STUDENT NO. 17890522

Master's Thesis

Annotation and Classification of Toxicity for Thai Twitter

Sugan Sirihattasak

February 22, 2019

Graduate School of System Design
Tokyo Metropolitan University

A Master's Thesis
submitted to Graduate School of System Design,
Tokyo Metropolitan University
in partial fulfillment of the requirements for the degree of
MASTER of SYSTEM DESIGN

Sugan Sirihattasak

Thesis Committee:

Associate Professor Mamoru Komachi	(Supervisor)
Professor Hiroshi Ishikawa	(Co-Supervisor)
Associate Professor Shohei Yokoyama	

Annotation and Classification of Toxicity for Thai Twitter*

Sugan Sirihattasak

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. Moreover, 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. we also estimate semantic orientation Turney [1] of words to rank words according to toxicity. As the result, we got precision@k for 0.58@40 and 0.41@80. Finally, we launched our demo application for the public feedback and our dataset is available on GitHub.

Keywords:

toxicity, corpus, Thai, Twitter

*Master's Thesis, Graduate School of System Design System Design,
Tokyo Metropolitan University, Student No. 17890522, February 22, 2019.

Contents

List of Figures

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 more difficult to control. In addition, the victims of toxic messages are not always the intended targets of those messages. According to Wang et al. [2], many people regret their negative posts because of problems they face later, such as being terminated from employment or losing other opportunities. 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 little possibility of deleting it and undoing its effects.

Consequently, there have been many research efforts in various fields, such as the social sciences, psychology, and natural language processing, to improve the quality of online conversion while considering the right to freedom of speech.

One of the challenges of studying toxicity in online communication is arriving at a clear common definition of toxicity of language. Toxic comments are often sarcastic and indicate aggressive disagreement; in Kolhatkar and Taboada [3], the relationship between constructiveness and toxicity, including in comments on news stories, was studied. In our study, we define toxicity from a more general perspective to include any messages that can imply toxic behavior [4], antisocial behavior [5], or online harassment [6]; hate speech [7]; or cyberbullying [8], or any type of offensive language [9]. In particular, a toxic message is any message that may hurt or harm an individual or a generalized group, challenge societal norms, or negatively affect the entire community. As toxic words, we consider any negative words such as those associated with profanity and obscenity, or those which are offensive.

Though there has been an increase in studies related to toxicity, freely available resources related to it are still limited. There are several corpora for major languages like English, including a harassment dataset [10], a hate speech Twitter annotation corpus [11], and a personal attacks comment corpus [12]. Unfortunately, studies related to this topic do not yet 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 the Thai language is complicated due to ambiguity of segmentation in the written language [13]; for example, without segmentation, "ปลาตากลมตัวนี้น่ารัก (This round-eyes (ตา ∣ กลม) fish is cute.)" reads much like "ขอเดินออกไปตากลม (Let me go out to have some fresh air (ตาก ∣ ลม))." Likewise, sentence boundary detection is difficult [14] because the space used for differentiating sentences is not appropriate in some cases, such as in "โอ๊ย! เจ็บ (Ouch! it hurts)."

For the above reasons, we present an annotation and classification of toxicity on Twitter in the Thai language as a preliminary exploration to support further toxicity analysis in the Thai language in general.

The main contributions of this study are as follows:

1. We constructed a dictionary of Thai toxic words, which we use as keywords for the annotation.
2. We built a toxicity corpus using Twitter messages.
3. We used our abovementioned dataset to conduct supervised classification and obtained an accuracy of 0.80.
4. We applied semantic orientation Turney [1] in order to extend our dictionary.
5. We provided public access to the demo application.

Our dictionary and corpus are available on GitHub*.

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 shows the construction of our dictionary and Section 4 presents the

*https://github.com/tmu-nlp/ThaiToxicityTweetCorpus/

corpus analysis. Section 5 reports the experimental results of supervised classification using our dictionary and corpus. Section 6 is about the demo application. Finally, Section 7 presents the conclusions of our study and indicates the scope of 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 acquaintances, especially close friends, to casually communicate each other using potentially toxic words without intending any harm [15]. Likewise, the factors used to identify politeness in Thai male conversation depend on situational context, such as the relationship between the speaker and listener or the location at which the conversation takes place, rather than strictly linguistic aspects [16]. Moreover, the keyword-based approach does not seem flexible enough for a non-segmenting language like Thai. The following two examples contain a toxic word "หอก*" (The original meaning is "spear"; however, the slang meaning is an insulting phrase, roughly "Damn, bitch.")

