Annotation and Classification of Toxicity for Thai Twitter

Sugan Sirihattasak, Mamoru Komachi, Hiroshi Ishikawa
Tokyo Metropolitan University
6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
sirihattasak-sugan@ed.tmu.ac.jp, {komachi, hiroshi-ishikawa}@tmu.ac.jp

Abstract

In this study, we present toxicity annotation for a Thai Twitter Corpus as a preliminary exploration for toxicity analysis in the Thai language. We construct a Thai toxic word dictionary and select 3,300 tweets for annotation using the 44 keywords from our dictionary. We obtained 2,027 and 1,273 toxic and non-toxic tweets, respectively; these were labeled by three annotators. The result of corpus analysis indicates that tweets that include toxic words are not always toxic. Further, it is more likely that a tweet is toxic, if it contains toxic words indicating their original meaning. Moreover, disagreements in annotation are primarily because of sarcasm, unclear existing target, and word sense ambiguity. Finally, we conducted supervised classification using our corpus as a dataset and obtained an accuracy of 0.80, which is comparable with the inter-annotator agreement of this dataset. Our dataset is available on GitHub.

Keywords: toxicity, corpus, Thai, Twitter

1. Introduction

With the rise of social media in Thailand, it has become an integral part of the daily lives of Thai people, providing various opportunities for education, relationships, and career development. Despite these benefits, online toxicity is not only becoming harsher, but also difficult to control. Furthermore, the victims of toxic messages are not always the intended targets of those messages. According to Wang et al. (2011), many people regret their negative posts because of problems they face later, such as being terminated from employment and losing other opportunities. The instances of bullying or any similar toxic behavior are not easy to delete once they are posted publicly. In particular, any post shared on social media can potentially spread widely across an entire community with a considerably small possibility of deleting it and undoing its effects.

Consequently, there have been many research efforts among various fields such as social science, psychology, and natural language processing, to improve the quality of online conversion while considering the right to freedom of speech. For example, the Google Jigsaw Team launched the Perspective API[1] to identify toxic comments.

Figure 1: Example of toxicity evaluation from Perspective API.

One of the challenges in studying toxicity in online communication is a clear common definition of toxicity in the case of language. Toxic comments are often sarcastic and indicate aggressive disagreement; in Kolhatkar and Taboada (2017), the relationship between constructiveness and toxicity including toxicity levels in news comments was studied. In our study, we define toxicity with a more general perspective to include any messages that can imply toxic behavior (Kwak and Blackburn, 2014), antisocial behavior (Cheng et al., 2017), online harassment (Yin et al., 2009), hate speech (Davidson et al., 2017), cyberbullying (Van Hee et al., 2015), and any type of offensive language (Razavi et al., 2010). In particular, a toxic message is any message that may hurt or harm an individual or a generalized group, may challenge the societal norms, or negatively affect the entire community. In terms of toxic words, we consider any negative words, such as those associated with profanity and obscenity, or those which are offensive.

Though there is an increase in the studies related to toxicity, open resources related to it are still limited. There are several corpora for major languages like English, including a harassment dataset (Kennedy et al., 2017), hate speech Twitter annotation corpus (Waseem and Hovy, 2016), and personal attacks comment corpus (Wulczyn et al., 2017). Unfortunately, researches related to this topic do not include minor languages, such as the Thai language. To our best knowledge, there is no public Thai resource related to online toxicity. Furthermore, text analysis in Thai language is complicated due to ambiguity in segmentation (Cooper, 1996); for example, “ปลาตากลมตัวนี้น่ารัก (This round-eyes (ตา | กลม) fish is cute.)” and “ขอเดินออกไปตากลม (Let me go out to have some fresh air (ตาก | ลม)).” Likewise, sentence boundary detection is difficult (Zhou et al., 2016) because the space which is used for differentiating sentences is not appropriate in some cases such as in “โอ๊ย! เจ็บ (Ouch! it hurts).”

