Identifying the Presence of Cyberbullying in Tamil-English Phonetic Words Using Browser Plugin

Rina Md. Anwar[1](✉), Puven Alvin Victor[1], Fiza Abdul Rahim[1,2], Marina Md Din[1], Asmidar Abu Bakar[1], and Aliza Abdul Latif[1]

[1] Institute of Informatics and Computing in Energy (IICE), Universiti Tenaga Nasional, 43000 Putrajaya Campus, Jalan Ikram-Uniten, 43000 Kajang, Selangor, Malaysia
{mrina,marina,asmidar,aliza}@uniten.edu.my,
fiza.abdulrahim@utm.my

[2] Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia

Abstract. While social media provides excellent communication opportunities, it also exposes people to potentially threatening situations online. The rising popularity of various social media platforms has enabled people worldwide to freely exchange views, ideas, and interests. But that does not come without consequences, including an increase in cyberbullying. Early identification of cyberbullying has been proven to be beneficial in preventing it from spreading. We have tested a web browser plugin over few social media platforms to identify offensive comments or posts. Furthermore, research on cyberbullying identification has been done in many languages, but none has been done on Tamil and English phonetic words until the time of conducting this study. Therefore, this study attempts to elicit keywords or phrases relating to cyberbullying incidents in both Tamil and English phonetic words. The study results might be helpful to the research community to create further tools and prepare a training dataset for cyberbullying identification.

Keywords: Social networking · Cyberbullying · Detection · Language

1 Introduction

The rise of the Internet has sparked a debate over the influence of social interactions in cyberspace. When you have a platform in which its community can communicate smoothly using social media, there are numerous classes of netizens who, by means of different forms of behaviours, can or may not contribute to a healthy atmosphere. One such dark side of human behaviour is cyberbullying perpetration.

Cyberbullying is a phrase often used amongst netizens, but only a few fully comprehend its true meaning and its detrimental impacts on any victim, regardless of age or background. Cyberbullying occurs when a person or group utilizes the Internet or technologies with the intent to harass, intimidate, disgrace, or abuse someone else [1]. Because humiliation of the victim can "go viral" and become public in a short time, cyberbullying may do more damages than traditional bullying [2].

H. Badioze Zaman et al. (Eds.): IVIC 2021, LNCS 13051, pp. 90–97, 2021.
https://doi.org/10.1007/978-3-030-90235-3_8

Worldwide statistics reveal that cyberbullying incidents have increased gradually between 2011 to 2018, as reported in [3]. This raised a lot of concern for the public as social media gets toxic every day. Cyberbullying has brought a swirl of adverse effects from suicide to increased and persistent mental health issues such as social anxiety disorder and depression worldwide.

As cyberbullying occurs around the globe, it immediately diversifies regarding its inflected language depending on the nation it "visits". Cyberbullying can take place in different languages, depending on one's language proficiency. It can also be presented in a mixture of languages, especially in a multiracial country such as Malaysia. A mixture of languages is utilized for hurting or reaching a wider "audience" with various language proficiency.

Indeed, Malaysian cyberbullies tend to communicate and, in some instances, employ all their harmful words in a range of languages. However, there is limited research in cyberbullying that uses different types of languages or a mixture of multiple languages. The likelihood of the cyberbullying victim feeling the same or perhaps worse effects using a mix of language that may convey more hurtful meanings. Hence, there is a serious need to investigate methods for identifying potentially offensive mixed languages on social media platforms.

2 Background

There have been numerous instances of cyberbullying in the digital world, and the constant and intensive use of social media worldwide makes it unlikely that these figures will fall. In addition, people have more time spent on social media through the Covid-19 pandemic, raising their chances of becoming cyberbullies or victims.

A 2018 poll was carried out with 20,793 people worldwide on cyberbullying by IPSOS [3]. The results showed that 3 out of 10 young Malaysians had been victims of cyberbullies in the country. Malaysia ranks second in Asia, where parents have reported their children experiencing cyberbullies.

