Identification of Risk of Cyberbullying from Social Network Messages

Thitiphan Semangern, Wachit Chaisitsak, and Twittie Senivongse

***Abstract*—Social networks have revolutionalized how people communicate. People can create content while others can share, react to, or express opinions about the content. Since communication can take place more freely and more anonymously, social network users are more vulnerable to cyberbullying. The problem is a threat to the victim's several mental and physical health conditions such as low self-esteem, depression, drug and alcohol abuse, and suicidal thoughts. In Thailand, cyberbullying is a widespread problem as the number of incidents is among the top lists in the world. To increase awareness and try to prevent the problem, this paper presents a method to identify risk of cyberbullying to an individual via an analysis of social network messages in Thai. The presented method collected training data from Twitter and used several machine learning algorithms to classify textual tweets into four cyberbullying categories, i.e. sexual harassment, insult and threat, race and religion, and intelligence, appearance, and social status. An experiment showed that Logistic Regression performed best when the problem of imbalanced data set was handled with the F1 score of 73.89 and accuracy of 73.61. In addition, a tool to visualize cyberbullying incidents of any individual has been implemented. The method and accompanying tool can help to monitor potential risk of cyberbullying to an individual so that appropriate care can be sent in a timely manner.**

***Index Terms*—cyberbullying, machine learning, social networks, text classification, Twitter**

I. Introduction

SOCIAL networks have revolutionalized how people communicate as the places for people to engage in social interaction by posting content, sharing information, and expressing opinions. Using their digital devices, people can communicate freely, anytime anywhere, and anonymously. Reaching the large audience, information and opinions posted on social networks may cause negative impact on another person who is the subject of the post. The person could become the victim of cyberbullying. The term "cyberbullying" is conceptualized around the definition of traditional bullying, but the activities are through electronic means, specifically through mobile phones, tablets, computers, and the Internet. Smith et al. [1] define cyberbullying as "an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself". Menesini et al. [2] discuss the criteria that define the term cyberbullying in its virtual context, i.e. 1) *intentionality* – The bully intends to hurt, harm, or make negative impact on the victim, 2) *repetition* – The victim is bullied repeatedly over time, not necessarily by a single bully, 3) *imbalance of power* – The victim is overpowered by the bully and feels powerless and defenceless in the face of attack, 4) *anonymity* – The possible anonymity of the bully may intensify negative feelings in the victim such as defenceless, and 5) *public versus private* – The victim may consider the attack as more serious when it is made public because of the potentially large audience.

The common places where cyberbullying occur are social networks (e.g. Facebook, Instagram, and Twitter), text messages and social media messages sent through devices, and emails [3]. Cyberbullying has unique characteristics that make it have a great impact on the victim [3], i.e. 1) *persistent* – With widespread use of digital devices, it is more convenient for the bully to reach the victim any time of the day, unlike traditional bullying, so it can be hard for the victim to find relief 2) *permanent* - Most electronic information made by the bully is permanent and public, if not reported and removed, and it can impact several areas of the victim's life, and 3) *hard to notice* – It might be hard to recognize a cyberbullying case as it might not be seen or heard by those who can give help, when it occurs.

Cyberbullying incidents have been reported worldwide. For example, US federal surveys [3] reported that 15% of students ages 12-18 who were bullied during the school year were bullied online or by text. In addition, 14.9% of high school students were cyberbullied in the past 12 months before the survey. Athanasiou et al. [4] reported that cyberbullying is subject to country-specific socio-demographic factors and patterns of current Internet use and its development. In their research on students ages 14-17 in seven European countries, Romania had the highest proportion of youth that had been bullied online in past 12 months (37.3%) whereas Spain had the lowest proportion (13.3%). In some countries like Romania, Poland, and Germany, the duration of social network use was associated with cyberbullying victimization. In Thailand, cyberbullying is widespread although research on this topic is relatively limited. A study of Pornnoppadol et al. [5], [6] on the

Manuscript received June 28, 2019; revised July 30, 2019.

T. Semangern is with the Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, 10330, Thailand (email: thitipanjen@icloud.com).

W. Chaisitsak is with the Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, 10330, Thailand (email: ruby.pwn@hotmail.com).