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

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

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 Thai because that may lead to ambiguity about what word is meant.

Thus, 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 examples presented below.

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

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

3. เหี้ยเป็นสัตว์เลื้อยคลาน
   เหี้ย | เป็น | สัตว์เลื้อยคลาน
   varanus salvator | is | 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 [17]; we do this by creating a corpus that contains a variety of examples of toxicity in the Thai language.

3. Keyword Dictionary Construction

3.1. 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 Conceptual Metaphor of Thai Curse Words [18] and rechecked spelling using the (Thai) Royal Institute Dictionary*. 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 4.1.

In order to calculate semantic orientation of words, we formed a new dictionary of 44 positive words, which are contrary in valence to the toxic ones and are often used in encouragement and compliment. The list is as below.
ดี (good), สวย (pretty), หล่อ (handsome), รัก (love), เก่ง (skillful), น่ารัก (cute), สุข (happy), สบาย (comfortable), งาม (pretty), ใจดี (kind), อ่อนโยน (mild), สุภาพ (gentle), เห็นใจ (sympathy), ปลื้ม (overjoy), ชอบ (like), สนุก (fun), พอใจ (favor), ประเสริฐ (sublime), ฉลาด (clever), สูง (high), เลิศ (great), อุ่น (warm), มิตร (friend), นับถือ (respect), ใส่ใจ (considerate), สุดยอด (supreme), หลง (passionate), ประทับใจ (impressive), ปกป้อง (protect), สนับสนุน (support), นิยม (adore), กำลังใจ (encourage), อภัย (forgive), ชื่นชม (praise), ความหวัง (hope), ห่วงใย (care), ศรัทธา (faith), เข้มแข็ง (strong), แข็งแกร่ง (sturdy), กล้า (brave), สว่าง (bright), สำเร็จ

*http://www.royin.go.th/dictionary

(success), สดใส (cheerful), ฝัน (dream)

3.2. Semantic Orientation

Our initial dictionary is still small and does not contain Twitter-specific keywords, so that it is insufficient for downstream applications. In order to extend our dictionary and adapt it to this domain, we estimate the semantic orientation [1] of words to rank them according to toxicity. The dataset includes 175,366 Thai tweets from January to August 2018, which were collected by using Twitter Search API without specifying keywords. Then, we rank extracted words based on pointwise mutual information, as below.

$$PMI(word_1, word_2) = \log \frac{p(word_1 \& word_2)}{p(word_1)p(word_2)}$$

$$SemanticOrientation(word) = \sum_i PMI(word, toxic_i) - \sum_j PMI(word, positive_j)$$

Due to errors in auto-tokenization, which we will discuss in the Section 5, we did some adjustment such as excluding stopwords, named entities, phrases, and unmeaningful words from the ranking manually.
As shown in A.1, we found 6 already identified toxic keywords (bold text) and 3 new alternative toxic words (text in '[]'). The top-10 in semantic orientation for toxic words are shown in Table 3.1. Besides, semantic orientation scores for positive words are shown in A.2. Furthermore, we used precision@k to learn how relevant the results were, and got 0.58@40 and 0.41@80.

Table 3.1: Top-10 in semantic orientation for toxic words.

Thai word	English Meaning	Score	Is this word toxic or likely for offensive use?
กู	impolite form of I	54.8	✓
ด่า	damn	42.0	✓
คาว	fishy	38.6	✓
มึง	impolite form of you	33.7	✓
555	laugh sound	31.9	

ขี้	defecate	31.5	✓
ประเภท	type	28.5	
ควย*	genitalia/ f*ck	28.3	✓
ฉีด	inject	27.6	
กิน	eat	25.6	

4. Corpus

4.1. Dataset Construction

We used the public Twitter Search API to collect 9,819 tweets from January–December 2017 that contained one or more of the 44 toxic keywords in our dictionary. From those, 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 would also select "บ้านนอก" (countryside) which was not a target word.)

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. 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.

5. 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.

4.2. Annotation

We manually annotated our dataset with two labels: Toxic and Non-Toxic. We define a message as toxic if it indicates any harmful, damaging, or negative intent, based on our definition of toxicity given above. Furthermore, all the tweets were annotated for toxicity by three annotators; the conditions they 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 must have a target or consequence. It can either be an individual or a generalized group based on a commonality such as religion or ethnicity, or an entire community.
- Self-complaint 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 toxic.
- Both direct and indirect messages, including those using sarcasm, are taken into consideration.