Most research in cyberbullying identification was based on either a filters program or machine learning approaches [4–7]. As multiple uncertainties are inherent in one language, several identification cyberbullying techniques for different contexts have been investigated by scholars. For example, Saravanaraj et al. (2016) suggested a cyberbullying identification model using Naïve Bayes, where the presence of an abusive word indicates cyberbullying, and the absence indicates otherwise [7].

In addition to content, researchers have examined user-related factors in cyberbullying identification systems. This includes age, gender, sexual orientation, and race, as discovered by [8, 9]. Many studies also were conducted to classify between cyberbully or non-cyberbully [10–12].

It is believed that the integration of user knowledge, features, and post-harassing behaviour, for example, by readdressing their harassment experiences by placing a new status on another social network, will increase the cyberbullying identification's accuracy [13]. These strategies have over the years been repeated and enhanced to better match modern language nuances and the expansion of social media features.

In this study, we focused on the identification of the most commonly used Tamil-English phonetic offensive words. A wide search of existing journals and publications was conducted, and no previous work was found on the identification of cyber-bullying in used Tamil-English phonetic words. Since Tamil is also a language that transcends the multiracial people of Malaysia, it may also be utilized as a medium for harm and distress online.

3 Browser Plugin Testing

The browser plugin used in this study was adapted from two other Google Chrome Browser Plugins or Extensions, "NO SPOILERS" [14] and "CODAR" [15]. "NO SPOILERS!" was created by a Malaysian named Ng Khai Yong, which will automatically blur out any posts talking about Avengers: Endgame movie. The other plugin is called "CODAR" or "Cyber Offense Detecting and Reporting Framework" which performs Text Toxicity Prediction on public Facebook posts or comments using BeautifulSoup and Facebook API. Both plugins are integrated to test the sample dataset in this study.

Figure 1 illustrates the workflow of the testing activities, from preparing the dataset preparation to configuring the manifest file, modifying the sample dataset and finally testing the adapted plugin.

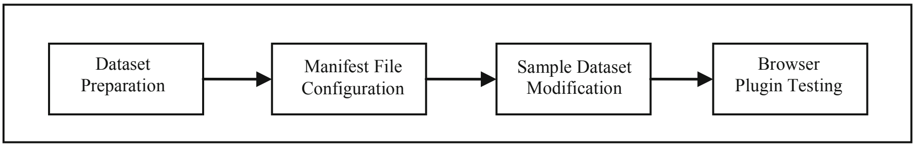
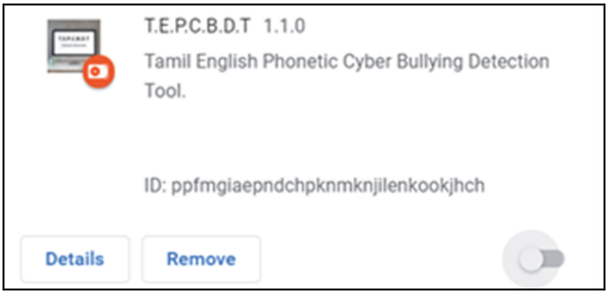


Fig. 1. Browser plugin testing workflow

Dataset Preparation. The key challenge in this study is the availability of a suitable dataset, which is necessary to identify the presence of cyberbullying. In this study, we have developed a Tamil-English phonetic words dataset obtained from the comment sections of public posts of some popular Facebook pages and Twitter accounts. We select a sample of 45 offensive Tamil-English phonetic words from the dataset for testing purposes.

Configuration, Modification and Testing. The manifest file was coded and tested, similar to all Google Chrome Browser plugins. The manifest is the Google Chrome Browser plugin' backbone or the central nervous system. If the manifest file is successfully coded and loaded into the Google Chrome web browser, this is visible in the plugins menu at chrome://extensions on the same browser, as shown in Fig. 2. When the plugin is selected, it is displayed the entire name, version number, description, and unique ID number. Each plugin is allocated an ID number since each plugin has its own unique ID.