Some toxic tweets that are typical in the case of bullying messages, such as “ไอ้ห่า! ไปตายซะ คนไร้ประโยชน์แกก็เหมือนพ่อแก” (Damn you! Just go to die. You are useless just like your father.), may not only affect an individual, but also his or her family. Thus, we present annotation and classification of toxicity on Twitter in the Thai language as a preliminary exploration for toxicity analysis in the Thai language in general. The main contributions of this study are as follows:

1. We construct a dictionary of Thai toxic words that we use as keywords for annotation.

[1] http://www.perspectiveapi.com/

2. We build a toxicity corpus based on Twitter messages or tweets, because these messages represent the daily-life conversations of the Thai people.

3. We used our abovementioned dataset to conduct supervised classification and obtained an accuracy of 0.80 for it.

Our dictionary and corpus are available on GitHub[2].
The remainder of this paper is organized as follows. Section 2 introduces the definition of toxicity and describes some difficulties with respect to Thai tweet analysis. Section 3 explains our corpus construction and annotation process including the construction of our dictionary of Thai toxic words. Then, Section 4 presents the analysis of the resulting corpus, while Section 5 provides classification results and discussion. Finally, Section 6 presents the conclusions of our study and indicates future work.

2. Toxicity and Thai Language

Many social media platforms and websites use embedded keyword-based approaches to automatically filter out toxic messages. However, it is possible for individuals who are close friends to casually communicate using toxic words without intending any harm (Nand et al., 2016). Likewise, the factors used to identify politeness in Thai male conversation depend on the situational context such as the relationship between the speaker and listener, and the location at which the conversation takes place, rather than the linguistic aspects (Mekthawornwathana, 2011).
Moreover, the keyword-based approach does not seem flexible for a non-segmenting language like the Thai language. The following two examples contain a toxic word "หอก[3]" (The original meaning is "spear"; however, the slang meaning is an insulting phrase, "Damn, Bitch.")

(a) นักการเมืองหอกเลวมากสมควรตาย
นักการเมือง (politician) | หอก (damn) | เลว (bad) | มาก (very) | สมควร (deserve) | ตาย (die)
The damn Politician deserves to die.
(This is a toxic message.)

(b) ที่หอกล้องวงจรปิดเยอะจึงไม่มีหัวขโมย
ที่ (at) | หอ (dormitory) | กล้องวงจรปิด (security camera) | เยอะ (many) | จึง (therefore) | ไม่ (no) | มี (have) | หัวขโมย (thief/thieves)
There are no thieves because there are a lot of security cameras at the dormitory.
(This is a non-toxic message.)

Therefore, not only ambiguity in segmenting as shown above, but also word variations and homonyms are inevitable obstacles in Thai tweet analysis. For example, the toxic word "เหี้ย" has several homonyms including the following examples presented below.

(a) นักกีฬาประเทศนี้เหี้ยโกงตลอด
นักกีฬา (athlete) | ประเทศ (country) | นี้ (this) | เหี้ย (damn/bad) | โกง (cheat) | ตลอด (always)
An athlete from this country always cheats.
(This is a toxic message.)

(b) อากาศร้อนเหี้ย
อากาศ (weather) | ร้อน (hot) | เหี้ย (damn/very)
The weather is very hot.
(This is a non-toxic message.)

(c) เหี้ยเป็นสัตว์เลื้อยคลาน
เหี้ย (varanus salvator) | เป็น (is) | สัตว์เลื้อยคลาน (reptile)
Varanus salvator is a reptile.
(This is a non-toxic message.)

Thus, the classification of toxicity should not only depend on a word, but also the context in which it is used. In order to achieve this, we need to apply a data-driven approach because a keyword-based approach is insufficient (Saleem et al., 2016); we do this by creating a corpus that contains a variety of examples of toxicity in the Thai language.

3. Dataset Construction and Annotation

3.1. Keyword Dictionary Construction

Because toxic posts often contain toxic words, we used toxic words as the keywords to retrieve the data for our dictionary. We selected some toxic words from the Conceptual Metaphor of Thai Curse Words (Orathai Chinakarapong, 2014) and rechecked spelling using the Royal Institute Dictionary[4]. Then, we added some well-known variations of these toxic words such as "สัส," which is a spelling variation of "สัตว์" (The original meaning of this word is animal and its slang meaning is similar to "damn."). Finally, we included a few negative words, for example, "ฆ่า" (kill) and "แช่ง" (curse), into the set. In total, we included 44 keywords in this dictionary, which are shown in Figure 2.