We carefully instructed all the candidate 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 annotators who obtained a full score to serve as annotators for the actual annotation in the second round. From among the candidate annotators, 20% failed the first round and were not involved in the final annotation.

4.3. Corpus Analysis

As previously mentioned, the corpus consists of 3,300 tweets, divided into 2,027 toxic tweets and 1,273 non-toxic tweets; these labels are assigned based on majority decisions. There were 1,692 toxic and 1,093 non-toxic tweets that were

considered "gold standard tweets," entailing perfect agreement among all raters. Overall, inter-annotator agreement (Fleiss' Kappa) [19] is 0.78, which shows that the agreement was significant. There were three primary reasons for disagreement. First, more than 35% of tweets that annotators disagreed upon are difficult to judge 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 covertly, to avoid defamation. Last, 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," can imply that the writer 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 4.1. Interestingly, the tweets that contain toxic words and are 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 a dish made from horseshoe crabs that is popular in Thailand, while "ควาย" seems to be rarely used for its literal meaning of buffalo. Moreover, words related to physical bottomness or lowness, 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; the head, for example, is treated with special respect. The word "โง่" (stupid) seems to be used in a non-toxic manner rather than for toxic purposes; in the corpus data, 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 to toxic contexts, but is used in everyday conversation, for example in teaching or reporting a situation, as well. Finally, the word "สัตว์" (animal) is commonly used for its original non-toxic meaning. This is in contrast to 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

Table 4.1: 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

toxicity.

Some tweets are difficult to label, leading to inconsistency in annotation as shown in Table 4.1. Moreover, Thai people often use metaphors in their conversations, as indicated in the example below.

กินกะหรี่ปั๊บอร่อยไม่เหมือนกินกะหรี่

กิน ׀ กะหรี่ปั๊บ ׀ อร่อย ׀ ไม่ ׀ เหมือน ׀ กิน ׀ กะหรี่

eat | curry puff | yummy/delicious | not | similar to | eat | curry? whore?

Eating curry puff is yummy not like eating curry (whore?).

In such cases, it is difficult to ascertain the meaning of the 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 on Thai Twitter because people avoid mentioning the target of the message directly to prevent legal repercussions or other issues.