T. Senivongse is with the Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand (corresponding author phone: +66 2 2186996; fax: +66 2 2186955; e-mail: twittie.s@chula.ac.th).

prevalence and related factors of cyberbullying reported that 45% of Thai adolescents were cyberbullied at least once. This percentage is four times higher than that of the US, Europe, and Japan. Sittichai [7] surveyed students ages 14-17 in southern Thailand and reported that 14.9% were cyberbullied once or twice and 3.7% were cyberbullied 2 or 3 times a month or more. Cyberbullying victimization in males was higher than that in females but it was not due to the frequency of their Internet use, as females used the Internet more frequently. However male students spent more time surfing the net and playing games, whereas female students used the Internet more for schoolwork and Facebook. The nature of the Internet use could be a factor for the opportunities for cyberbullying, rather than the frequency or activeness of use. In addition, about half of all cyberbullied victims were not bullied in a traditional way, and this suggests the importance of cyberbullying victimization in the Thai context.

Songsri and Musikaphan [8] reported seven categories of cyberbullying which were experienced by Thai youth in Bangkok, i.e. 1) the use of messages with angry and vulgar language, 2) repeatedly posting mean, nasty, and insulting messages, 3) talking someone into revealing secrets or embarrassing information, then sharing it online, 4) sending gossip or rumors about a person to damage his/her reputation or friendships, 5) pretending to be someone else and sending material to get that person in trouble or to damage that person's reputation or friendships, 6) repeated intense harassment and denigration including threats, and 7) intentionally and cruelly excluding someone from an online group. Of these seven categories, the first two were the most common. Cyberbullying is a threat to the victims' mental and physical well-being [9]. For example, the victims may have trust issues or have more trouble getting along with others. They may abuse alcohol or drugs or have frequent headaches and stomach pain due to nervousness. They may also turn into self-harm. Long-term effects include low self-esteem that may cause fatigue, insomnia, and poor performance at work or in school. The victims may have depression, the feeling of worthlessness about their lives, and suicidal thoughts. Several cases of suicide have been reported [10]. For example, a boy killed himself after being called a retard. Another boy who was a cheerleader attempted a suicide because his friends called him a hermaphrodite. A young man jumped off a building because he was teased by friends as being fat. Although cyberbullying is often associated with children in school, the attack can happen to anyone. Incidents involving celebrities are common. An actress committed suicide after being called too ugly [11], while other celebrities who were attacked for their weight, race, sexual preference, or disappointing performance had to leave their social network accounts for a period of time, check themselves into a rehab, or even take a legal action against the bullies [12].

Due to the importance of the problem, this paper presents a method to identify risk of cyberbullying to an individual via an analysis of Twitter messages in Thai. As cyberbullying in the written form is the most common in the Thai context [8], the method uses text classification [13] to classify tweets about an individual into four categories, i.e. sexual harassment, insult and threat, race and religion, and intelligence, appearance, and social status. Several machine learnings algortihms are used including Multinomial Naive Bayes (MultinomialNB), Linear Support Vector Classification (LinearSVC), Random Forest, and Logistic Regression [13]. Unlike other research that also captures the bullying intention in any single post on a social network, this paper also considers the repetition criterion and tries to make the attack more noticeable. That is, the method is accompanied by a tool to visualize cyberbullying incidents of an individual over a period of time. The method and tool can help to monitor potential risk of cyberbullying to an individual and to determine how repetitive the attack is, so that appropriate actions can be taken.

The rest of the paper is organized as follows. Section II discusses related work. The dataset construction is described in section III and the experiment in section IV. Section V presents the tool, and the paper concludes in section VI.