3.2. Data Collection

We used the public Twitter Search API to collect 9,819 tweets from January–December 2017 based on our keyword dictionary. Then, we selected 75 tweets for each keyword. In total, we collected 3,300 tweets for annotation.
To ensure quality of data, we set the following selection criteria.

1. All tweets are selected by humans to prevent word ambiguity. (The Twitter API selected the tweets based on characters in the keyword. For example, in the case of "บ้า(crazy)," the API will also select "บ้านนอก" (countryside)" which is not our target.)

[2] https://github.com/tmu-nlp/ThaiToxicityTweetCorpus/

[3] This paper contains several inappropriate, impolite, and harsh words in both the Thai and English languages. We rewrite some English toxic words using "*" for some characters or replacing these words with appropriate substitutes. However, we could not rewrite such words for the Thai language because that may lead to an ambiguous word.

[4] http://www.royin.go.th/dictionary

2. The length of the tweet should be sufficiently long to discern the context of the tweet. Hence, we set five words as the minimum limit.

3. The tweets that contain only extremely toxic words, (for example: "damn, retard, bitch, f*ck, slut!!!") are not considered.

4. In addition, we allowed tweets with English words if they were not critical elements in the labeling decision, for example, the word "f*ck." As a result, our corpus contains English words, but they are less than 2% of the total.

All hashtags, re-tweets, and links were removed from these tweets. However, we did not delete emoticons because these emotional icons can imply the real intent of the post owners. Furthermore, only in the case of annotation, some entries such as the names of famous people were replaced with a tag <ไม่ขอเปิดเผยชื่อ>, for anonymity to prevent individual bias.

3.3. Annotation

We manually annotated our dataset with two labels: Toxic and Non-Toxic. We define a message as toxic if it indicates any harmful, damage, or negative intent based on our definition of toxicity. Furthermore, all the tweets were annotated by three annotators to identify toxicity; the conditions used for this identification are presented in the following list.

- A toxic message is a message that should be deleted or not be allowed in public.

- A message's target or consequence must exist. It can either be an individual or a generalized group based on a commonality such as religion or ethnicity, or an entire community.

- Self-complain is not considered toxic, because it is not harmful to anyone. However, if self-complain is intended to indicate something bad, it will be considered as toxic.

- Both direct and indirect messages including those with sarcasm are taken into consideration.

We strictly instructed all the annotators about these concepts and asked them to perform a small test to ensure they understood these conditions. The annotation process was divided into two rounds. We asked the candidates to annotate their answers in the first round to learn our annotation standard. Then, we asked them to annotate a different dataset and selected the ones who obtained a full-score for the second round as an annotator. From among these annotators, 20% of the annotators failed the first round and were not involved in the final annotation.

4. Corpus Analysis

As previously mentioned, the corpus consists of 3,300 tweets divided into 2,027 toxic tweets and 1,273 non-toxic tweets. The labels are assigned based on majority decisions. The numbers of tweets with perfect agreement, referred to as gold standard tweets, are 1,692 and 1,093 for toxic and non-toxic cases, respectively. The inter-annotator agreement (Fleiss' Kappa) (Carletta, 1996) is 0.78, which shows that the agreement is significant.

There are three primary reasons for disagreement. First, more than 35% of tweets that annotators disagreed upon are difficult to judge as toxic or non-toxic because of sarcasm. Second, it is ambiguous whether a message owner is self-complaining or referring to someone else or some group by cunning to avoid defamation. Lastly, there are some cases where word sense ambiguity is affected by the annotation. For example; "Damn it, I want to commit arson on the university," which can imply that he/she is very stressed out and just wants to complain. This kind of sarcastic expression is quite common in Thailand. However, there is a possibility that the owner of the comment really intends to commit such a crime.