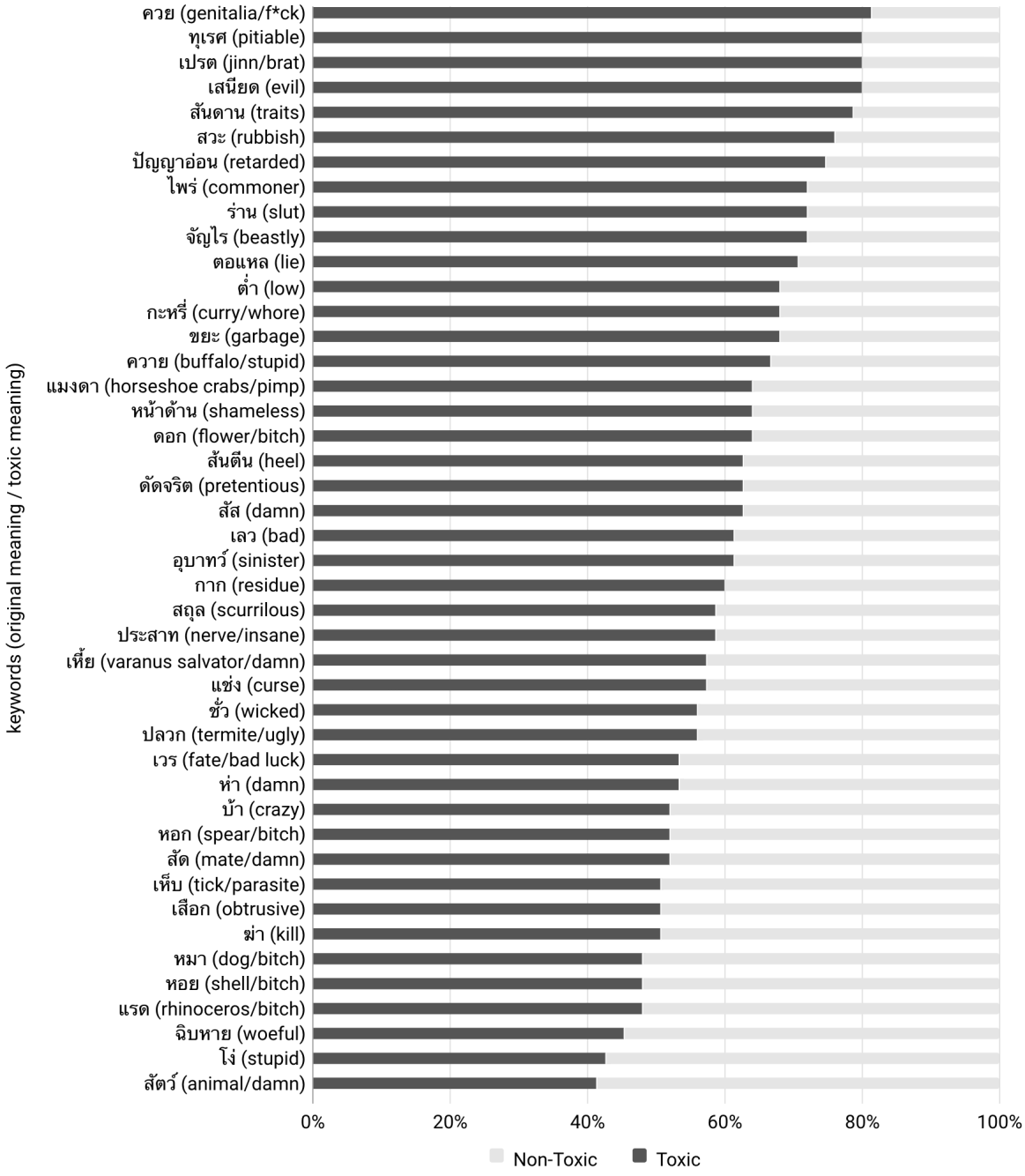

Figure 4.1: Distribution of toxic and non-toxic tweets based on keywords.

5. Classification

5.1. Train and Test Dataset

Aside from the steps performed for annotation, we conduct supervised classification using the dictionary and corpus constructed in this study. First, we conduct further tweet data cleaning after segmenting the tweets into tokens using Deepcut (library version) 0.6*.

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

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

3. We removed non-Thai words.

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

*https://github.com/rkcosmos/deepcut

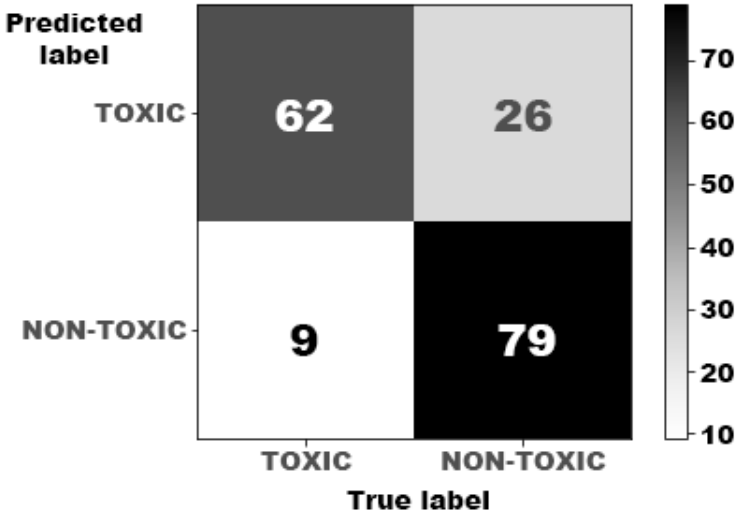

Figure 5.1: Confusion matrix of toxicity classification.

5.2. Experimental Settings

For classification, we use the CountVectorizer method from the scikit-learn library version 0.19[†] 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 GridSearchCV, as follows.

1. C value: 0.001, 0.01, 0.1, 1, 10.
2. Fit intercept: True or False.
3. Penalty: L1 or L2.

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

[†]`https://github.com/scikit-learn/scikit-learn`

Table 5.1: 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

Table 5.1 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 5.1. Compared with the keyword baseline method, our classification results are better in terms of precision and F1-score.

Table 5.2: Examples of false positives.

Tweet text (English translation)	Toxic keyword	True label	Predicted label
Since this morning, the dormitory internet isdamn 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 5.3: 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

5.4. Discussion

Although the keyword-based approach is popular for performing this type of classification, it failed to correctly classify some tweets, as in the following translated example, which it labeled toxic: "Damn, just finished laundry and it's raining." In contrast, our approach correctly classified this example 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 5.3. 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-frequency words in our model.

