Journal of Information Science Research and Practice

Research Articles

Identifying Text-based Online Thai Hate Speech in Social Media

Siranuch Hemtanon [1], Ketsara Phetkrachang [2] and Wachira Yangyuen [3,*]

[1] Management Program, Faculty of Business Administration Rajamangala University of Technology Srivijaya in Songkhla, Thailand

[2] Computer Engineering, Faculty of Engineering Rajamangala University of Technology Srivijaya in Songkhla, Thailand

[3] College of Industrial Technology and Management Rajamangala University of Technology Srivijaya Nakhon Sri thammarat, Thailand

***** Corresponding author email: Wachira.y@rmutsv.ac.th

Abstract

Purpose: This work proposes a method to detect Thai online hate speech which can be categorized to 5 types, including ethnic-based, gender-based, ableism, belief-based, and social status-based hate speech. Online comments from famous social network services in Thailand are collected and annotated for training data.

Methodology: Machine learning approaches are employed to perform multiclass classification for identifying the hate speech. Moreover, we exploit the information gain score to determine which terms are significant to relay hateful intent of each hate speech class.

Findings: The results of hate speech detection reveal that a language model of combining TF-IDF and trigram using with SVM technique obtained the best performance in detection for 0.76 F-measure score in average. The use of IG score also provides a list of significant terms that related to a specific hate speech class.

Applications of this study: Hate speech detection helps to analyze Thai text messages that may be hurtful to recipients. It can actively filter and disallow the message before posting to prevent online cyber bullies in social media platforms, and it reminds users who may unintentionally choose Thai risky words that may cause emotional wound to readers.

Keywords: Online hate speech, Classification, Social network service, Keyword detection, Text mining

Received: May 21, 2024 Revised: July 31, 2024 Accepted: September 7, 2024 https://doi.org/10.14456/jiskku.2024.27

1. Introduction

Online harassment refers to any form of aggressive, abusive, or threatening behavior that takes place on the internet (Melander, 2010; Willard, 2007). The behavior aims to target individuals by intimidating, threatening, or harming someone through online or digital communication platforms such as social media, online forums, and online gaming environments. The harassment can manifest in various forms, including impersonation/identity theft, stalking, doxing, hate speech, or spreading malicious rumors. Online harassment can have significant psychological, emotional, and even physical consequences for the victims. It is important to recognize and address online harassment to ensure the safety and well-being of individuals in the digital space.

Among the online harassment acts, the uses of hate speech and the spreading of malicious rumors are most frequent in Thailand. Many incidents can be sighted in Thai social network media, especially among celebrities and well-known persons. There are several topics of hate speech, such as ethnicity, gender, ableism, and the incident of the target. For malicious rumors, it refers to false or misleading information intentionally spreading to damage someone's reputation, cause harm, or create a negative perception about the victim. Malicious rumors are often spread with ill intentions, such as to incite conflict, generate controversy, or undermine someone's credibility.

The number of online hate speech and malicious rumors spreading in Thailand is constantly increasing relative to the greater number of social network services (SNS) users. Some may come from malicious intentions, while some are not intended as unintended hate speech. Unintended online hate speech can happen for various reasons, including a lack of awareness or understanding, poor choice of words or lack of communication skills, and cultural or societal influences.

As online harassment occurs globally and adversely impacts the mental health of victims, the detection of such behaviors is crucial as a preventive measure, particularly on social networks. Research on systems for detecting online bullying and harassment (Kanan, Aldaaja & Hawashin, 2020; Islam et al., 2020; Milosevic, Van Royen, & Davis, 2022) has focused on identifying and classifying online activities intended to harass other social network users. Common methodologies involve the application of text mining techniques (Kusal et al., 2021; Xiong, Yan, Yao, & Liang, 2022; Dang & Ahmad, 2014) or automated classification (Yuvaraj et al., 2021; ALBayari, Abdullah & Salloum, 2021; Neelakandan et al., 2022) to categorize

maliciously intended words on prominent social network platforms such as Facebook, Twitter, and Reddit.

As there are numerous existing works of online harassment detection on English messages in social networks, it may not be able to directly apply to other languages with different cultures. The meaning of insulting and embarrassing is different from culture to culture based on what people of the culture hold important. Thus, this work aims to detect online hate speech and malicious rumors in the Thai language and to classify them into types. Furthermore, we plan to find terms used to deliver the malicious meaning of each type. The target SNS for data collection is Facebook and Pantip, Thailand's most frequently used social network platforms. The rest of this paper is organized as follows. Section 2 provides background knowledge about existing research on detecting online harassment. Section 3 describes the details of online harassment detection of Thai text in Thai on social network platforms. Section 4 gives the experiment setting and evaluation results. Lastly, Section 5 gives a conclusion and future work.

