Research Article

CYBERGUARD: A Mobile Cyber bullying Detector

Siok Yee Tan[1*], Kevin Tze Herng Choo[2]

[1,2]Center for Artificial Intelligence Technology, Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
E-mail: *esther@ukm.edu.my

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Social media is something that is used by a lot of people in Malaysia, especially among youths today. Social media can come in many different platforms such as facebook and twitter. However, due to the convenience given, a person can send a message to another person through social media in the blink of an eye. Many cases of bullying occur online, and these cases are known as cyber bullying. Cyber bullying happens when a person bullies and harasses another user online by sending harassing messages such as messages with rude words or messages with sexual harassment elements. Cyber bullying can be dangerous, affecting a person's emotion, especially among youths, because they have no experience with cyber bullying and doesn't know how to overcome this problem. As such, an android-based application named cyber guard has been developed to overcome this problem. This application is developed by connecting the user's twitter profile to retrieve the messages from the user's twitter messages. This application will alert the user if the application detects a negative implication's word retrieved from their twitter account. This application will also explain the meaning of the negative implication's word and their threat level. For users experiencing cyber bullying, the application provides advice and ways to overcome cyber bullying issues. Fifteen users have tested this application and the application has received good feedbacks from the users. This application is vital to everyone, especially youths, to know whether they are being cyber bullied and the steps that can be taken to overcome it**.**

Keywords: Cyber bully, Mobile application, Social Media, Text Retrieval

1. Introduction

Social media is an online media used by users to easily participate in social groups, create content, cover blogs, etc. Users use many types of social media based on their needs and interest. Due to the increasing ease of digital use, even children can engage in social media activities from a young age. Although many users use social media all around the world; however, social media can also be used as a medium for bullying. Due to the ability to be anonymous on social media, many users are taking advantage of social media to harass or bully others without being known publicly. This bullying case is known as cyber bullying, which are bullying cases that happen through social media using electronic devices such as hand phones and computers. Cyber bullying cases frequently occur in social media through text messages or social media websites such as Facebook or Twitter.

Cyber bullying cases are one of the biggest worries of Malaysians, especially among the teenager group because they are the highest social media users (Ismail 2014) and are also the easiest target for cyber bullying (Vaillancourt et al. 2017). A survey study carried out by Ipsos in 2018 showed that 23% of parents who answered the survey have their children experienced cyber bullying (Ipsos, 2018). Cyber bullying is also a dangerous situation as this type of bullying can disrupt their lifestyle since cyber bullying cases could be a persistent problem unless action is taken to resolve it (Peebles 2014). Cyber bullying can be a silent killer as it can lead to cases for victims being depressed and humiliated. The victims may find themselves helpless as they cannot solve the issues from these cases and might not know who to open up to help solve these issues. As such, cyber bullying may lead to the user’s mental health deterioration, which can cause many problems.

An Android-based application, namely Cyber guard, is proposed to overcome cyber bullying issues by detecting the negative implication's word retrieved from their Twitter account. The application will also describe each warning word and categorized the word accordingly. The application will provide advice for users who are currently experiencing cyber bullying to help users combat cyber bullying cases.

2. Literature Review

A few different applications were used as references for the Cyber guard Application features: the Bark, the Pocket Guardian, the Bully Alert and the Bull Stop application.

Bark

Bark Application is a mobile application developed by Bark Technologies, Inc. for Android and iOS (Bark Technologies 2018). The Bark application can be downloaded through Google Play Store for Android devices or iTunes for iOS devices. Bark Application is an application where parents can connect this app with their children's social media account so that parents can monitor their social media activity. This application will also alert the parents if their children's social media account contains signs of cyber bullying, sex implications, depression and more. The application can be connected to nearly 24 different social media websites that include popular websites like Facebook, Twitter and Instagram (T. Padmapriya et al).

The parents must first add and pair their children's social media account with this application to use this application. After the account has been paired with the application, the parents will monitor their children's social media activity. In addition to that, the parents will receive email or SMS notifications should the Bark application's algorithm detect any word that contains risk factors. The application will also give out a few advice about the known issues and help solve the problems.

The advantage of using this application is that this application can be connected to many social media websites. The application also allows the user to connect to multiple social media accounts at once. The application will also give detailed information about the words with negative implications if an alert is provided regarding any potential issues that could be happening. The disadvantage of using this application is that the application will not detect the problems if the user's children decide to change their social media account to prevent their parents from finding their accounts.

Pocket Guardian

Pocket Guardian Application is a mobile application developed by Pocket Guardian LLC for Android and iOS (Llc 2018). This application can be downloaded through Google Play Store for Android devices or iTunes for iOS devices. Pocket Guardian application is an application that can be used by parents to monitor their children's hand phone as well as their social media account without breaking their privacy. This application will only alert the parents if they detect cyber bullying implications or sexual implications in their children's hand phone.

The parents must open an account for this application and connect their email with this application to use this application. The parents must then input their children's information and connect their phone with this application. The application will then run in the background of their children's smartphones. When the Pocket Guardian application runs, an email alert notification will be sent to the parent's email account if the application detects cyber bullying implications in their children's text message or social media account. The email will inform the parents on which social media's account got alerted with cyber bullying, the type of bullying that could be potentially happening, and how their children use social media.