Examples of false positives are shown in Table 5.3. Because our corpus is small, its surface features are insufficient to properly cover 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 acknowledge that the auto-segmentation is not perfect, which affects the classification. For example, a tweet that includes incorrect 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.
Despite some errors, our auto-segmentation method is quite effective (considerably more than alternatives), as seen in the examples below.

1. ถึงคุณรวยล้นฟ้าแต่ไร้น้ำใจก็ยากที่คนจะศรัทธา (Despite being a millionaire, without kindness, nobody will respect you.) Auto-segmentation and human-segmentation are the 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).

2. คนเห็นแก่ตัวที่ไม่เคยเห็นใจคนอื่น (A selfish person who never cares 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) ׀ เห็นใจ (sympathetic) ׀ คนอื่น (others)

6. Demo

The computed model we created in the previous section was implemented in our demo application*. This demo is the first public application related to toxicity in the Thai language. For example, when we put a message: “ไอ้ห่า! ไปตายซะ! เกลียดมากคนไร้ประโยชน์" which means “Damn you! Just go die! I hate useless person so much," the result is “Toxic" as shown in Figure 6.1.
Moreover, we checked the performance of our model by auto-labeling to unannotated tweets without using any keywords. There were 14,697 tweets labeled toxic and 160,669 labeled non-toxic. From this result, we sampled 50 toxic tweets and 50 non-toxic tweets randomly, and evaluated accuracy using 2 annotators. The results shows 56 correctly evaluated tweets (non-toxic: 50 tweets, toxic: 6 tweets). All the wrong tweets were non-toxic tweets labeled toxic. Due to the small size of the training corpus, it is difficult to cover various conversation topics.

