2014).

However, detecting profanity in Chinese is more complicated than in English. A Chinese character appearing in an offensive word may also appear in a mild word. For example, although the character "幹" means "fxxk" in some context, it also has meanings of "do", "work", and "stem", as its appearances in the words "幹活" (do works) or "枝幹" (tree branches and stems). Simple detection and masking will filter out inoffensive words.

Moreover, our system tries to offer an alternative way to express what a writer wants to say, rather than just applies masking and leave the offensive words there. There are two reasons that we want to rephrase the offensive expressions instead of masking them:

1. Sometimes in Chinese, the masking will make the sentence incomprehensible.

2. Hopefully, it will gradually make the writer to put words more politely.

To our best knowledge, there is no much research work in NLP discussing about detecting and rephrasing profanity in Chinese text. As a preliminary study, we only focused on the most frequent types of offensive words in Traditional Chinese, as well as some Mandarin transliterations in the Southern Min dialect (referred to as Taiwanese hereafter).

We have to apologize that this paper contains a lot of offensive words, both in Chinese and English. We will rewrite these words in the following ways to make them less offensive: a) for a single Chinese character, it will be replaced by an uppercase English letter, such as "F 你"; b) for a multi-character Chinese word, underlines are inserted between characters, such as 賤_人; c) for a phrase, plus signs are inserted between words, such as 你+奶奶+的; d) for an English word, some letter will be replaced by 'x', such as "fxxk". We hope the readers can feel less offended with these revisions.

Another challenge in English is to detect abusive languages in articles, including racist, sexual-oriented harassment, bullying, and hateful speech (Ross *et al*., 2016; Waseem, 2016; Waseem and Hovy, 2016; Wulczyn, *et al*., 2017). It will be our future work in Chinese.

This paper is organized as follows. Section 2 defines the categories of offensive expressions studied in this paper and explains how to build the experimental dataset. Section 3 describes the rephrasing rules. Section 4 delivers the evaluation and error analysis. Section 5 concludes the paper.

Proceedings of the First Workshop on Abusive Language Online, pages 18–24,
Vancouver, Canada, July 30 - August 4, 2017.

Figure 1. An Example of searching in Twitter

2 Profanity Data Collection

In order to observe all variations of Chinese profanity in real world, we built a data collection from social media. Offensive expressions were annotated and their milder paraphrases were also created by humans. The procedure is described briefly in the section.

2.1 Real-World Profanity Data Collection

Two popular social media, Twitter and PTT (a famous BBS in Taiwan), were chosen as the source websites to collect offensive expressions.

Twitter[1] provides a search tool to find tweets containing submitted keywords. A search example is illustrated in Figure 1. Because Twitter only returns a small set of results for one query, we only collected the top 30 results for study.

PTT[2], on the other hand, does not provide any tool to search posts. We used Google to do the searching by adding the option "site:ptt.cc", which restricts the source website of the search results. We then retrieved the full texts of the posts by visiting the URLs in the search results. At most top 300 results for each query were collected.

Queries were Chinese characters or words commonly used in Chinese profanity. We expanded the query terms with their synonyms in Tongyici Cilin (同義詞詞林, a thesaurus of Chinese synonyms) or well-known substitutions with similar pronunciation. These query terms belong to the following four categories.

1. Terms related to "sexual intercourse"
2. Terms related to sexual organs or substances
3. Terms synonymous to "bxtch"
4. Terms in the pattern of "one's relative's", a special pattern of profanity in Chinese

In case that tweets or posts written in Simplified Chinese were collected, we converted them into Traditional Chinese by three mapping dictionaries developed by Wikipedia with the longest-matching strategy.

We found that not all the search results from PTT were suitable for our research. Some users expressed their anger against some persons, teams or TV programs by changing their names into indelicate characters. These are not common cases thus should be filtered out. In order not to spend too much human effort on filtering, posts from the Gossip Board and the sex-related boards were discarded. The source board of a post can be determined from its URL.

There are totally 9,557 sentences in the test set.