The next step was to align manifest.json script with a jquery-3.6.0.min.js script. The jquery-3.6.0.min.js script is available to download on the jquery website. After downloading it, it is saved to the same directory as the manifest.json script, as shown in Fig. 3.

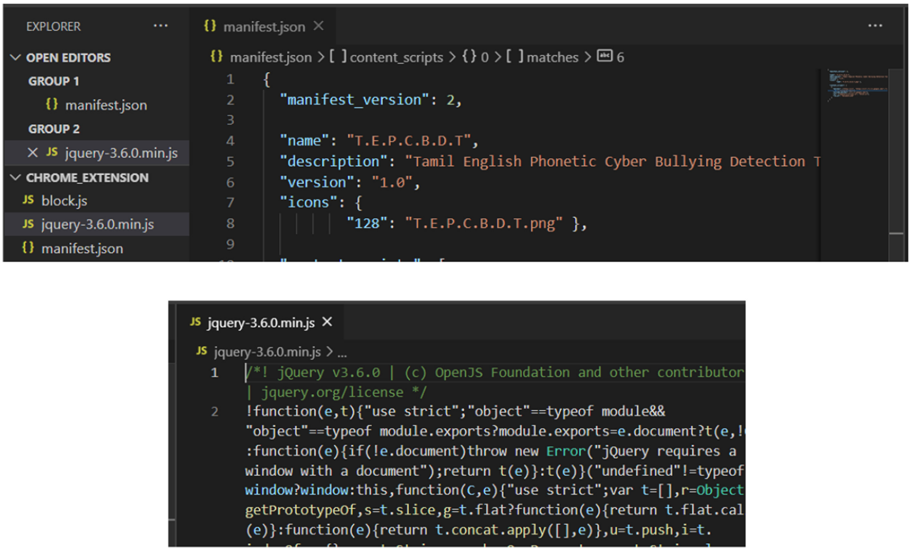


T.E.P.C.B.D.T 1.1.0

Tamil English Phonetic Cyber Bullying Detection Tool.

ID: ppfmgiaepndchpknmknjilenkookjhch

Fig. 2. The activation menu of the plugin



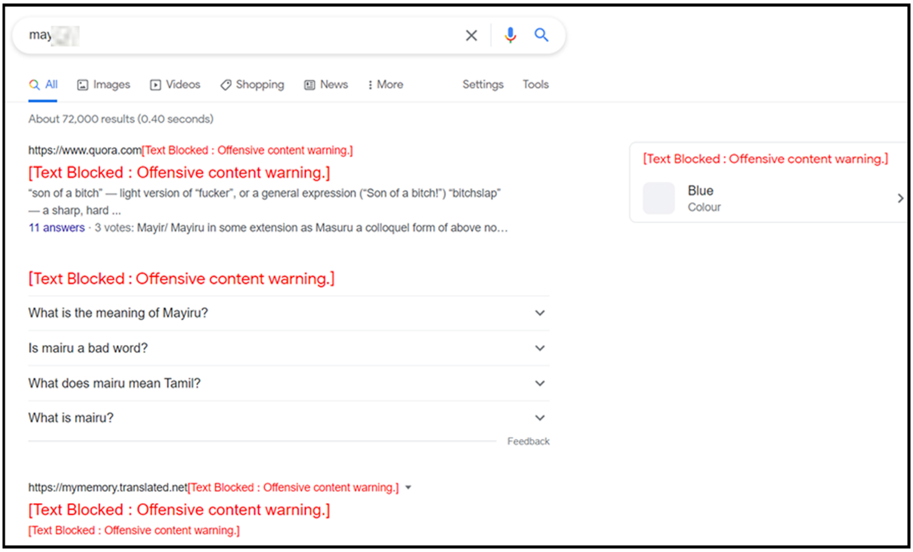
```
{
  "manifest_version": 2,

  "name": "T.E.P.C.B.D.T",
  "description": "Tamil English Phonetic Cyber Bullying Detection T
  "version": "1.0",
  "icons": {
          "128": "T.E.P.C.B.D.T.png" },