*http://cl.sd.tmu.ac.jp/thaitoxicity/

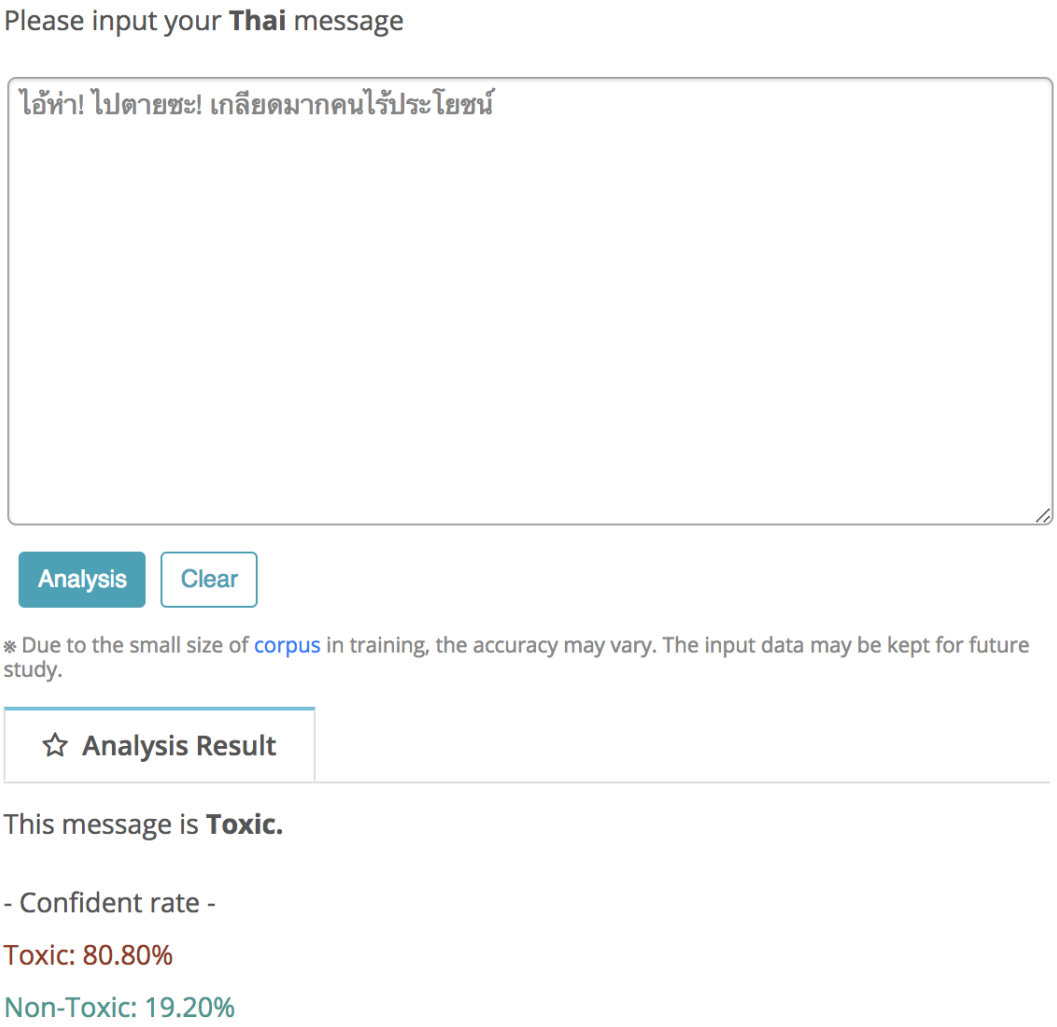

Figure 6.1: Demo Application of our computed model.

7. Conclusion

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, however, 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 future more extensive toxicity analysis in the Thai language. The corpus was formed using 44 toxic keywords. However, the present corpus is insufficient for various topics. Thus, we applied semantic orientation to expand the dictionary keywords, conducting supervised classification method using our dictionary and corpus and obtaining an accuracy of 0.80, which is comparable with the inter-annotator agreement on this dataset. Finally, we create a demo, which is the first public application related to toxicity in the Thai language.

In the future, we hope to create a sufficient, reliable resource for Thai language analysis by using other content such as re-tweets or previous to provide a better understanding regarding the intentions of the messages.

A. Keyword Dictionary Construction

A.1. Semantic orientation score for toxic words.

Table A.1: Semantic orientation score for toxic words.

Thai word	English Meaning	Score	Is this word toxic or likely for offensive use?
กู	impolite form of I	54.8	✓
ด่า	damn	42.0	✓
คาว	fishy	38.6	✓
มึง	impolite form of you	33.7	✓
555	laugh sound	31.9	
ขี้	defecate	31.5	✓
ประเภท	type	28.5	
ควย*	genitalia/ f*ck	28.3	✓
ฉีด	inject	27.6	
กิน	eat	25.6	
ชาบู	shabu	25.2	
ตลก	joke	23.3	
จัญไร*	beastly	23.0	✓
จอดำ	black screen/ involuntarily shutdown (slang)	22.8	✓
ตาย	die	22.1	✓
ตายโหง	die unnaturally	21.8	✓
นอน	sleepy	21.6	
สันดาน*	traits	21.5	✓

ตบ	slap	21.4	✓
ญ	abbreviation of woman	20.8	
ตอแหล*	lie	20.2	✓
กู	impolite form of I	20.1	✓
ผี	ghost	20.0	✓
นั่ง	sit	20.0	
พวก	gang	19.9	
ปวด	hurt	19.6	
middle_finger	(emoticon)	19.4	✓
กลัว	fear	19.2	
บังคับเลิก	force to quit	19.1	✓
ลูกสมุน	underling	19.1	
ติด	attach	19.0	
มึง	impolite form of you	18.7	✓
หยาบ	rude	18.7	✓
คดโกง	cheat	18.6	✓
เสือก*	obtrusive	18.5	✓
กัก	confine	18.5	✓
ขาด	lack	18.2	
ซื้อ	buy	18.2	
พงไพร	forest	18.1	
นรก	hell	17.8	✓
อีจิ้งจก	lizard	17.7	✓
สัตว์*	animal/damn (slang)	17.5	✓
ยังค่ำ	evening	17.5	
วัง	palace	17.4	
บ้าน	house	17.4	
สุรา	booze	17.3	
[สุดตีน*]	feet/damn (slang)	17.3	✓
คะแนน	score	17.2	
ตู้	cabinet	17.2	
ตั้ง	set	17.1	
กลับ	return	17.0	