2.2 Data Annotation

The main purpose of our system is to rephrase profanity into another meaningful text. It is important to find suitable substitutions so that the rephrased text is fluent and has the same (or similar) meaning of the original text. Therefore, we built a develop set as gold standard for observation.

Two annotators (college students) were asked to browse the posts collected from Twitter or PTT, extract the sentences containing profanity, and provide possible paraphrases. Besides, if they saw a sentence containing obscene keywords but was not offensive, they would tag the sentence as "no need to change".

Moreover, if two annotators had different opinions on the same sentence, the authors would discuss and make decisions. Most of the disagreements were about the determination of indelicate text. Some found a text offensive while the other could tolerate it.

3 Detection and Rephrasing Rules

The main purpose of this paper is to develop a system which can detect and rephrase profanity in a given input. This system can be integrated with social media. After a user writes down some words, our system can provide a more decent way to express the same thing before the message is submitted. An example is shown in Figure 2 when integrating with Facebook.

[1] https://twitter.com/

[2] https://www.ptt.cc/bbs/

Before rephrasing:

But... nonetheless, I still want to say...
❶Fxxk! It's all ❷bxllshxt!
I'm ❸pixsed off!

After rephrasing:

But... nonetheless, I still want to say...
❶Darn! It's all ❷trash talking!
I'm ❸very angry!

Figure 2. An example of profanity rephrasing on Facebook

After observing the real-world indelicate texts and their paraphrases, we invented 29 sets of detection patterns and rephrasing rules. Detection patterns consist of surface strings, word sets, and preconditions to apply the rules. Appendix A lists all the detection patterns and rephrasing rules. The following subsections explain the definitions and the challenges of the five major categories.

3.1 Phrases with F-words

The direct Chinese translation of "fxxk" is "幹" (masked by **F** hereafter in this paper). Appendix A also lists 3 other synonymous characters. Such characters have several usages.

1. It can be used as a verb as a swear word as in the phrase "**F** 你老師" ("fxxk your teacher"). The writer only wants to express his or her anger, so it can be replaced by "darn" or "oh no" (Rule #1).
   Note that the object of such a verb is often a relative or a close person to the hearer (such as mother or teacher).
2. A single word can form an exclamatory sentence "**F**!" ("Fxxk!"). It has the same meaning as the first case (Rule #2).
3. It can be used as an adjective as in the phrase "覺得很 **F**" ("feeling really fxxked"). Most of the time there will be an adverb preceding it. It can be replaced by an adjective synonymous to "angry" (Rule #3).
4. It can be used as a sentence opener as in the sentence "**F** 昨天忘了買鞋" ("Fxxk I forgot to buy shoes yesterday"). Replacing it with "oh no" is OK (Rule #4).
5. It may appear in an inoffensive word such as "幹活" (do works) or "操作" (operate). A list of formal words containing these offensive characters is maintained. Words in this list will remain unchanged in the input text (Rule #27).

3.2 Phrases Containing Relatives

The original phrase in this category is "他+媽+的" ("to his mother" / "his mother's"). Due to the explosion of social media, many similar phrases have been invented. They are all in the pattern of Pronoun + **RL** + 的, where Pronoun is a 2nd- or 3rd-person singular pronoun and **RL** is a relative title such as "奶奶" (grandma) or "妹妹" (sister). Note that "的" is a particle and carries no content information.

Such a phrase can also be used as a possessive form in a formal text, such as "他奶奶的拿手菜" ("his grandmother's specialty dish"). But when it is used alone, it becomes offensive (Rule #29).

3.3 Words Synonymous to "Bxtch"

Words in this category are used to scold somebody, so they can be replaced by phrases like "bad person" (Rules #5 ~ #7) which is less offensive.

3.4 Phrases with the Word "Semen"

The word "semen" has more than one translation in Chinese. Its formal term is "精液" and its obscene term is "洨" (siao2, Mandarin transliteration of Taiwanese dialect; masked by **X** hereafter).

Because the obscene term comes from Taiwanese, a lot of Taiwanese profanities are written down in many different ways of Mandarin transliterations, as shown in Appendix A. Their meanings are explained as follows.