II. Related Work

Researchers address the cyberbullying problem in social networks and use machine learning approaches to detect cyberbullying incidents. Van Hee et al. [14] used text classification to classify social network messages written in Dutch and English. The training data were collected from ASKfm. The messages were classified into fine-grained categories with regard to types of cyberbullying and roles in cyberbullying. They were threat/blackmail, insult (i.e. general insult, attacking relatives, discrimination), curse/exclusion, defamation, sexual talk, defense (i.e. bystander defense, victim defense), encouragement to the harasser, and other form of cyberbullying. Combinations of features for classification included n-gram bag of words, sentiment, specific term lists, and topic model features. Binary classification experiments using linear SVM were performed, and the maximum attained F1 scores were 64.26% for English and 61.20% for Dutch. Al-garadi et al. [15] detected cyberbullying (i.e. cyberbullying, non-cyberbullying) from Twitter messages in English. The features for classification included network (e.g. no. of followers and following), activity (e.g. no. of posted tweets and mentions), user (i.e the use of neurotic-, gender-, and age-related terms), and tweet content (i.e. the use of profane words, cyberbullying-related slangs, and first and second person pronouns). They experimented with NB, SVM (LibSVM), decision trees (DT), (random forest), and KNN algorithms. They considered feature selection, using SMOTE to handle imbalanced data set, and cost-sensitive methods, but the best classifier was random forest with SMOTE with the F1 score of 93.6%. Zhao et al. [16] proposed a representation of Twitter messages in English for binary classification (i.e. bullying, non-bullying). The tweet representation consisted of bag of words features, latent semantic features, and bullying features. The bullying features were created by defining a list of insulting words (as insulting seeds) and finding similar terms to the insulting words using word embeddings. Linear SVM was used and the attained F1 score was 78%. Hosseinmardi et al. [17] analyzed Instagram posts and associated English comments to detect cyberbullying incidents for the posting users (i.e. cyberbullying, non-cyberbullying). The features for classification included metadata (i.e. no. of associated commments, following, followed-by, likes, frequency of comments), comments (i.e. n-grams bag of words), and labeled image categories. They experimented with NB and linearSVM algorithms. The best classifier was linearSVM with the use of single value decomposition (SVD) to reduce dimensions of n-gram features and kernel principle component analysis (kernelPCA) to reduce dimension of the rest of the features. Its attained accuracy was 87%.

Unlike the related work above, we aim to provide an analysis of the written form of cyberbullying for Thai language. We are inspired by the work of Van Hee et al. [14] the most and the analysis is also performed for multiple categories of cyberbullying, but our training data were collected from Twitter instead as it is a good source for cyberbullying research [18]. Unlike the work [15] and [17] where metadata of posting users like user network (i.e. following/followers), activity of users (e.g. no. of posted tweets), and user profile (i.e. personality, gender, age) were considered, we opt for features extracted from textual comments only. This is because textual comments and such user metadata are not always easily obtained from all social network platforms, e.g. to collect training data, it would be more difficult to obtain comments and metadata of posting users from Facebook than from Twitter. Hence, it is possible that the classifier, built from Twitter data, would be used as a general cyberbullying detection model to classify comment data from different social networks. In that case, the classifier should be built on as minimal features as possible, i.e. textual comment only. In addition, the user metadata are features obtained from the users who tweet. We see that the classifier should be built without any knowledge of such user metadata so that the bullies are kept anonymous. In a real situation, a victim can feel the impact of the attack by only reading the messages and not knowing whom the messages are from.

III. Dataset Construction

This section describes how to construct the dataset for experiment which contains social network messages of different cyberbullying categories as well as non-cyberbullying content.

A. *Data Collection*

Twitter was used as the data source. Due to the limitations of Twitter search API (such as timeframe of tweet history that can be queried, number of allowable requests sent in a period, and number of tweets per request), we used a python program called GetOldTweets [19] to retrieve tweets. The program mimics Twitter search on browsers and can search old tweets in a more convenient way.

The fine-grained message categories proposed by Van Hee et al. [14] comprise types of cyberbullying and roles in cyberbullying. Here, only the types of cyberbullying are considered. Based on the literature review in section II about cyberbullying research and incidents in the Thai context, the written-form of cyberbullying is classified into four categories:

1) *Sexual Harassment*

This category refers to the expressions with "unwelcome sexual advances, requests for sexual favors, and other conduct of a sexual nature. Examples include 1) making sexual comments about a person's body, clothing, or looks, 2) making sexual comments or innuendos, 3) asking about sexual fantasies, preferences, or history, 4) asking personal questions about social or sexual life, 5) spreading rumors about a person's personal sex life" [20].

2) *Insult and Threat*

This category refers to the expressions containing "abusive, degrading, or offensive language [21], or an intention to inflict pain or damage physically or psychologically in order to hurt, offend, or intimidate a person."

3) *Race and Religion*

This category refers to the expressions of "discrimination that are based on the person's race, skin color, ethnicity, nationality, or religion" [21]. Discrimination cases are less common in Thailand. However, the majority of Thai people are Buddhists, but in the southern provinces of the country, there is a sizeable Malay Muslim population. Some people are inclined towards separatism due to cultural differences and violence that happened in the past. In recent decades, there have been violent incidents, including those involving Thai Muslim and Thai Buddhist, in the area.

4) *Intelligence, Appearance, and Social Status*

This category refers to the expressions containing remarks about the person's intelligence, physical appearance, or social status to hurt or offend the person.