The advantage of using this application is that the application will run on their children's smartphones and only have to be opened once. The application will automatically run its function in the background. The application will also send an email to whoever the user set up in the Pocket Guardian's account, even if the connected email is not a user of Pocket Guardian. The disadvantage of this application is that the application will not show the full context of the bullying that could be happening since it only shows the type of bullying without the message's full context. This disadvantage could potentially provide some misunderstanding as the application could detect cyber bullying keywords and send an email to the parents without telling the full context on why their children's social media postings contain the following keywords.

Bully Alert

Bully Alert application is a mobile application developed by CU Cyber Safety Lab as a research project for the University of Colorado Boulder (Lab 2018). Bully Alert application is designed for Android and can be downloaded from Google Play Store. Anyone can use this application to observe another person's social media account to get the notification from their user's profile. This application, as of now, can only be used for an Instagram account. To use this application, the user must input the user's Instagram profile name in the application. After that, the application will start observing the Instagram profile inputted into the application and then send a notification if a cyber bullying case is happening on that Instagram profile. Through the notification, the user can view the context and detail of the message being sent from that Instagram profile.

The advantage of using this application is that it is easy to use and only requires the profile name of the person's Instagram to start observing the profile movement. The disadvantage of this application is that it can only be used if the user knows the user's Instagram profile name and the warning message does not inform the user to who that message is being sent.

Bull Stop

Bull Stop application is a mobile application developed for Android devices as part of a Ph.D. study at the University of Aston, United Kingdom (Salawu, 2020). The application is designed to help young people combat cyber bullying proactively, with its target audience being children aged 13 above. The application works by using a deep learning-based algorithm to analyse messages and flag any offensive content such as cyber bullying, abuse or insults. For this application, if the application detects some offensive contents in the user's social media account, it will automatically delete the messages to prevent them from reaching the user. Bull Stop application also allows users to block other user's social media from the application directly. This application currently only works on Twitter social media account with plans for Facebook and Instagram in the future.

This application's advantage is that it is easy to set up and will automatically delete the cyber bullying content just by running this application in the background. Another advantage of this application will also be that this application does not allow parents to monitor their children's social media account to help preserve the user's social media account privacy. However, the disadvantage of this application is that it does not show the messages that are deleted, which could delete a message accidentally if the algorithm flagged the messages wrongly. Another disadvantage of this application is that the user won't see the cyber bullying context as the message will be deleted before it reaches the user.

After reviewing the four existing applications, a comparison is made to compare the features of the existing applications and the features used for the Cyber guard application. The application comparisons are showed in Table 1.

Table 1. Comparisons with Existing Applications

Application	Function	Explanation	Cyber guard Application
Bark	Full message	The application will show the full details of the social media user's message with the cyber bullying category	Cyber guard's application will display the user's Tweet with full message and highlight the cyber bullying keyword to show the warning keywords.
Bark	Advice	The application will advise users based on the bullying category that is happening to the user.	Cyber guard's application will advise the user based on the word that is detected in the Tweet user.
Bully Alert	User Profile Searching	The application can scan any user profile by entering the user profile name.	Cyber guard's application has a search function to enter other users' Twitter profile names to check their timeline. The application will also be able to check the user's profile.
Bully Alert	Listing the cyber bullying words	The application will list all the warning words detected in the user's profile send a notification.	Cyber guard's application will show and list the warning word that is detected in the Tweet that is chosen by the user.
Bull Stop	Detecting cyber bullying implications in Social Medias	The application will scan through the entire social media account after connecting and automatically delete any messages that have cyber bullying implications.	Cyber guard's application will also scan through the user's social media account and will display the user's social media content for the user to be able to check messages with cyber bullying implications.

3. Methodology

Functional Requirement

Multiple functional requirements are designed to ensure the Cyber guard Application can fulfill the user's needs. The functional requirement that will be used for this application are as following:

a. Detect words that contain cyber bullying implications from the user's social media
- Users can log in to their Twitter account and connect their user's profile with this application.
- The application will scan through the user's Twitter profile and list out the list of Tweets that are made by the user's account. The user can then select the Tweet and the application will retrieve the Tweet's message.
- The application will highlight the warning words if the Tweet contains warning words saved in the database. The user can then select the word to see the description, category, and advice on potentially solving the issues.

b. A detailed description of cyber bullying word
- Users should be able to view detailed explanations on every word that contains cyber bullying implications in the application.
- Users can view every word that contains cyber bullying implications that are detected in the application.
- Users can also view the type of bullying that could be happening based on the word and the bullying category of the word detected by the application.

c. Giving Advice
- Users should be able to get advice based on the words detected that contain cyber bullying implications.
- The application can detect words containing cyber bullying elements and give advice based on the words detected from the user's Twitter account.

Context Diagram

The context diagram for the Cyber guard application is shown in Fig. 1.