2. Purpose

1) To detect online hate speech and malicious rumors in the Thai language.

2) To identify and classify online activities meant to harass other social network users.

3) To analyze online comments to find significant features representing Thai online hate speech.

3. Methodology

3.1 Data Preparation

In this study, the data consists of Thai text collected from public posts on Facebook and Pantip, which are widely used social networking sites in Thailand. Only the text content was recorded, without any names or identifying information, in accordance with research license ID WUEC-23-039-01. The criteria for data collection are as follows.

Facebook data: we looked for comments made to news posted in famous news page with at least 50,000 followers. The posts in the page have at least 2,000 total reactions (like, love, and angry for example). We collected comments to the post that contained at least 40 characters and have at least 3 replied to the comment (replies are not collected).

Pantip data: we looked for forums with a tag of politics, celebrity, or sport with at least 50 replies, and collected the replies that contained at least 40 and at most 500 characters and had at least 3 replied to the comment (replies are not collected).

The collection was made between February to June 2023 for a total of 19,841 text instances. We then asked a team of mental health personnel and a linguist to assign the label(s) to text instances. For the methods to label the data, mental health personnel and linguist analyzes the collected comments and subjectively decided on the type of hate speech by reaching a consensus between them. The labels are given in Table 1.

Table 1 Types of online hate speech for label and their definition

Label	Definition
H-E	**Ethnic-based hate Speech:** Language that promotes discrimination, stereotypes, or prejudice against individuals or communities based on their race or ethnicity. This can include derogatory slurs, racial epithets, or generalizations that dehumanize or marginalize certain racial or ethnic groups.
H-G	**Gender-based hate Speech:** Speech that targets individuals or perpetuates stereotypes based on their gender and choice of gender including transgender. It can involve sexist remarks, objectification, or demeaning language that undermines or belittles individuals based on their gender identity.
H-A	**Ableism:** Hate speech directed towards individuals with disabilities or impairments. This can involve derogatory language, mocking or belittling comments, or expressions that devalue or exclude people with disabilities.
H-B	**Belief-based hate Speech:** Expressions that target individuals or groups based on their religious or cultural beliefs. This includes the topics of religious difference, or a group of a particular belief/faith such as being vegan, flat earther, supernatural belief, and anti-vaccination.
H-S	**Social status-based hate Speech:** Speech or language that targets individuals or groups based on their socioeconomic status or social standing within a society. It involves using derogatory, offensive, or demeaning language to discriminate against or marginalize individuals based on their economic circumstances, occupation, education, or perceived social hierarchy.
N	Not containing hate speech

The text instances that annotators from both healthcare personnel and a linguist agreed on the same label are kept. The text instances that were not in consensus were discarded. Text instances are allowed to be labeled more than 1 label. As a result, we obtained the dataset with an online hate speech label as shown in Figure 1.