```

```
/*! jQuery v3.6.0 | (c) OpenJS Foundation and other contributor
| jquery.org/license */
!function(e,t){"use strict";"object"==typeof module&&
"object"==typeof module.exports?module.exports=e.document?t(e,!
:function(e){if(!e.document)throw new Error("jQuery requires a
window with a document");return t(e)}:t(e)}("undefined"!=typeof
window?window:this,function(C,e){"use strict";var t=[],r=Object
getPrototypeOf,s=t.slice,g=t.flat?function(e){return t.flat.cal
(e)}:function(e){return t.concat.apply([],e)},u=t.push,i=t.
```

Fig. 3. The manifest.json and jquery-3.6.0.min.js scripts working under the same directory

After the manifest file is configured, the block.js file, adapted from the NO SPOILERS! and CODAR extensions, is equipped with a new dataset that consists of 45 offensive Tamil-English phonetic words. However, certain concerns with the security policies on Facebook, Twitter, Instagram, and Youtube have prevented the detection procedure from operating on these websites. Despite that, the Google Chrome search bar detects and blocks photos from the same search. After several adjustments have been made, the adapted plugin still could not bypass Facebook, Twitter, Instagram, and Youtube's security policies.

If the offensive words are found, a message will appear in red font colour showing "[Text Blocked: Offensive content warning.]", indicating a clear warning that the word or phrase being searched contains an offensive word. As shown in Fig. 4 and Fig. 5, the plugin can identify and block out the offensive phrases in the Chrome and standard Incognito web browsers.


may[redacted]

About 72,000 results (0.40 seconds)

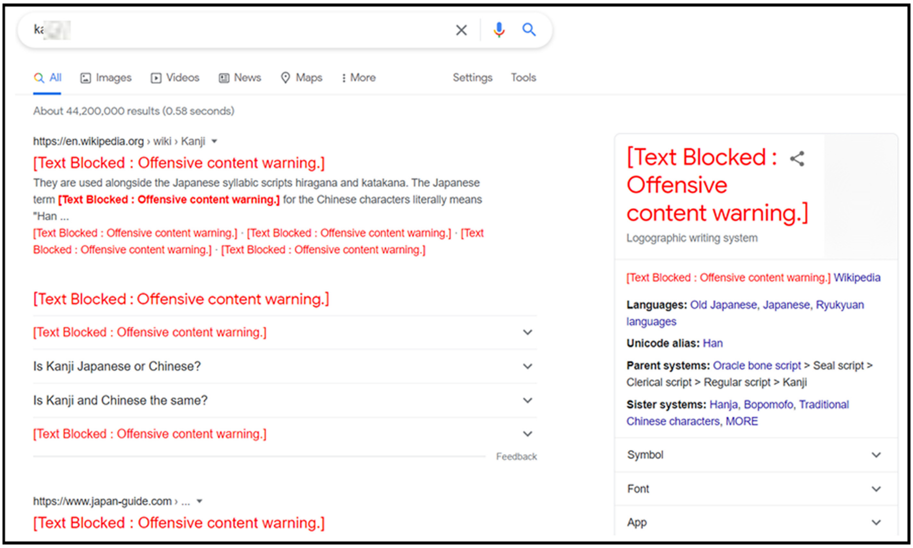
https://www.quora.com[Text Blocked : Offensive content warning.]
[Text Blocked : Offensive content warning.]
"son of a bitch" — light version of "fucker", or a general expression ("Son of a bitch!") "bitchslap" — a sharp, hard ...
11 answers · 3 votes: Mayir/ Mayiru in some extension as Masuru a colloquel form of above no...

[Text Blocked : Offensive content warning.]
What is the meaning of Mayiru?
Is mairu a bad word?
What does mairu mean Tamil?
What is mairu?

https://mymemory.translated.net[Text Blocked : Offensive content warning.]
[Text Blocked : Offensive content warning.]
[Text Blocked : Offensive content warning.]