กรรม	karma	17.0	
unamused_face	(emoticon)	17.0	✓
ขบขัน	funny	16.7	
รถ	car	16.7	
ทอปฟอร์ม	top form	16.7	
อีคน	impolite form of human	16.7	✓
ยอม	surrender	16.5	
นักหนา	much	16.4	
หน้าตลก	funny face	16.4	
ดึก	late night	16.4	
จุด	spot	16.4	
พันธุ์	breed	16.3	
มัน	it	16.2	✓
ฉี่	peep	16.1	✓
ลิงค์	link	16.0	
เพรส	place	15.9	
ชและญ	man and woman	15.9	
[สัต*]	animal/damn (slang)	15.8	✓
หยาบสุด	the rudest	15.7	
พ้อง	same	15.6	
ร่าง	body	15.5	
สอบ	exam	15.5	
ง่วง	sleepy	15.4	
เซเลป	celebrity	15.4	
ลอยนวล	unleash (negative sense)	15.3	✓
พ่อง	damn	15.3	✓
สะใจ	satisfy (negative sense)	15.3	✓
ฉีก	rip	15.2	
กลาย	become	15.2	
คัน	itchy	15.2	
อ่อย	flirt	15.2	
กะลาแลนด์	foolish country (slang)	14.9	✓
คลาสพิเศษ	special class	14.8	
คณิต	math	14.8	

ฝน	rain	14.7	
หลุดปาก	make a slip	14.5	
กฎหมาย	law	14.5	
สนิท	close to	14.4	
คุก	jail	14.3	
อนาคต	future	14.2	
ขนลุก	gooseflesh	14.2	
ปัญญา	intelligent	14.2	
[อีเหี้ย*]	damn	14.1	✓
ทราม	bad	14.1	✓
จุดจบ	ending	14.1	
กลุ่มไลน์	group	14.1	
ทะเลเน็ต	network	13.9	
มหาลัย	university	13.9	
คบ	dating	13.9	

A.2. Semantic orientation score for positive words.

Table A.2: Semantic orientation score for positive words.

Thai word	English Meaning	Score
ขอบคุณ	thanks	45.19
thumbs_up	(emoticon)	37.43
growing_heart	(emoticon)	31.98
ดูดี	look good	31.87
ท่าน	respect form of "you"	31.22
นิ่ม	soft	30.66
คอนโซล	console	30.49
กำลังใจ	encouragement	30.36

กตัญญู	gratitude	30.12
lion_face	(emoticon)	29.46
ตะวัน	sun	29.33
กันและกัน	together	28.72
ชื่นชอบ	like	28.38
พรม	splash/carpet	27.48
ชุมชน	community	26.38
folded_hands	(emoticon)	26.27
คอย	wait	26.03
light_skin_tone	(emoticon)	25.74
ชื่นชม	admire	25.65
น่ารัก	cute	25.52
ฟุตบอล	football	25.40
face_blowing_a_kiss	(emoticon)	25.33
heart_suit	(emoticon)	25.19
พลัง	power	25.18
ชุ่มชื่น	fresh	24.86
ยิ้ม	smile	24.77
จินตนาการ	imagine	24.67
smiling_face_with_smiling_eyes	(emoticon)	24.28
จนหลง	be enchanted	24.28
ชนะ	win	24.15
คุณ	you	23.83
กันเอง	intimately	23.53
วัย	age	23.44
ขนมดี	nice dessert	23.40
พลังใจ	spirit	23.35
ผิว	skin	23.02
ปีน	climb	22.99
คึกคัก	lively	22.98
นุ่ม	soft	22.96

ย้อน	return	22.79
musical_note	(emoticon)	22.73
medium-light_skin_tone	(emoticon)	22.67
ขอ	ask	22.53
นุ้งหมา	puppy	22.46
sparkling_heart	(emoticon)	22.37
ติ๊	he	22.11
ทุ่มเท	devote	22.10
บาน	bloom	22.08
ขี้เล่น	playful	22.05
สนาน	fun	21.95
ระบาย	let it out	21.94
ดารา	actor/actress	21.91
รองเท้า	shoes	21.88
คริสตัล	crystal	21.87
ครอบครอง	possess	21.82
ขยาย	extend	21.77
งาม	beautiful	21.77
musical_notes	(emoticon)	21.71
ต่างๆ	etc	21.60
red_heart	(emoticon)	21.57
บุรุษ	gentleman	21.48
สดใส	cheerful	21.24
ศิลปะ	art	21.07
ร่ม	umbralla/shady	20.96
sheaf_of_rice	(emoticon)	20.78
birthday_cake	(emoticon)	20.65
ปรับ	adjust	20.61
ชมภาพ	watch	20.51
คว้า	take	20.51
มหาศาล	a lot	20.51