Based on the definition above, data collection for the dataset was guided by keywords and hashtags that could signify cyberbullying or the names of individuals who were likely to be the targets of cyberbullying. Some messages were obtained from known Twitter accounts whose tweets signify cyberbullying acts. We used the search terms in Table I for collecting old Twitter messages, using GetOldTweets and specifying 1000 tweets per request (i.e. per term). Note that the search terms that are not general terms and can identify specific persons are concealed.

B. *Data Annotation*

There were 12,547 tweets obtained from the term-based search. The tweets were dated from November 2011 to March 2019. We randomly sampled a subset for the manual annotation and used a peer manual content analysis to annotate the sampled tweets. Two coders who were senior computer engineering students were assigned to do the task. We introduced the definitions of cyberbullying categories to the coders and discussed examples to reduce disagreements. Each coder then read each tweet content carefully and specified one of the cyberbullying categories or non-cyberbullying. In the case of disagreement, the coders discussed to find an agreement. A summary and examples of the labeled data are shown in Table II.

IV. Experiment

The dataset was preprocessed and had features extracted for building cyberbullying classifiers. The best performing

TABLE I
SEARCH TERMS FOR COLLECTING TWEETS

Category	Search Term	Description
Sexual Harassment	@<account1>, @<account2>	Two accounts that post pictures of women and give sexual comments about those women (account names are concealed here)
	น่าเย็ด	(Vulgar) Fuckable, sexually desirable
Insult and Threat	#<name>	Name of a woman who criticized the look of other people and was condemned by the public (name is concealed here)
	#<singer>	Name of a singer who had two girlfriends at the same time (name is concealed here)
	#<actress>	Name of an actress who allegedly came between another actress and her husband (name is concealed here)
	<answer>	A singer’s silly answer to a game show question. She was largely criticized for pretending to be so cute and naive like a child, and not knowing the answer.
	เหี้ย	(Vulgar, very angry) Fuck, asshole
	ร่าน	(Vulgar) Slut
	แรด	(Vulgar) Bitch
	ควย	(Vulgar) Dick (for a very bad and mean person)
	หน้าหี	(Vulgar) Cunt face (for a very bad and mean person)
	อยากต่อย	Want to punch
Race and Religion	อิสลาม	Islam
	มุสลิม	Muslim
	ละหมาด	Worship in Islam
	พระพุทธเจ้า	Buddha
	สิทธัตถะ	Siddhartha (the Buddha’s name before entering the monkhood)
	เยซู	Jesus
Intelligence, appearance, and social status	ควาย	(Slang) Stupid person, idiot
	โง่	Stupid
	ปัญญาอ่อน	Retarded, idiotic
	ไร้การศึกษา	Uneducated
	ตลาดล่าง	(Slang) Low-end market (person having low social status or unacceptable bad behavior)
	หน้าเหี้ย	(Vulgar) Looking very ugly
	สารรูป	(Bad connotation) Appearance, look

classifier was identified for use in the visualization tool.

A. Data Preprocessing

We performed some data cleansing and preprocessed the tweets in the dataset as follows:

1) *Removal of irrelevant text*

The tweets may contain characters that are not suitable for further analysis, e.g. emoji, text in other languages, URLs, numbers, and symbols. These were removed from the tweets in the dataset. Also, the dataset contained no retweets.

TABLE II
SUMMARY OF LABELED DATA USED IN TRAINING

Category	No. of Tweets	Example
Sexual Harassment	484	ทรงนี้ น่าเย็ด ครับ^^ มีโดนทุกวันแน่นอน (Good figure. So fuckable. Could have a fuck everyday.)
Insult and Threat	497	สรุปง่ายๆ คือผู้ชายแม่งเหี้ย #ป๊อปปองกูล (Simply put, this man is an asshole. #the man’s name)
Race and Religion	102	ไม่มีมนุษย์คนไหนที่ มีสติปัญญาไปนับถือลัทธิ อิสลาม หรอก (There are no wise men who worship Islam.)
Intelligence, appearance, and social status	502	โง่ แล้วเสือกทำตัวกร่าง ไอ้ควาย (Stupid but with a swaggering act! Such an idiot!)
Other	1585	อิสลาม กินหมูไม่ได้ (Eating pork is prohibited in Islam.)
Total	3,170	

2) *Spelling correction*

The tweets may contain typos. Misspelled words in the dataset were corrected.