[Text Blocked : Offensive content warning.]
Blue
Colour

Fig. 4. Search result using standard Chrome tab


ka[redacted]

About 44,200,000 results (0.58 seconds)

https://en.wikipedia.org › wiki › Kanji
[Text Blocked : Offensive content warning.]
They are used alongside the Japanese syllabic scripts hiragana and katakana. The Japanese term **[Text Blocked : Offensive content warning.]** for the Chinese characters literally means "Han ...
[Text Blocked : Offensive content warning.] · [Text Blocked : Offensive content warning.] · [Text Blocked : Offensive content warning.] · [Text Blocked : Offensive content warning.]

[Text Blocked : Offensive content warning.]
[Text Blocked : Offensive content warning.]
Is Kanji Japanese or Chinese?
Is Kanji and Chinese the same?
[Text Blocked : Offensive content warning.]

https://www.japan-guide.com › ...
[Text Blocked : Offensive content warning.]

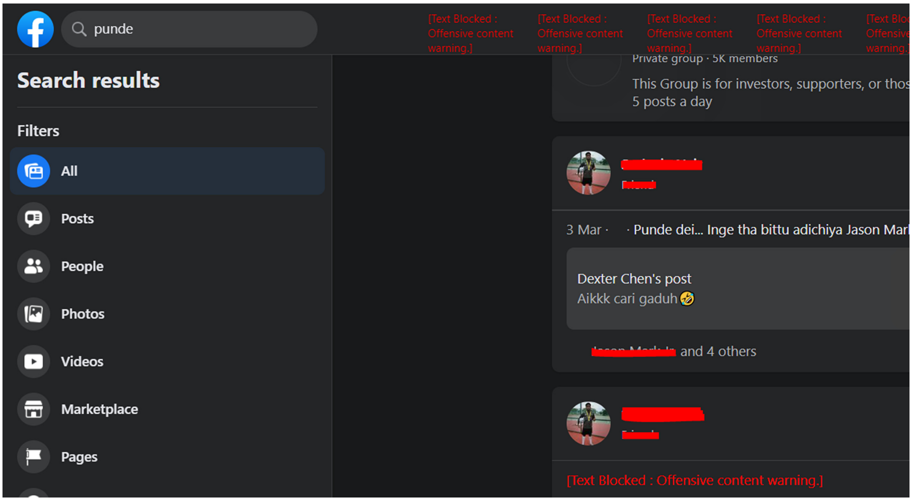
[Text Blocked : Offensive content warning.]
Logographic writing system

[Text Blocked : Offensive content warning.] Wikipedia

Languages: Old Japanese, Japanese, Ryukyuan languages
Unicode alias: Han
Parent systems: Oracle bone script > Seal script > Clerical script > Regular script > Kanji
Sister systems: Hanja, Bopomofo, Traditional Chinese characters, MORE

Fig. 5. Search result using standard Incognito tab

Figure 6, Fig. 7 and Fig. 8 show the adapted plugin running on Facebook, Twitter, and YouTube. The adapted plugin works to identify harmful phrases on Facebook, but it does not work on Twitter and YouTube, which is due to the security policies established by these social networking sites to ensure that no foreign application or plugins can



[Text Blocked : Offensive content warning.] [Text Blocked : Offensive content warning.] [Text Blocked : Offensive content warning.] [Text Blocked : Offensive content warning.] [Text Blo Offensive warning.]

Private group · 5K members
This Group is for investors, supporters, or thos
5 posts a day

[redacted]