1. The Taiwanese word "hau-siau5" means "exaggerating" or "trash talking". Its second character has no corresponding Chinese character therefore is usually written as **X** in text (Rules #12 and #13).
2. The Taiwanese phrase "jia7 siau5" means "eat shxt" and its second character is indeed **X**. We suggest a milder term to expression the same feeling (Rule #14).
3. The Taiwanese phrase "siaN2 siau5" means "what the hxll" and its second character is indeed **X**. We suggest a milder term to express the same feeling (Rule #15).
   Note that "三小" (san-siao3, three + little, what the hxll) is one of the expressions in this category. However, the string may also appear in a common phrase such as "三小時" (three hours). It is rephrased only when it follows a verb (Rule #11).
4. The Taiwanese word "lu5-siau5" means "annoying". Its second character has no corresponding Chinese character therefore is usually written as **X** in text (Rule #17).
5. If the term really means "semen", it should be replaced with its formal term (Rule #16).
6. The character **X** can also be used to replace any character with a sound similar to "siao3" when haters write person names or show names. It is not easy to recover the correct characters in the names. The proposed rule is a baseline rule (Rule #18).

3.5 Phrases Containing Sex Organs

Sex organs often have several names in Chinese. Their obscene terms 屄 (female genital, masked as **B** hereafter), 屌 (male genital, masked as **D** hereafter), and the Taiwanese word "lam7-pha" (scrotum, masked as **LP** hereafter) have developed different meanings in the Internet as listed here.

1. The word "牛 **B**" and the character **D** itself mean "awesome" in some context (Rules #19 and #23).
2. The word "傻 **B**" means "fool" (Rule #20).
3. The character **D** can be an adverb meaning "greatly" as in the phrase "**D** 打" (to defeat greatly) (Rule #24).
4. The character **D** can also be a verb meaning "to pay attention" as in the phrase "**D** 你" (to pay attention to you) (Rule #25).

Rule	Y	N	Acc	Rule	Y	N	Acc
1	85	15	0.85	16	39	5	0.89
2	91	9	0.91	17	98	2	0.98
3	98	0	1.00	18	78	22	0.78
4	76	24	0.76	19	56	44	0.56
5	98	2	0.98	20	99	1	0.99
6	97	3	0.97	21	30	11	0.73
7	70	30	0.70	22	47	53	0.47
8	93	5	0.95	23	98	2	0.98
9	81	19	0.81	24	93	1	0.99
10	88	12	0.88	25	30	14	0.68
11	2	0	1.00	26	68	32	0.68
12	12	0	1.00	27	93	7	0.93
13	96	4	0.96	28	10	0	1.00
14	42	4	0.91	29	86	14	0.86
15	90	10	0.90	**Total**	2044	345	0.86

Table 1. Evaluation results of rephrasing rules

5. The Taiwanese phrase "gui **LP** hoe2" (all + scrotum + fire) means "being pixsed off" or "very angry" (Rule #8).
6. If a word **B**, **D**, or **LP** really refers to "genital", a more decent expression is provided (Rules #9, #21, #22, and #26).

4 Evaluation

As a preliminary experiment, we evaluated our system in a small test set constructed by the following steps. For each of the 29 rephrasing rules, we randomly selected at most 100 sentences containing corresponding keywords to do the evaluation. Note that some groups were infrequent so we only had less than 100 sentences. There are totally 2,389 sentences in the test set.

Each rephrased (or detected but remain unchanged) part was assessed by two assessors in terms of both correct and fluent. The evaluation metric is the ratio of the correctness of the processing by the rephrasing rules. Note that if two or more parts in a sentence were detected, they were assessed separately.

The evaluation result is shown in Table 1, where Acc denotes accuracy. We can see that 15 of 29 groups of rules achieved accuracy above 90% and only 4 groups did not achieved accuracy better than 70%. The overall accuracy was 85.56%.

The main error types are discussed as follows.