3) *Tokenization*

The tweets are transformed into lists of words (or tokens) that would be the basis for feature engineering in the next step. We transformed each tweet in the dataset into a list of tokens using the PyThaiNLP library [22]. For example, after data cleansing, the tweet โง่ แล้วเสือกทำตัวกร่าง ไอ้ควาย (Stupid but with a swaggering act! Such an idiot!) was tokenized into [‘โง่’, ‘แล้ว’, ‘เสือก’, ‘ทำตัว’, ‘กร่าง’, ‘ไอ้’, ‘ควาย’].

B. Feature Engineering

After preprocessing, each tweet is represented as a vector of features. The features are characteristics of the tweets which should be able to distinguish between different cyberbullying categories. Using scikit-learn 0.21.2, we represented each tweet by the following features based on the tokens of all tweets in the dataset.

1) *Word unigram bag-of-words*

These features are unique single words (unigrams) in the documents in a corpus, i.e. tweets in the dataset. Each tweet was converted into a vector by counting the number of times each unigram appeared in the tweet.

2) *Word unigram bag-of-words with TF-IDF*

Term frequency-inverse document frequency (TF-IDF) is a weighted value that indicates how important any single word is to a document in a corpus. It is computed by

$$TF - IDF\ weight = TF * IDF \tag{1}$$

where

$$TF(w) = \frac{number\ of\ times\ word\ w\ appears\ in\ a\ document}{total\ number\ of\ words\ in\ a\ document} \tag{2}$$

$$IDF(w) = \log_e\left(\frac{total\ number\ of\ documents}{number\ of\ documents\ with\ word\ w\ in\ it}\right). \tag{3}$$

A word is important to a document if the frequency of the word in the document is high. However, the importance is offset by the frequency of the word across documents in the corpus. If the word appears in many documents, its importance to those documents would be low. Each tweet in

ISBN: 978-988-14048-7-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the dataset was converted into a vector by calculating a TF-IDF weight of each unigram with respect to the tweet.

3) Word bigram bag-of-words

These features are unique two-word pairs (bigrams) in the documents in a corpus, e.g. the bigrams of the tokenized text [‘โง่’, ‘แล้ว’, ‘เสือก’, ‘ทำตัว’, ‘กร่าง’, ‘ไอ้’, ‘ควาย’] are [‘โง่_แล้ว’, ‘แล้ว_เสือก’, ‘เสือก_ทำตัว’, ‘ทำตัว_กร่าง’, ‘กร่าง_ไอ้’, ‘ไอ้_ควาย’]. Each tweet was converted into a vector by counting the number of times each bigram appeared in the tweet.

4) Word bigram bag-of-words with TF-IDF

In this case, a TF-IDF weight indicates how important any bigram pair is to a document in a corpus. Each tweet in the dataset was converted into a vector by calculating a TF-IDF weight of each bigram with respect to the tweet.

When these features were applied to the training data, the dimension of unigram features was 4,880 and that of bigram features was 25,868.

C. Handling Imbalanced Class Distribution

The dataset in Table II showed that the distribution of the data classes was imbalanced, especially having the Race and Religion class being the smallest minority. Imbalanced class distribution can prevent the classifier from accurately classifying the tweets. A technique such as the Synthetic Minority Over-sampling Technique (SMOTE) can be used to create synthetic samples of the minority classes [23]. We used SMOTE to make all data classes distributed more equally. There are 9,315 samples of training data after applying SMOTE, 1863 for each class.

D. Experimental Setting

We ran four machine learning algorithms, i.e. MultinomialNB, LinearSVC, Random Forest, and Logistic Regression [13] to build multiclass classifiers for cyberbullying tweets. Different sets of lexical features were tested, i.e. word unigram, word unigram with TF-IDF, word bigram, and word bigram with TF-IDF. Also, the algorithms were run with and without using the SMOTE technique. The performance of the classifiers was evaluated by stratified 5-fold cross validation.