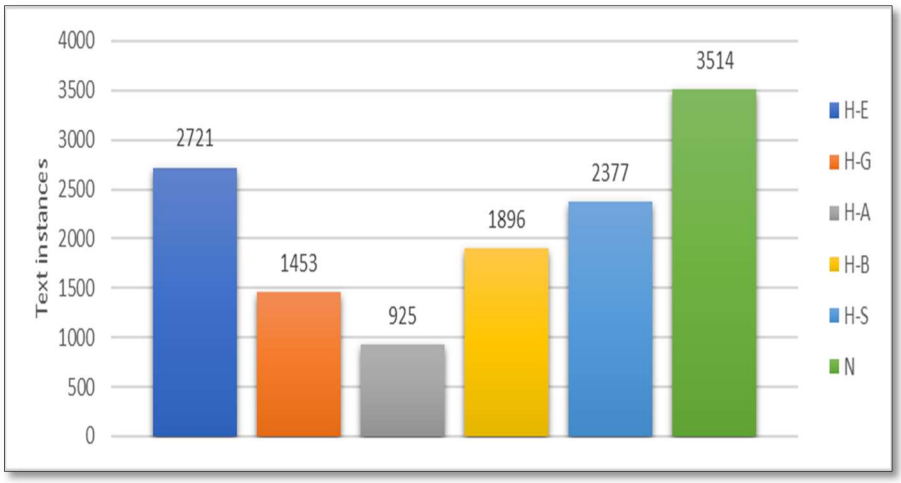

Figure 1: Show about Text instances are allowed to be labeled more than 1 label

3.2 Text Pre-processing

The collected Thai texts are to be processed as follows. The typos and misspelling words are corrected to reduce noises and scattering of the same word. To define words in the given text, word segmentation is applied. The word segmentation service selected in this work is Lexto-plus (Haruechaiyasak & Kongthon, 2013) for the longest conceptualized word. However, the performance of the word segmentation can be incorrect from ambiguity and unknown words, especially the new word recently emerged by teenagers. Thus, post-edit is applied to correct segmented words to improve text input quality. To remain anonymous, all name-entities such as proper noun and tagged username in the context are removed. Last, functional words used for representing grammatical function with little to none meaning including number expression, conjunction, preposition, interjection, ending particle, auxiliary verb, and pronoun are removed to maximize text processing performance in terms of computational complexity from lowering search space.

3.3 Feature and Classification to Detect Hate Speech

For feature representation, we select 2 approaches as term frequency weighting and text sequence. To represent terms in the text instances, Term Frequency-Inverse Document Frequency (TF-IDF), which is a commonly used technique in representing and analyzing the importance of words across a collection of text corpus, are chosen as term weighting approach. It provides a numerical representation of words that reflects their significance in distinguishing between different classes. Term Frequency (TF) measures the frequency of a word within a text instance. It is calculated as the number of times a word appears in an instance divided by the total number of words in that context. Inverse Document

Frequency (IDF) measures the importance of a word across a corpus. It is calculated as the logarithm of the total number of text instances divided by the number of text instances that contain the word. As a result, IDF assigns higher weights to words that appear in fewer text instances and are hence more informative or discriminative. TF-IDF is obtained by multiplying the TF value of a word by its IDF value. This product reflects the importance of the word within the text-instance and the entire corpus. Words with higher TF-IDF scores are those that are frequent within a text instance but infrequent across a corpus, making them potentially more indicative of the class text instances belong to. By calculating TF-IDF scores for all words in the corpus, each instance is represented as a numerical feature vector. The dimensions of the vector correspond to number of the present unique words, and the TF-IDF scores indicate the importance of each word in the collected text instance. To represent text sequence, n-gram is a selected technique to capture the contextual information and sequential patterns of words within a text instance. An n-gram represents a contiguous sequence of n word. n-grams are created by sliding a window of size n over the segmented text. n-grams provide a representation of the sequential patterns and context in the text. Each n-gram is considered as a separate feature. In this work, the generated n-grams are transformed into numerical vectors using one-hot encoding technique.

The classification model utilized in this study is based on a supervised learning technique. With labeled data, the task is to classify text instances into a specific category, employing various machine learning techniques to develop the classification model. In this research, we implement several well-known approaches, including support vector machines (SVM), random forests, and neural network models.

3.4 Finding Significant Terms

To find the most signified terms towards malicious intent among the hate speech text, information gain is calculated to represent a measure of how much information a feature provides about a class. In this work, features are words that appear in a training text. Hence, words can be ranked following their obtained IG score to represent how much impact the words signify the positive of being cyberbullying. IG score is calculated as a measure of the difference in entropy values from before to after the set S is partitioned regarding a word A by (1).

$$IG(A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S|A) \qquad (1)$$

where H(S) is entropy of set S. T refers to subsets from separating set S by a feature A, and p(t) is a proportion of the number of items in subset t to the number of items in set S. Last, H(S) refers to entropy of the subset t. The higher the IG score of a word, the more significance of the word leading to be positive of cyberbullying. In this work, we use the IG score to rank top-n of words as a list of significant terms leading to incur online hate speech intention.