3 Mar · · Punde dei... Inge tha bittu adichiya Jason Mar

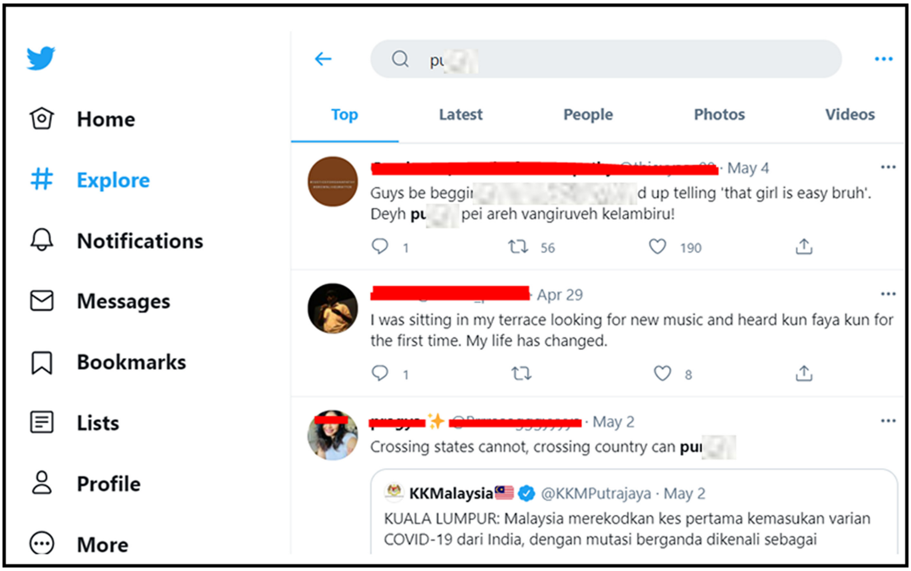
Dexter Chen's post
Aikkk cari gaduh 🤣

[redacted] and 4 others

[redacted]

[Text Blocked : Offensive content warning.]

Fig. 6. The adapted plugin running on Facebook



[redacted] · May 4
Guys be beggi[redacted]d up telling 'that girl is easy bruh'. Deyh **pu**[redacted] pei areh vangiruveh kelambiru!
1 56 190

[redacted] Apr 29
I was sitting in my terrace looking for new music and heard kun faya kun for the first time. My life has changed.
1 8

[redacted] ✨ [redacted] · May 2
Crossing states cannot, crossing country can **pu**[redacted]

KKMalaysia 🇲🇾 @KKMPutrajaya · May 2
KUALA LUMPUR: Malaysia merekodkan kes pertama kemasukan varian COVID-19 dari India, dengan mutasi berganda dikenali sebagai

Fig. 7. The adapted plugin running on Twitter

comb through any information, even in the background on their websites. The constant changes in their privacy policies can also be attributable to this.

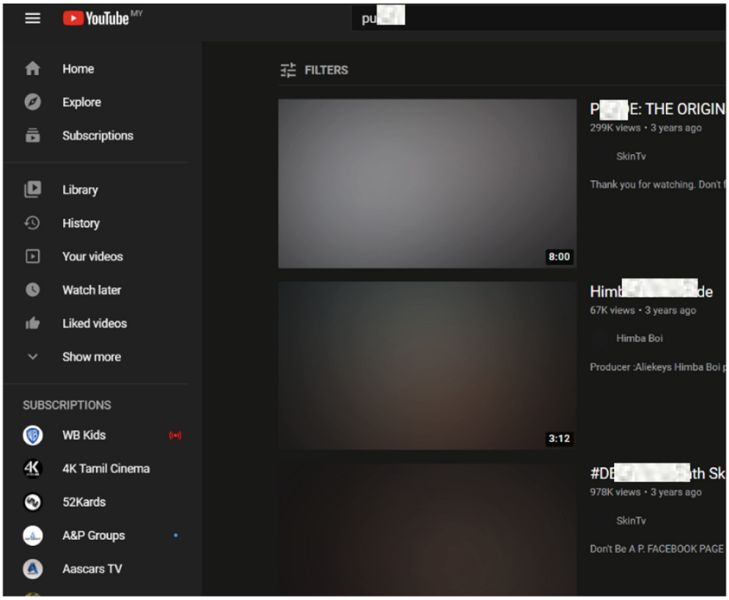


Fig. 8. The adapted plugin running on YouTube

4 Conclusion

The adapted browser plugin focused on identifying inappropriate words and phrases in a mixture of Tamil and English phonetic words. However, much future work is needed to improve the usability and performance of the plugin or develop sophisticated tools that can automatically identify offensive words. Developing a complete dataset is also essential in multiple languages for conducting tests and development to identify cyberbullying.