E. Result and Discussion

Table III shows the average performance of the multiclass classifiers in terms of precision (P), recall (R), F-measure (F1), and accuracy (Acc) from the stratified 5-fold cross validation in different settings. The performance of different settings was quite consistent across different algorithms. For the Unigram models, the F1 scores were between 49-62% and the accuracy scores were between 63-70% for all algorithms. However, the F1 and accuracy dropped to 32-54% and 58-66% respectively when TF-IDF weights were used with the Unigrams. This is because the IDF of the word that is found in many documents will be low and this word will have a very small TF-IDF. This TF-IDF weight is used by the classifier and if this word is a good feature for the classification task, the performance can drop. This happened to be the case for our classification task as different cyberbullying categories were likely to be distinguishable by specific words. When such words were found in most tweets of certain classes, the performance dropped. Class imbalance also impacts the performance since the good word features of the majority classes may have lower IDF and hence lower TF-IDF weights. When SMOTE was applied to the Unigram model, the F1 scores improved significantly to 67-69% and accuracy improved in some cases to 66-69%. Similarly, when SMOTE was applied to the Unigram with TF-IDF model, the F1 scores improved substantially to 68-74% and accuracy increased to 69-74%. For the Bigram models, TF-IDF weights got the F1 scores decreased in all algorithms and the accuracy dropped in some cases. When SMOTE was applied to the Bigram model, F1 and accuracy decreased in all algorithms, but when applied to the Bigram with TF-IDF model, F1 and accuracy dropped in all algorithms except MultinomialNB. Overall, the performance of all Unigram-based models was better than that of the Bigram-based models. The likely reason was that, single words or unigrams were likely to be predictive of the cyberbullying categories in most cases, rather than bigrams. Nevertheless the written form of cyberbullying is not easy to identify in an automated manner without the current context of what is happening to

TABLE III
PERFORMANCE OF CYBERBULLYING CLASSIFIERS

Feature/ Technique	MultinomialNB				LinearSVC				Random Forest				Logistic Regression			
	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
Unigram	51.25	49.98	**49.34**	**63.26**	65.69	60.79	**62.10**	**68.55**	66.46	52.11	**55.67**	**65.99**	67.72	59.51	**61.98**	**69.68**
Unigram + TF-IDF	42.92	34.19	**31.64**	**57.92**	65.03	51.37	**54.49**	**66.01**	66.59	47.84	**51.50**	**64.76**	58.48	42.69	**44.55**	**63.87**
Unigram + SMOTE	69.84	66.11	**66.90**	**66.11**	70.30	67.45	**67.84**	**67.45**	69.74	67.73	**67.46**	**67.73**	71.62	68.69	**69.07**	**68.69**
Unigram + TF-IDF + SMOTE	72.94	72.39	**70.93**	**72.39**	75.00	68.39	**68.46**	**69.39**	77.49	71.88	**72.32**	**71.88**	77.35	73.61	**<u>73.89</u>**	**<u>73.61</u>**
Bigram	46.08	51.22	46.68	54.37	56.93	44.42	46.27	60.25	56.18	41.16	42.66	58.29	54.08	42.59	43.74	60.49
Bigram + TF-IDF	47.90	30.90	28.99	55.79	55.53	33.25	33.29	57.57	56.34	35.40	36.18	59.06	47.44	30.75	29.34	55.55
Bigram + SMOTE	40.18	28.37	32.24	28.37	53.22	39.88	38.59	39.88	51.69	41.79	40.49	41.79	53.93	42.67	41.91	42.67
Bigram + TF-IDF + SMOTE	65.33	64.92	63.72	64.92	61.73	45.25	42.03	45.25	70.53	52.63	51.83	52.63	67.58	49.48	48.37	49.48

the person. For example, the text อ่านหนังสือเยอะ ๆ นะ (Read a lot of books) could be read as a suggestion from a well-wisher, but if the context is known, the text could be interpreted as a sarcastic form of saying someone is stupid.

Logistic Regression works well for text classification as it can handle sparse data like we had with text. In most cases of the experiment, logistic regression performed better than other algorithms. The Unigram with TF-IDF and SMOTE model was the winning classifier and was exported for use in the visualization tool in the next step.

V. Supporting Tool

Social networks are convenient means for cyberbullying to intensify. A bullying tweet from one Twitter user can reach many people who may also retweet or give additional bullying comments. Cyberbullying is not necessarily a private attack from one bully or a few but can grow into a public attack. Such repetition of cyberbullying acts can occur 24/7 until the incidents die down after a while or, in some cases, may continue over a longer period. The proposed method is supposed to be used to identify the risk of cyberbullying attack on an individual. For example, family and friends might need to know if their loved ones have become targets of cyberbullying. The proposed method comes with a tool to visualize cyberbullying incidents of an individual over a period of time, using the Matplotlib library for Python [24]. The tool can help to monitor potential risk of cyberbullying to an individual and to determine how repetitive the attack is, so that appropriate actions can be taken.