4. Research results

4.1 Evaluation Results of Classifying Online Hate Speech

In this experiment, we evaluate performance of classification of online hate speech. The evaluation measurement is F-measure (F1) score. The collected data were prepared for 5-fold cross validation for evaluation. 5-fold cross validation is a resampling technique that splits the dataset into five equal-sized subsets and then performs training and validation in 5 iterations by switching the training subsets in each run. Each of the separated folds is made sure to contain the similar number of instances for each class given in Table 1. For classification model generation, there are 3 sets of features as 1. TF-IDF, 2. n-gram, and 3. TF-IDF + n-gram. In the experiment, trigram is chosen for n-gram method. For machine learning techniques of classification, support vector machines (SVM), random forests (RF), and neural network (NN) models are selected. All the models in this experiment are trained using Scikit-learn, an open-source machine learning library. All parameters for the machine learning are set to default. The results of classifying online hate speech are given in Figure 2

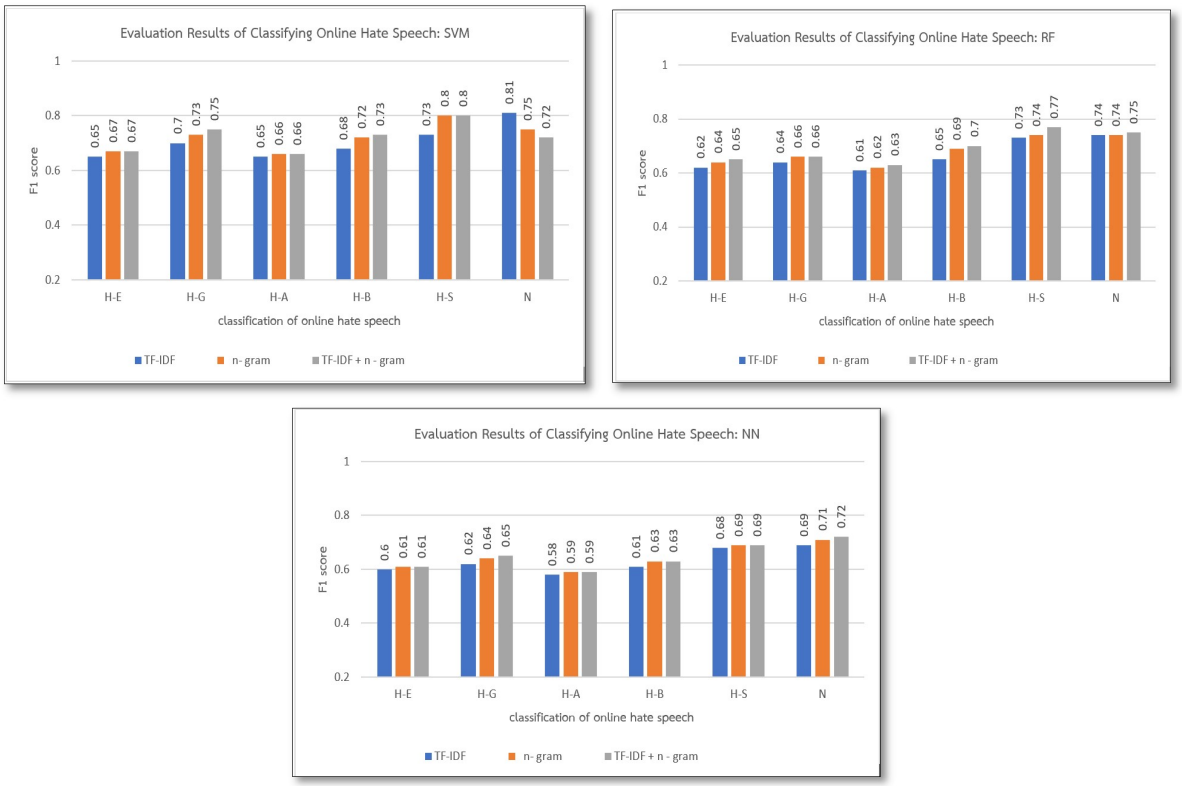

Figure 2 Show about Evaluation Results of Classifying Online Hate Speech

The results show that the best performance is obtained with the model from SVM and TF-IDF + n-gram combination for classifying most classes. In overall, SVM yields the highest F1 score in average, followed by RF. For feature setting, results from TF-IDF + n-gram combination is similar or slightly better than using n-gram in most of the cases, but these 2 settings yield higher F1 score than using TF-IDF solely in all classes.

4.2 Significant Words Representing Online Hate Speech

In this experiment, we want to find terms that are informative for representing an intention of the hate speech for each type, we apply all sentence instances in the dataset to calculate for IG score. The applied feature representation is 3-gram. The found terms commonly used to indicate online hate speech based on a specific type. To scope the terms, we selected the top-20 words with highest IG score of each class, but some found words were inappropriate to display. Thus, we categorized the terms into a semantic group based on intended meaning and exemplified some terms in the group in Table 3. The exemplified terms are in Thai, so we provide POS and literally translation.