1. Out-of-vocabulary problem
   Although we have tried to collect as many variants as possible, there are still newly invented ways to transliterate Taiwanese profanity. For example, "**F** 0 糧" has similar sound to "**F** 您_娘" (fxxk your mother) but does not appear in our dictionary. It is the major error of Rules #4 and #7.
2. Similar sound substitution
   Haters usually like to disparage the targets whom they are criticizing by replacing characters in the names with profane characters **B**, **D**, or **X**. It is not easy to recover the original names and becomes the major errors of Rules #18, #22, and #26.
3. Proper names containing obscene words
   There is a hamburger restaurant in Taiwan whose name is "牛逼洋行". The term "牛逼" inside its name should not be changed by Rule #19. The accuracy of Rule #19 becomes 94% if our system can recognize this name.
4. Sentence segmentation
   Some writers are too lazy to use punctuation marks to separate sentences. Words in different sentences are incorrectly adjoined and matched with wrong rephrasing rules. For example, "你 **D**" should be rephrased into "你厲害" (you are awesome). But "你 **D** 你 **D** 你 **D**…" matches Rule #25 and is incorrectly rephrased as "你理你理你…" (you notice you notice you…).

5 Conclusion

This paper proposes a system to deal with profanity in Chinese text. The system does not only detect profanity, but also provide rephrased text which is less offensive.

Nearly ten thousand sentences containing Chinese profanity were collected from real-world social websites. After human annotation, 29 groups of detection and rephrasing rules were invented. The overall accuracy of our system was 85.56% when evaluating on a test set of 2,389 sentences.

Now we have handled five main types of Chinese profanity. We need to look for a larger dataset in order to expand our rephrasing rules and find more types of profanities in the future.

Moreover, the proposed rephrasing rules were hand-crafted. We should try to discover more rules by machine learning.

References

Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. 2014. Detecting Online Harassment in Social Networks. In *Proceedings of the Thirty Fifth International Conference on Information Systems*, pages 1-14. http://aisel.aisnet.org/icis2014/proceedings/ConferenceTheme/2/

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive Language Detection Using Multi-level Classification. In *Proceedings of Advances in Artificial Intelligence* (*AI 2010*). *Lecture Notes in Computer Science*, 6085. http://link.springer.com/chapter/10.1007/978-3-642-13059-5_5

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*. *Bochumer Linguistische Arbeitsberichte*, 17:6-9. https://www.linguistics.ruhr-uni-bochum.de/bla/nlp4cmc2016/ross.pdf

Baptist Vandersmissen. 2012. *Automated detection of offensive language behavior on social networking sites*. Master Thesis, Universiteit Gent. http://lib.ugent.be/catalog/rug01:001887239

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science. Association for Computational Linguistics*, pages 138-142. http://aclweb.org/anthology/W16-5618.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop. Association for Computational Linguistics*, pages 88-93. http://www.aclweb.org/anthology/N16-2013.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. To appear in *Proceedings of the 26th International Conference on World Wide Web – WWW 2017*.

Guang Xiang, Bin Fan, Ling Wang, Jason I. Hong, and Carolyn P. Rose. 2012. Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (*CIKM 2012*), pages 1980-1984. https://doi.org/10.1145/2396761.2398556