Using the tool, a user can specify a keyword, such as a person name, and starting and ending time in order to retrieve relevant messages that were tweeted during that period. The tweets are preprocessed, represented by their feature vectors, and classified by the cyberbullying classifier from the previous section. Fig. 1 shows a cyberbullying classification graph for a male singer. The tool retrieved messages that were tweeted about the singer during the month of February 2019. As shown in the graph, there were a lot of cyberbullying tweets around the end of the month because the secret about him living a double life with two girlfriends was revealed and he was strongly criticized by the public. Most of the tweets were insults while there were traces of sexual harassment tweets also. Since he is a big man, some of the tweets attacked his physical appearance and weight. The incident led him to see a psychiatrist.

VI. Conclusion

This paper has presented an intial attempt in the Thai context to identify cyberbullying incidents that occur to an individual. Machine learning algorithms were used to learn lexical feaures of Twitter messages and build classifiers to classify four cyberbullying categories, i.e. sexual harassment, insult and threat, race and religion, and intelligence, appearance, and social status. The accompanying tool can give a view of the bullying comments that an individual has received over a period of time. The method and tool can only identify the "risk" of cyberbullying as they do not identify whether the person feels defenceless and overpowered by the comments, and really becomes a cyberbullying "victim".

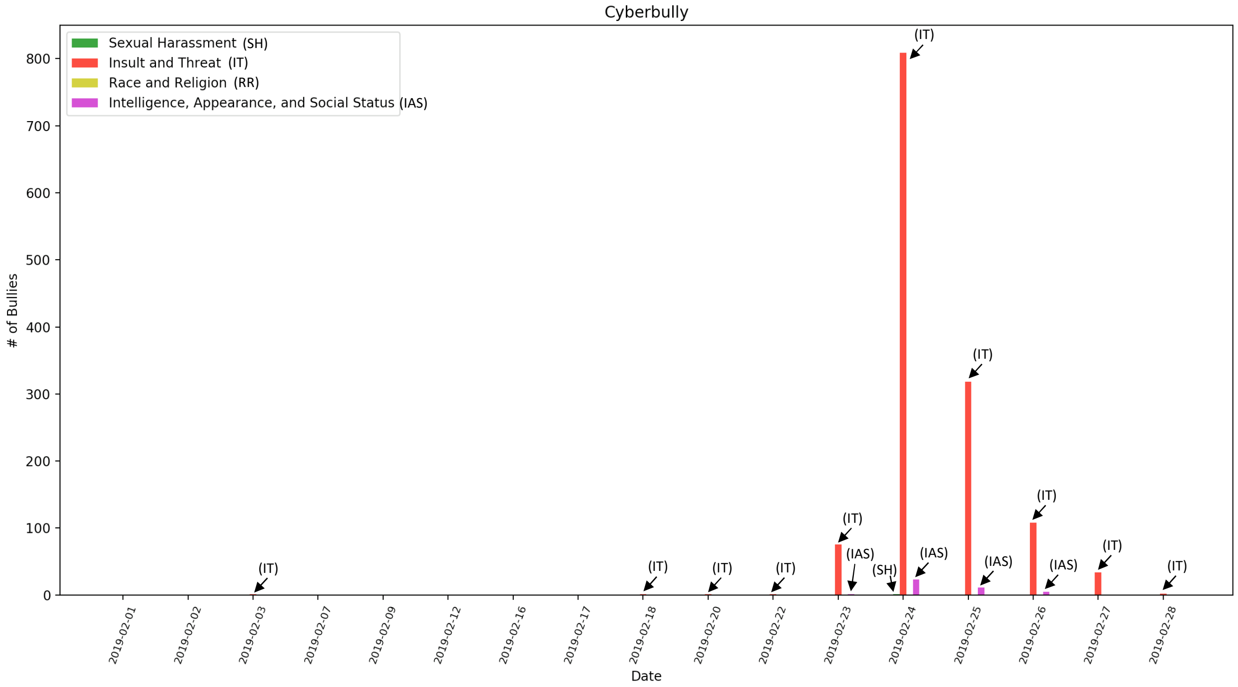

Fig. 1. Example of classification of cyberbullying tweets reported by the tool.

To improve the method, more training data should be collected, and other features can be experimented such as character n-grams, part of speech, and sentiments [25]. As suggested by Van Hee [14], the detection of messages from bystanders who give support to the person should give more insight into the severity of the incident. In addition, the person's reaction and how he/she copes with the incident should be analyzed.

References

[1] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: its nature and impact in secondary school pupils," *The Journal of Child Psychology and Psychiatry*, vol. 49, issue 4, Apr. 2008, pp. 36-385.