The distribution of toxic and non-toxic tweets is shown in Figure 2. Interestingly, the tweets that contain toxic words related to animals are less likely to be toxic than the rest except in the cases of "แมงดา" (pimp/horseshoe crabs) and "ควาย" (stupid/buffalo). Most of the non-toxic cases for "แมงดา" refer to one of Thailand's popular dish that is made from horseshoe crabs while "ควาย" seems to be rarely used for its literal meaning of buffalo. Moreover, the words that related to bottom like "ต่ำ" (low) and "ส้นตีน" (heel) are commonly used in a toxic manner because they are antonyms to the words "top" or "high" which Thai people believe indicate a sacred position like a head. The word "โง่" (stupid) seems to be used in a non-toxic manner rather than for toxic purposes. Based on the non-toxic tweets from our corpus, we found that people tend to use the word "stupid" whenever they want to blame themselves. Moreover, as part of everyday conversation, people use the word "หมา" (dog) not only as an insult, but also to refer to a pet or as an adorable joke. Surprisingly, the usage of the word "ชั่ว" (wicked) is not limited as a toxic word, but we found that, in everyday conversation, like in the case of teaching or reporting a situation, it is used in a non-toxic manner as well. Finally, the word "สัตว์" (animal) is used by people for its original non-toxic meaning. This is in contrast to its variations such as "สัส" and "สัด," which are more likely to be used in a toxic manner.

In the case of toxic tweets, we found that a word, "ควย," which refers to f*ck or genitalia, is highly toxic and unpleasant regardless of the level of contextual toxicity.

Some tweets are difficult to label leading to inconsistency in annotation as shown in Table 1. Moreover, Thai people often use metaphors in their conversations as indicated in the example below.

กินกะหรี่ปั๊บอร่อยไม่เหมือนกินกะหรี่

กิน (eat) | กะหรี่ปั๊บ (curry puff) | อร่อย (yummy/delicious) | ไม่ (not) | เหมือน (similar to) | กิน (eat) | กะหรี่ (curry? whore?)

Eating curry puff is yummy not like eating curry (whore?). In such cases, it is difficult to ascertain the meaning of the

Table 1: Top three conflicts in annotation agreement.

Keywords (original/toxic meaning)	Disagreement of tweets (%)
กะหรี่ (curry/whore)	22.7
ห่า (damn) หอก (spear/bitch) ฉิบหาย (woeful)	21.3
ตอแหล (lie) เห็บ (tick/parasite) ปลวก (termite/ugly) ประสาท (nerve/insane) ส้นตีน (heel) ดัดจริต (pretentious) แช่ง (curse) จัญไร (beastly)	20.0

word "กะหรี่"; thus, its purpose is vague and could either indicate a warning or be an attack against someone. These types of tweets are common in Thai Twitter because people avoid mentioning the target of the message directly to prevent legal repercussions or other issues.

5. Classification Experiment

5.1. Data

Aside from the steps performed for annotation, we conduct further tweet data cleaning after we have segmented the tweets into tokens using the Deepcut library version 0.6[5].

1. We normalized repetitive letters, for example, "มากกก" to "มาก" and "5555..." to "555." The pronunciation in Thai for number 5 is "Ha," therefore, people always use it as a substitute for the laugh sound.

2. We removed stopwords and punctuation marks except "?" and "!" because they may be related to some emotions.

3. We removed non-Thai words.

In order to make a fair comparison, the training data is created by selecting equal number of toxic and non-toxic instances from the corpus; in particular, we selected 1,888 tweets with 944 toxic tweets and 944 non-toxic tweets. All of these tweets were selected randomly. Furthermore, each keyword must have an equal number of tweets for both labels and the maximum number of tweets per label is 30. For test data, we used 176 tweets from among the gold standard tweets with 2 toxic tweets and 2 non-toxic tweets per keyword.

5.2. Setting

For classification, we use the CountVectorizer method from the scikit-learn library version 0.19[6] to create bag-of-word features and set the threshold to 10 for minimum document frequency. From the same library, we tuned hyper-parameters for the LogisticRegression method using the GridSearchCV method. We setup the hyper-parameters as follows.

1. C value: 0.001, 0.01, 0.1, 1, 10.

2. Fit intercept: True or False.

3. Penalty: L1 or L2.

Finally, our baseline is to set all predictions of toxic tweets according to the keyword-based approach, because all tweets contain toxic keywords.

[5] https://github.com/rkcosmos/deepcut

[6] https://github.com/scikit-learn/scikit-learn

Table 2: Classification result.