Appendix A. Chinese Profanity Detection and Rephrasing Rules

#	Detection Patterns	Example	Rephrasing Rules
1	**F** (*fxxk*)+ Pronoun + **RL** (*relative title*)+ [老 (*old*)] + [的 (*'s*)] + [**JB** (*genital*)]	**F** 你娘 (*fxxk your mother*)	真是可惡 (*Darn*)
2	**F** (single word in one sentence)		可惡 (*Oh no*)
3	Adverb + **F**	很 **F** (*really fxxked*)	Adverb + 可惡
4	**F** (in the beginning of a sentence)	**F** 我今天... (*Fxxk! Today I...*)	可惡
29	他 + **RL** (*relative title*) + 的 你 + **RL** (*relative title*) + 的	他_媽_的 (*to his mother*) 你_媽_的 (*to your mother*)	Removed 你的 (*your*)
5	**BW** (*bxtchy whxre*)		壞女人 (*bad woman*)
6	賤_人 (*bxtch*)		壞人 (*bad person*)
7	**BT** (*bxtchy*)		機車 (a mild term)
11	Verb + 三小 (*what the hxll*)	你在講三小 (*what the hxll are you talking about*)	Verb + 什麼 (*what*)
12	在 (*be*) + **H** (*to lie*) + **X** (*semen*)	在_豪_泛 (*is shxt talking*)	在 + 瞎扯 (*trash talking*)
13	**H** + **X**	豪泛的劇情 (*ridiculous plot*)	唬人 (*to bluff* or *to lie*)
14	**J** (*to eat*) + **X** (*semen*)	你_甲_泛_啦 (*Eat shxt!*)	撞牆 (*run into the wall*)
15	**S** (*what*) + **X** (*semen*)	我到底在玩三_泛 (*what the hxll am I playing at all*)	什麼 (*what*)
16	的 (Chinese particle) + **X** (*semen*) (at the end of a sentence)		的 + 精液 (formal term)
17	**L** (*annoy*)+ **X** (*semen*) **L** + **X** + **X**	魯_泛 (*to annoy*) 魯_泛_泛 (*annoying person*)	糾纏 (*to annoy*) 煩人精 (*annoyer*)
18	**X** (sia02, not the cases above)	泛_明 (*Little Min*, a name)	小 (sia03, *little*)

Synonym sets:

F = 幹, 操, 肏, (Taiwanese) 賽

RL = 娘 (*mother*), 祖母 (*grandma*), 老師 (*teacher*), 全家 (*whole family*)...

JB =機_掰, 雞_掰

BW = 賤_婊, 婊_子, 破_麻, 賤_婊_子, 淫_蕩, 淫_娃, 賤_貨, 賤_女人

BT = **JB**, 機_八, 雞_八, 機_歪, 雞_歪, 機機_歪歪, 雞雞_歪歪

X = (Taiwanese) 泛

H = (Taiwanese) 豪, 唬, 虎, 毫

J = (Taiwanese) 甲, 假, 呷

S = (Taiwanese) 三, 撒, 殺, 啥, 沙

L = (Taiwanese) 魯, 盧, 嚕

Format:

SET (transliteration, *English meaning,* condition or note) [optional]

Appendix A. Chinese Profanity Detection and Rephrasing Rules (*Cont.*)

#	Detection Patterns	Example	Rephrasing Rules
8	**ALL** + **LP** (*scrotum*) + **FIRE** **LP** (*scrotum*) + **FIRE**	歸_覽_趴_會 (*fire full of my scrotum; pixsed off*)	滿肚子氣 一肚子氣 (*very angry*)
9	**PN** (*pxnis*) **LP** (*scrotum*)		那話兒 (a mild term)
10	去死 (*to go to hxll*)		去撞牆 (*to go to bump into the wall; to punish yourself*)
19	牛 (*cow*) + **B** (*puxs*)	牛_屄	厲害 (*awesome*)
20	傻 (*stupid*) + **B** (*puxs*)	傻_屄	傻子 (*fool*)
21	臭 (*stinky*) + **B** (*puxs*)	臭_屄	臭 + 下體 (*private part*)
22	**B** (not the cases above)		女生下體 (*female genital*)
23	Adverb + **D** (*dxck*)	特_屌 (*very impressive*)	Adverb + 厲害 (*awesome*)
24	**D** (*dxck*) + 打 (*to beat*)	屌_打 (*to defeat*)	打爆 (a mild term)
25	**D** (*dxck*) + Pronoun	不_屌_你 (*don't give you a shxt*)	理 (*to notice*) + Pronoun
26	**D** (not the cases above)		那話兒 (a mild term)
27	Formal words containing in-decent characters	幹部 (*manager*) 幹活 (*do work*)	Unchanged
28	A list of direct mappings	怪洨 睡懶覺	怪咖 (*weirdo*) 睡覺 (*to sleep*)

Synonym sets:

ALL = (Taiwanese) 歸, 規, 龜

LP = (Taiwanese) 覽_趴, 懶_趴, 攬_趴

FIRE = 火, (Taiwanese) 會

PN = (Taiwanese) 懶_覺, 懶_較, 覽_覺, 覽_較

B = 屄, 逼

D = 屌