[2] E. Menesini, A. Nocentini, B. E. Palladino, A. Frisén, S. Berne, R. Ortega-Ruiz, J. Calmaestra, H. Scheithauer, A. Schultze-Krumbholz, P. Luik, K. Naruskov, C. Blaya, J. Berthaud, and P. K. Smith, "Cyberbullying definition among adolescents: a comparison across six European countries," *Cyberpsychology, Behavior and Social Networking, 15(9)*, Sep. 2012, pp. 455-463.

[3] stopbullying.gov, What Is Cyberbullying [Online]. Available: https://www.stopbullying.gov/cyberbullying/what-is-it/index.html

[4] K. Athanasiou, E. Melegkovits, E. K. Andrie, C. Magoulas, C. K. Tzavara, C. Richardson, D. Greydanus, M. Tsolia, A. K. Tsitsika, Cross-national aspects of cyberbullying victimization among 14–17-year-old adolescents across seven European countries, *BMC Public Health* (2018) 18:800, 15 pp.

[5] brandinside.asia. (2017, Jun 15). Stop Bullying Thailand Top 5 [Online]. Available: https://brandinside.asia/stop-bullying-thailand-top5/ (in Thai)

[6] P. Cheyjunya, "A meta-analysis on cyberbullying factors correlation in Thai academic research," *Journal of Communication Arts, vol. 36, no. 2*, May-Aug. 2018, 13 pp.

[7] R. Sittichai, "Information technology behavior cyberbullying in Thailand: incidence and predictors of victimization and cyber-victimization," *Asian Social Science, vol. 10, no. 11*, 2014, pp. 132-140.

[8] N. Songsiri and W. Musikaphan, "Cyber-bullying among secondary and vocational students in Bangkok," *Journal of Population and Social Studies, vol. 19, no. 2*, Jan. 2011, pp. 235-242.

[9] Touro University Worldwide. (2015, Jul 15). Digital Threats: The Impact of Cyberbullying [Online]. https://www.tuw.edu/health/impact-of-cyberbullying/

[10] Stop Cyberbullying Thailand [Online]. https://www.facebook.com/pg/stopcyberbullyingthailand/posts/ (in Thai)

[11] star2.com (2018, Aug 30). More and More Celebrities Are Being Bullied Online [Online]. Available: https://www.star2.com/entertainment/2018/08/30/celebrity-cyberbullying-victims/

[12] Net Nanny. (2016, Oct 9). 9 Celebrities You Never Knew Were Cyber-Bullied [Online]. https://www.netnanny.com/blog/9-celebrities-you-never-knew-were-cyber-bullied/

[13] A. C. Müller and S. Guido, Introduction to Machine Learning with Python. Sebastopol, CA: O'Reilly, 2016.

[14] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V.´ Hoste, "Automatic detection of cyberbullying in social media text," *PLoS ONE 13(10)*, Oct. 2018, 22 pp.

[15] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Computers in Human Behavior, vol. 63*, Oct. 2016, pp. 433-443.

[16] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proc. 17th Int. Conf. Distributed Computing and Networking*, Jan. 2016, 7 pp.

[17] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. (2015, Mar 12). Detection of Cyberbullying Incidents on the Instagram Social Network [Online]. Available: https://arxiv.org/abs/1503.03909

[18] C. Barncard. (2012, Aug 1). Learning Machines Scour Twitter in Service of Bullying Research [Online]. Available: https://news.wisc.edu/learning-machines-scour-twitter-in-service-of-bullying-research/

[19] J. Henrique. GetOldTweets-python [Online]. Available: https://github.com/Jefferson-Henrique/GetOldTweets-python

[20] What is Sexual Harassment? [Online]. Available: https://www.un.org/womenwatch/osagi/pdf/whatissh.pdf

[21] C. Van Hee, B. Verhoeven, E. Lefever, G. De Pauw, W. Daelemans, and V. Hoste, "Guidelines for the fine-grained analysis of cyberbullying, version 1.0," Tech. Rep. No. LT3 15-01, LT3, Language and Translation Technology Team, Ghent University, Aug. 2015.

[22] PyThaiNLP [Online]. Available: https://github.com/PyThaiNLP/pythainlp

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, 16(1), 2002, pp. 321-357.

[24] Matplotlib [Online]. Available: https://matplotlib.org/

[25] R. Arreerard and T. Senivongse, "Thai defamatory text classification on social media," in *Proc. 2018 IEEE Int. Conf. on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, Jul. 2018, pp. 77-82.