Method	Precision	Recall	F1-Score
Logistic Regression	0.87	0.70	0.78
Keyword Baseline	0.50	1.00	0.67

5.3. Results and Discussion

Table 2 shows the experimental results. The best accuracy is 0.80, when the hyper-parameters are C = 0.1, Fit intercept = True, and Penalty = L2. We obtained 9 false negatives and 26 false positives, as can be seen in Figure 3. Compared with the keyword baseline method, our classification results are better in terms of precision and F1-score.

Although the keyword-based approaches are popular for performing this type of classification, it failed to correctly classify some tweets, as in the following example, which is a Thai-English translated tweet: "Damn, just finished laundry and it's raining." In contrast, our approach correctly classified it as non-toxic.

Furthermore, in our approach, the primary reason for an error in the case of a false positive is complaining in a tweet, examples of which are given in Table 3. The cases of false negatives are primarily because of the following two reasons.

1. Tweets that contain both toxic words and positive words such as "good" or "beautiful."

2. Tweets that contain unknown or low document frequency words in our model.

The examples of false negatives are shown in Table 4.

Because our corpus is small, surface features are insufficient for abbreviation, slang, and unknown words; thus, we need to increase the size of our dictionary to let the model learn more words. In addition, we are aware that using only bag-of-word features is not sufficient for tweet classification; therefore, we will explore more efficient approaches in a future study.

Furthermore, we admit that the auto-segmentation is not perfect, which affects the classification. For example, a tweet that includes a wrong word segmentation like "อะอีดอก" gets incorrectly predicted as non-toxic. The right segmentation should be "อะ (affix) | อี (impolite prefix) | ดอก (bitch)" and with this, the prediction is toxic.