Hence, this study will be extended to apply different machine learning models for identification purposes. Secondly, we intend to complete the dataset development of the words and phrases in a mixture of Tamil and English phonetic words. Finally, we plan to validate our approach and generalize the results using different datasets from social media sites.

Acknowledgment. This study was funded by BOLD Research Grant 2021 Universiti Tenaga Nasional (J510050002/2021046). We would like to thank UNITEN Innovation & Research Management Centre (iRMC) for fund management.

References

1. Park, M.S., Golden, K.J., Vizcaino-vickers, S., Jidong, D., Raj, S.: Sociocultural values, attitudes and risk factors associated with adolescent cyberbullying in east Asia : a systematic review. Cyberpsychol. J. Psychosoc. Res. Cybersp. **15** (2021)

2. Sood, S.M.M., Hua, T.K., Hamid, B.A.: Cyberbullying through intellect-related insults. Malaysian J. Commun. **36**, 278–297 (2020)
3. IPSOS: Malaysian and Global Views on Cyberbullying
4. Ghosh, R., Nowal, S.: Social media cyberbullying detection using machine learning in Bengali language. Int. J. Eng. Res. Technol. **10**, 190–193 (2021)
5. Van Hee, C., et al.: Automatic detection of cyberbullying in social media text. PLoS ONE **13**, 1–22 (2018). https://doi.org/10.1371/journal.pone.0203794
6. Saravanaraj, A., Sheeba, J.I., Devaneyan, S.P.: Automatic detection of cyberbullying from twitter. Int. J. Comput. Sci. Inf. Technol. Secur. **6**, 2249–9555 (2019)
7. Balakrishnan, V., Khan, S., Arabnia, H.R.: Improving cyberbullying detection using twitter users' psychological features and machine learning. Comput. Secur. **90**, 101710 (2020). https://doi.org/10.1016/j.cose.2019.101710
8. Salawu, S., He, Y., Lumsden, J.: Approaches to automated detection of cyberbullying: a survey. IEEE Trans. Affect. Comput. **11**, 3–24 (2020). https://doi.org/10.1109/TAFFC.2017.2761757
9. Özel, S.A., Saraç, E., Akdemir, S., Aksu, H.: Detection of cyberbullying on social media messages in Turkish. In: 2017 International Conference on Computer Science and Engineering (UBMK), pp. 366–370 (2017). https://doi.org/10.1109/UBMK.2017.8093411
10. Al-Garadi, M.A., Varathan, K.D., Ravana, S.D.: Cybercrime detection in online communications: the experimental case of cyberbullying detection in the Twitter network. Comput. Human Behav. **63**, 433–443 (2016). https://doi.org/10.1016/j.chb.2016.05.051
11. Al-Ajlan, M.A., Ykhlef, M.: Optimized twitter cyberbullying detection based on deep learning. In: 2018 21st Saudi Computer Society National Computer Conference (NCC), pp. 1–5 (2018). https://doi.org/10.1109/NCG.2018.8593146
12. Chatzakou, D., et al.: Detecting cyberbullying and cyberaggression in social media. ACM Trans. Web. **13** (2019). https://doi.org/10.1145/3343484
13. Dadvar, M., de Jong, F.: Cyberbullying detection: a step toward a safer internet yard. In: Proceedings of the 21st International Conference on World Wide Web. pp. 121–126. Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2187980.2187995
14. Ng, K.: NO SPOILERS!. https://chrome.google.com/webstore/detail/no-spoilers/anbpdfddbjchiihmibgakojddmhbfmeb?hl=en
15. Krishnakanth, A., Mahalakshumi, V., Vignesh, S., Nivetha, M.: CODAR – Cyber Offense Detecting and Reporting Framework. https://github.com/axenhammer/CODAR. Last Accessed 30 June 2021