Table 3 Significant terms from IG Score indicating hate speech based on class

Classes	Semantic Group	Example
H-E	Naming	ลาว (noun: northeastern, Laotian), กะเหรี่ยง (noun: karen ethnic), เคลมโบเดีย (noun: over-claimer)
	Characteristics	ดั้งแหมบ (adjective: low nose bridge), เสี่ยว (adjective : unfashionable, outdated), หน้าอีสาน (adjective : ugly, face of the poor)
	Other	เข้ากรุง (verb: chasing metropolitan, wishy-washy), บ้านนอก (adjective: countryside, unfashionable)
H-G	Naming	ชะนี (noun: female called by transgender), กระเทยควาย (noun: fat transgender)
	Body part	หน้าปลอม (adjective: face with plastic surgery), หน้าลอย (adjective: cakey), กระเดือก (adjective: noticeably trans gendering)
	Other	แร่ด (adjective: overshowing sexual appeal, bitchy), ปลอม (adjective: fake, plastic surgery), ดอกทอง (adjective: slut, bitchy)
H-A	Disability Characteristics	ง่อย (adjective: crippled, handicapped), เป๋ (adjective: leg-crippled), ใบ้ (adjective: mute, silenced), พิการ (adjective: handicapped).
	Belittling	รกโลก (verb: be useless), เศษสวะ (noun: trash)
H-B	Believer Characteristics	มารศาสนา (noun: religious devil), แครอท (noun: monk), หิวบุญ (verb: hungry for merit) ไร้สมอง (noun: brainless)
	Belittling Belief	งมงาย (adjective: credulous, gullible), เพ้อเจ้อ (adjective: delusional, silly), ไร้สาระ (adjective: non-sense).
	Other	ขอส่วนบุญ (verb: ask for merit), ตกนรก (verb: go to hell), เกิดเป็นเปรต (verb: born as a ghost), นรกส่งมา (verb: hell sent)
H-S	Socioeconomic	ขอทาน (noun: beggar), หนักหัว (verb: heavy headed), สร้างภาพ (verb: create an image)
	Social Standing	ขี้ข้า (noun: lapdog), ไพร่ (noun: servant), กรรมกร (noun: worker), ตลาดล่าง (noun: lower market), ขยะสังคม (noun: dregs of society)

5. Discussion

The contribution of this research is a new method in the detection of Thai hate speech by applying traditional classification methods to categorize hate speech into types and word selection using calculated information gain to list out hate speech relevant terms. The results of hate speech detection reveal that a language model of combining TF-IDF and trigram using with SVM technique obtained the best performance among all selected algorithm in detection

for 0.76 F-measure score in average. The use of information gain score calculated within the classification process effectively provides a list of significant terms that related to a specific hate speech class.

From the significant word extraction results, many words are detected to be relevant to hate speech. We found that some words may not have a direct meaning towards a negative sense such as “แครอท” (noun: monk) and “กรรมกร” (noun: worker). For the former, the word refers to a Buddhist monk as euphemist name to avoid directly referring. The word does not contain any negative sense towards the monk, but it often exists in negative comments when social media users mention misdeeds made by monks. Hence, the word received higher score and selected as a significant term for hate speech. For the latter term, the word itself is a common word referring to workers in the original meaning, but similarly to the former word it is used in negative comments with different surrounding words. Thus, the list of found terms should not be used directly in a dictionary to filter comment hate speech as they may also remove some comments that contain non-negative meaning terms, but they may need to analyze further and group into in inappropriate word group and sensitive word group to manage the content effectively.

The remaining challenge in this work includes polysemous terms and euphemistic hidden meaning. Polysemous terms are words that have multiple meanings or interpretations depending on context. They are a common challenge in language comprehension and processing as they can lead to ambiguity or misunderstandings. As this work exploits term frequency and term surface as representation, terms that have different meanings with same word form are weighted together based on their overall frequency in the corpus leading to inaccurately reflect their relevance to a specific context or meaning. For euphemistic hidden meanings, the issue involves the use of softer or less direct language to convey sensitive, taboo, or potentially offensive ideas. These hidden meanings often rely on implication or inference rather than explicit expression. Understanding euphemistic hidden meanings requires sensitivity to context and cultural nuances. While euphemisms can serve important social functions, such as maintaining politeness or navigating sensitive topics, they also obscure text analysis and tern distribution if not interpreted correctly. To solve the issues, advance word sense disambiguation techniques for Thai are required to identify the correct sense of a polysemous term in a given context. The methods may utilize knowledge bases, corpus statistics, or machine learning algorithms to determine the most appropriate sense of the word based on its surrounding context. However, advance word sense disambiguation techniques

for Thai are another major challenge in Thai natural language processing task including using models such as ELMo (Embeddings from Language Models) and BERT (Bidirectional Encoder Representations from Transformers) to generate word embeddings that are sensitive to context and able to capture the meaning of a word in different contexts to better handle polysemy.