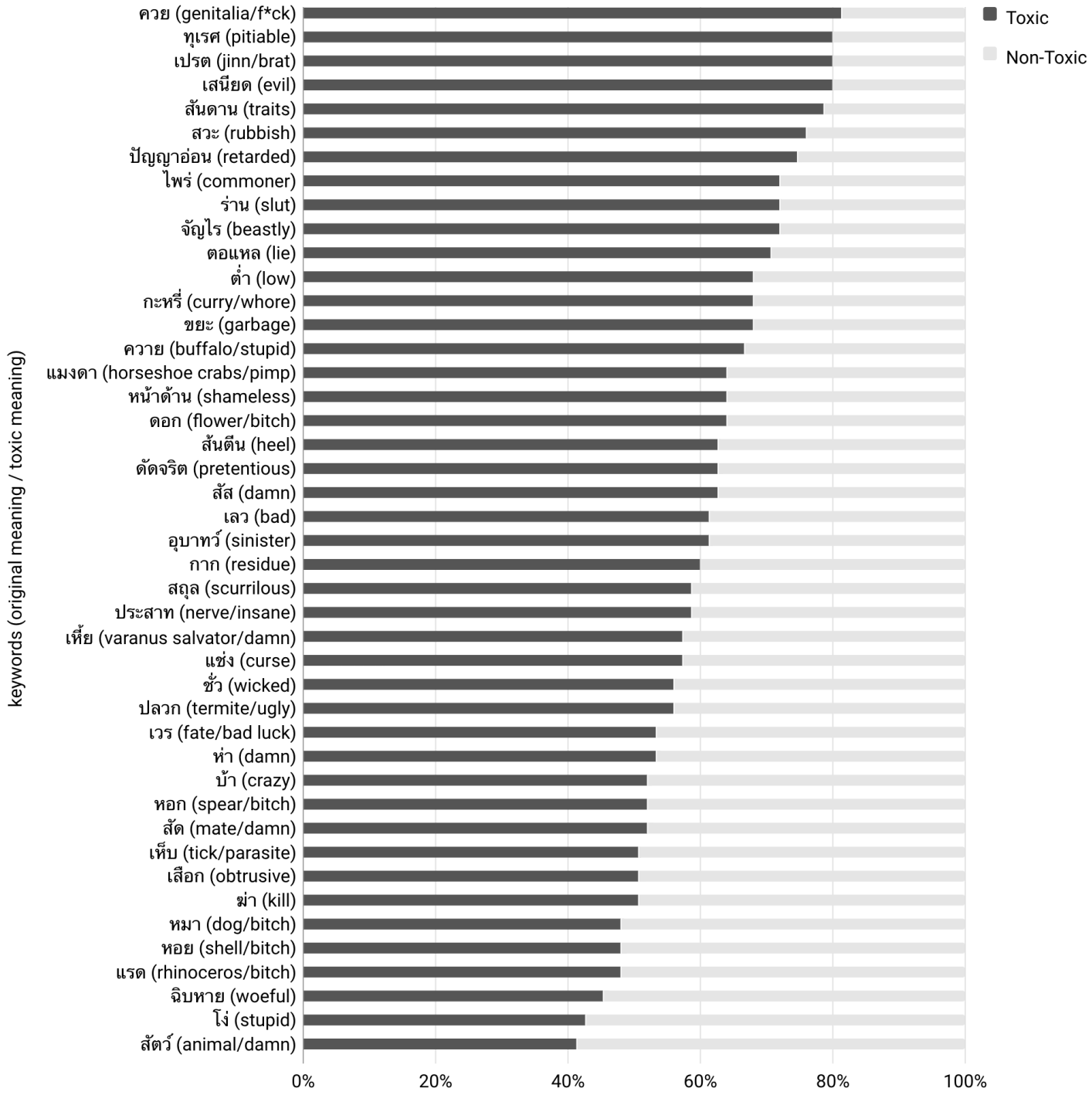

Figure 2: Distribution of toxic and non-toxic tweets based on keywords.

Despite some errors, our auto-segmentation method is considerably effective referring to the examples below.

(a) ถึงคุณรวยล้นฟ้าแต่ไร้น้ำใจก็ยากที่คนจะศรัทธา (Despite of being a millionaire, but without kindness, nobody will respect you.) which auto-segmentation and human-segmentation are same.
ถึง (to/although) ׀ คุณ (you) ׀ รวย (rich) ׀ ล้น (overflow) ׀ ฟ้า (sky) ׀ แต่ (but) ׀ ไร้ (without) ׀ น้ำใจ (kindness) ׀ ก็ (then) ׀ ยาก (hard) ׀ ที่ (at/that) ׀ คน (person/people) ׀ จะ (will) ׀ ศรัทธา (faith).

(b) คนเห็นแก่ตัวที่ไม่เคยเห็นใจคนอื่น (A selfish person who never care for others.)
auto-segmentation: คน (person/people) ׀ เห็น (see) ׀ แก่ (for) ׀ ตัว (self) ׀ ที่ (at/that) ׀ ไม่ (no) ׀ เคย (ever) ׀ เห็นใจ (sympathetic) ׀ คน (person/people) ׀ อื่น (another)
human-segmentation: คน (person/people) ׀ เห็นแก่ตัว (selfish) ׀ ที่ (at/that) ׀ ไม่เคย (never) ׀

Table 3: Examples of false positives.

Tweet text (English translation)	Toxic keyword	True label	Predicted label
Since this morning, the dormitory internet is damn and even now, it is still damn.	damn	Non-toxic	Toxic
I want to shout f*ck but all I can say is yes sir.	f*ck	Non-toxic	Toxic

Table 4: Examples of false negatives.

Tweet text (English translation)	Toxic keyword	True label	Predicted label
You damn, Just go to die for better.	damn	Toxic	Non-toxic
Damn, you're annoying. You are just pretty but stupid.	damn, stupid	Toxic	Non-toxic

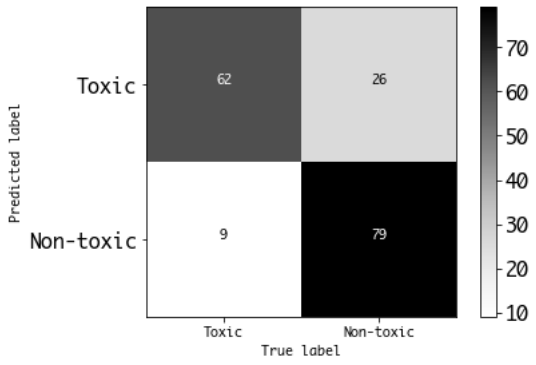

Figure 3: Confusion matrix of toxicity classification.