ตื้นตัน	speechless	20.48
กว้างใหญ่	big	20.43
สามี	husband	20.39
radio	(emoticon)	20.21
ผู้คน	crowd	20.21
ยิ่งใหญ่	mighty	20.16
camera_with_flash	(emoticon)	20.09
กะเผลก	stumblingly	20.09
ม๊าก	a lot	20.04
คูล	cool	20.01
ชุ่ม	moist	19.94
ค่อนข้าง	rather	19.94
ย้อนยุค	retro	19.84
กะหล่ำปลี	cabbage	19.83
smiling_face_with_heart-eyes	(emoticon)	19.81
woman	(emoticon)	19.81
ทักทาย	greet	19.76
ขี้งอน	touchy	19.65
น้ำตาล	sugar	19.61
ละลาบละล้วง	trespass	19.60
ฟอลโลวเวอร์	follower	19.60
blue_heart	(emoticon)	19.52
ขอบ	edge	19.51
กาย	body	19.41
ประกอบ	consist of	19.31
บรรยาย	explain	19.29
ขอบพระคุณ	thanks	19.27
อ่อนเยาว์	young	19.27
man	(emoticon)	19.20
มุม	viewpoint	19.17

Bibliography

[1] P. Turney, "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*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 417–424. [Online]. Available: http://www.aclweb.org/anthology/P02-1053

[2] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor, ""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*. Association for Computing Machinery, 2011, p. 10.

[3] V. Kolhatkar and M. Taboada, "Constructive Language in News Comments," in *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics, August 2017, pp. 11–17. [Online]. Available: http://www.aclweb.org/anthology/W17-3002

[4] H. Kwak and J. Blackburn, "Linguistic Analysis of Toxic Behavior in an Online Video Game," in *Proceedings of the 1st Exploration on Games and Gamers Workshop, EGG 2014*, 2014.

[5] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "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*. Portland, OR, USA: Association for Computing Machinery, February 2017, pp. 1217–1230.

[6] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0," in *Proceedings of the Content Analysis in the WEB*, vol. 2. Madrid, Spain: International World Wide Web Conference 2009, April 2009, pp. 1–7.

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

[8] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Detection and Fine-grained Classification of Cyberbullying Events," in *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar, Bulgaria: Association for Computational Linguistics, September 2015, pp. 672–680. [Online]. Available: http://www.aclweb.org/anthology/R15-1086

[9] A. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive Language Detection Using Multi-level Classification," in *Proceedings of the 23rd Canadian conference on Advances in Artificial Intelligence*. Ottawa, Canada: Canadian Conference on Artificial Intelligence 2010, June 2010, pp. 16–27.

[10] G. Kennedy, A. McCollough, E. Dixon, A. Bastidas, J. Ryan, C. Loo, and S. Sahay, "Technology Solutions to Combat Online Harassment," in *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics, August 2017, pp. 73–77. [Online]. Available: http://www.aclweb.org/anthology/W17-3011

[11] Z. Waseem and D. Hovy, "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*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 88–93. [Online]. Available: http://aclanthology.coli.uni-saarland.de/pdf/N/N16/N16-2013.pdf

[12] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina: Personal Attacks Seen at Scale," in *Proceedings of the 26th International Conference on World Wide Web*. Perth, Australia: International World Wide Web Conference 2017, April 2017, pp. 1391–1399.

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

[14] N. Zhou, A. Aw, N. Lertcheva, and X. Wang, "A Word Labeling Approach to Thai Sentence Boundary Detection and Pos Tagging," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*. Osaka, Japan: The COLING 2016 Organizing Committee, December 2016, pp. 319–327. [Online]. Available: http://aclweb.org/anthology/C16-1031

[15] P. Nand, R. Perera, and A. Kasture, ""How Bullying is this Message?": A Psychometric Thermometer for Bullying," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, December 2016, pp. 695–706. [Online]. Available: http://aclweb.org/anthology/C16-1067

[16] T. Mekthawornwathana, "The Factors used for Identifying "Politeness" in Male and Female Conversations among Thai Undergraduate Students," *NIDA Development Journal*, vol. 51, no. 3, pp. 142–166, 2011.

[17] H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths, "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 2016.

[18] Orathai Chinakarapong, "Conceptual Metaphor of Thai Curse Words," *Journal of Humanities Naresuan University*, vol. 11, no. 2, pp. 57–76, August 2014.

[19] J. Carletta, "Assessing Agreement on Classification Tasks: The Kappa Statistic," *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, 1996. [Online]. Available: http://aclanthology.coli.uni-saarland.de/pdf/J/J96/J96-2004.pdf

Publication List

[1] S. Sirihattasak, M. Komachi, H. Ishikawa, "Annotation and Classification of Toxicity for Thai Twitter," in *Proceedings of Second Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2018)*, 2018