6. Conclusion

Online platforms often provide a sense of anonymity, which can embolden individuals to express hateful and discriminatory views they may not express in face-to-face interactions. To prevent online hate speech, this work proposes a method to detect Thai online hate speech. In this study, we categorize hate speech into 5 types as ethnic-based, gender-based, ableism, belief-based, and social status-based hate speech. The main task is to classify Thai online text that belongs to the type of hate speech as multiclass text classification. The data for machine learning in this work are online comments from famous social network services in Thailand. The data are collected and annotated. For Thai text representation, term frequency and inverse document frequency (TF-IDF) and n-gram are selected. Machine learning approaches including Support Vector Machine (SVM), Neural Network (NN), and Random Forest are employed to perform multiclass classification for identifying the hate speech. We exploit the IG score to determine terms that are significant to relay hateful intent of each type of hate speech. The results of hate speech detection reveal that a language model of combining TF-IDF and trigram using with SVM technique obtained the best performance in detection for 0.76 f-measure score in average. The use of IG score also provides a list of significant terms that related to a specific hate speech class, and the list thus can be used as profanity list for preventing future hate speech.

Reference

ALBayari, R., Abdullah, S., & Salloum, S. A. (2021). Cyberbullying classification methods for Arabic: A systematic review. **Proceedings of the International Conference on Artificial Intelligence and Computer Vision.** (375-385). Settat, Morocco: Springer International Publishing.

Dang, S., & Ahmad, P. H. (2014). Text mining: Techniques and its application. **International Journal of Engineering & Technology Innovations, 1**(4), 22-25.

Haruechaiyasak, C., & Kongthon, A. (2013). LexToPlus: A Thai lexeme tokenization and normalization tool. **Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing.** (9-16). Nagoya, Japan.

Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020). Cyberbullying detection on social networks using machine learning approaches. **Proceedings of 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE).** (1-6). Gold Coast, Australia: IEEE.

Kanan, T., Aldaaja, A., & Hawashin, B. (2020). Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents. **Journal of Internet Technology, 21**(5), 1409-1421. https://doi.org/10.3966/160792642020092105016

Kusal, S., Patil, S., Kotecha, K., Aluvalu, R., & Varadarajan, V. (2021). AI based emotion detection for textual big data: Techniques and contribution. **Big Data and Cognitive Computing, 5**(3), 43. https://doi.org/10.3390/bdcc5030043

Melander, L. A. (2010). College students' perceptions of intimate partner cyber harassment. **Cyberpsychology, Behavior, and Social Networking, 13**(3), 263-268. https://doi.org/10.1089/cyber.2009.02

Milosevic, T., Van Royen, K., & Davis, B. (2022). Artificial intelligence to address cyberbullying, harassment and abuse: new directions in the midst of complexity. **International journal of bullying prevention, 4**(1), 1-5. https://doi.org/10.1007/s42380-022-00117-x

Neelakandan, S., Sridevi, M., Chandrasekaran, S., Murugeswari, K., Pundir, A. K. S., Sridevi, R., & Lingaiah, T. B. (2022). Deep learning approaches for cyberbullying detection and classification on social media. **Computational Intelligence and Neuroscience, 2022**, 1-13. https://doi.org/10.1155/2022/2163458

Willard, N. E. (2007). **Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress**. Champaign: Research Press.

Xiong, Z., Yan, Z., Yao, H., & Liang, S. (2022). Design demand trend acquisition method based on short text mining of user comments in shopping websites. **Information, 13**(3), 110. https://doi.org/10.3390/info13030110

Yuvaraj, N., Chang, V., Gobinathan, B., Pinagapani, A., Kannan, S., Dhiman, G., & Rajan, A. R. (2021). Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. **Computers & Electrical Engineering, 92**, 107186. https://doi.org/10.1016/j.compeleceng.2021.107186