เห็นใจ (sympathetic) ı คนอื่น (others)

6. Conclusions and Future work

With the increasing popularity of social media in Thailand, the growth of toxicity in online conversation is a growing concern. To the best of our knowledge, there is no public Thai resource related to online toxicity. In this study, we present toxicity annotation for a Thai Twitter Corpus along with a supervised classification method as a preliminary exploration for toxicity analysis in the Thai language.

In the future, we plan to not only enhance the classification method, but also improve our model and use streaming data for the dataset to eliminate bias involved with using keywords. Our improved model will be used to extend the volume of the Thai toxicity corpus.

Furthermore, aside from the corpus, we intend to increase, both, the size and content of our dictionary to include various other language entities, such as word variations and abbreviations by applying semantic orientation (Turney, 2002). Our dictionary will not only provide the English translation for Thai toxic words, but also examples for each word. We hope to enlarge our corpus with this new dictionary to make it a sufficient and reliable resource for Thai language analysis in the future. Finally, we might consider using other content such as re-tweets or previous conversations to provide a better understanding regarding the intentions of the messages in a future study.

7. Acknowledgements

This research was (partly) supported by Grant-in-Aid for Research on Priority Areas, Tokyo Metropolitan University, Research on social bigdata.

8. Bibliographical References

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1217–1230, Portland, OR, USA, February. Association for Computing Machinery.

Cooper, D. (1996). Ambiguous (((Par(t)(it))((ion))(s))(in)) Thai Text. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, pages 109–118, Seoul, South Korea, December. Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International Conference on Web and Social Media*, Montreal, Canada, May. Association for the Association for the Advancement of Artificial Intelligence.

Kennedy, G., McCollough, A., Dixon, E., Bastidas, A., Ryan, J., Loo, C., and Sahay, S. (2017). Technology Solutions to Combat Online Harassment. In *Proceedings of the First Workshop on Abusive Language Online*, pages 73–77, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Kolhatkar, V. and Taboada, M. (2017). Constructive Language in News Comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Kwak, H. and Blackburn, J. (2014). Linguistic Analysis of Toxic Behavior in an Online Video Game. In *Pro-*

ceedings of the 1st Exploration on Games and Gamers Workshop, EGG 2014.

Mekthawornwathana, T. (2011). The Factors used for Identifying "Politeness" in Male and Female Conversations among Thai Undergraduate Students. *NIDA Development Journal*, 51(3):142–166.

Nand, P., Perera, R., and Kasture, A. (2016). "How Bullying is this Message?": A Psychometric Thermometer for Bullying. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 695–706, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Orathai Chinakarapong. (2014). Conceptual Metaphor of Thai Curse Words. *Journal of Humanities Naresuan University*, 11(2):57–76, August.

Razavi, A., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive Language Detection Using Multi-level Classification. In *Proceedings of the 23rd Canadian conference on Advances in Artificial Intelligence*, pages 16–27, Ottawa, Canada, June. Canadian Conference on Artificial Intelligence 2010.

Saleem, H. M., Dillon, K. P., Benesch, S., and Ruths, D. (2016). A Web of Hate: Tackling Hateful Speech in Online Social Spaces. In *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016)*, Portorož, The Republic of Slovenia, May.

Turney, P. (2002). Thumbs Up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., and Hoste, V. (2015). Detection and Fine-grained Classification of Cyberbullying Events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria, September. Association for Computational Linguistics.

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., and Cranor, L. F. (2011). "I regretted the minute I pressed share": A Qualitative Study of Regrets on Facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, page 10. Association for Computing Machinery.

Waseem, Z. and Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399, Perth, Australia, April. International World Wide Web Conference 2017.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. (2009). Detection of Harassment on Web 2.0. In *Proceedings of the Content Analysis in the WEB*, volume 2, pages 1–7, Madrid, Spain, April. International World Wide Web Conference 2009.

Zhou, N., Aw, A., Lertcheva, N., and Wang, X. (2016). A Word Labeling Approach to Thai Sentence Boundary Detection and Pos Tagging. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 319–327, Osaka, Japan, December. The COLING 2016 Organizing Committee.