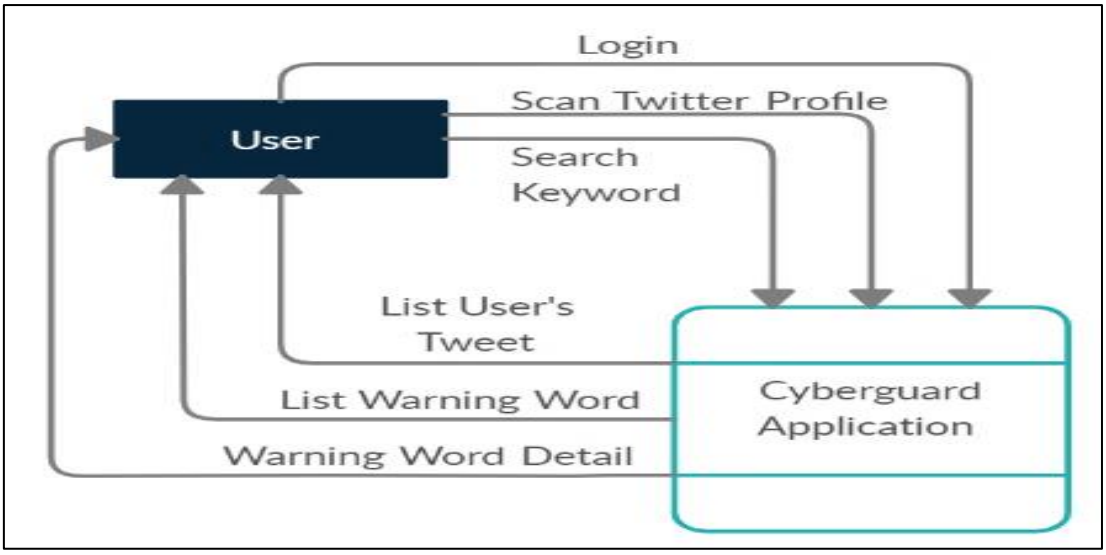

Figure 1. Context Diagram

The context diagram shows the context on how the application user connects with the Cyber guard application system. In this diagram, the user will use the application to log in to their profile. After the user has logged into their Twitter profile, they can scan their Twitter profile and use the application to search for a warning keyword. After scanning the warning word from Twitter's profile, the application will return a list of the user's profile Tweets. The list of warning words containing cyber bullying implications and the details for the cyber bullying warning word is displayed.

Use Case Diagram

The use case diagram for the Cyber guard application is shown in Fig. 2.

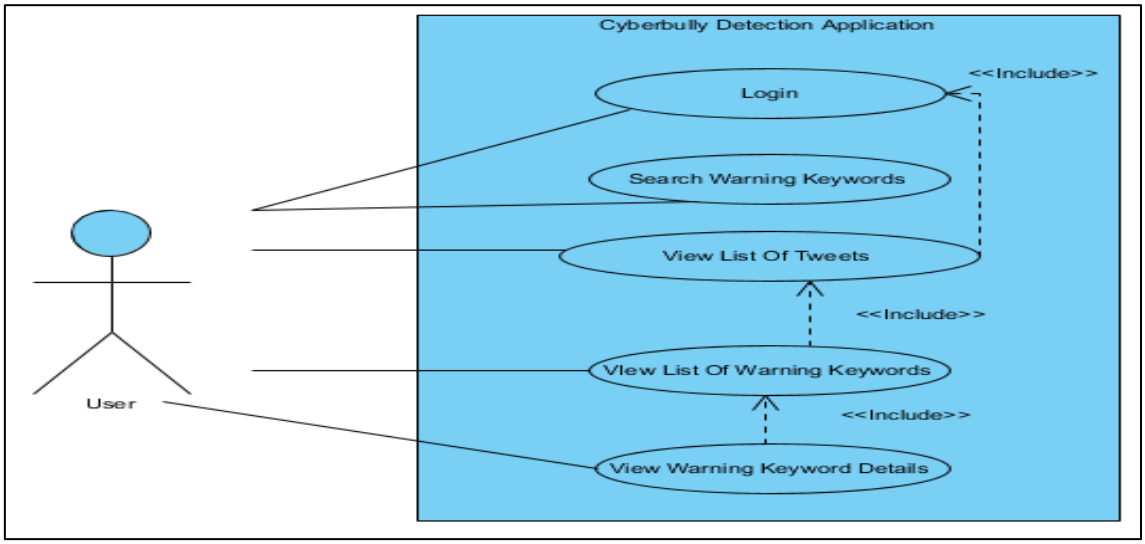

Figure 2. Use Case Diagram

The use case diagram above shows how the user can use the application. The user can start using the application by using the login function. After that, the user can search warning keywords that contain cyber bullying context manually. The user can also use this application to view Tweets' list from its own Twitter account or other people's Twitter Account. The user can then select the Tweet from the list to observe if that Tweet contains any keywords that have cyber bullying implications. The user can then view the warning keyword's details based on the selected keyword that the user chooses.

Use Case Specification

a. Table 2 shows the use case specification for login.

Table 2. Use Case Specification for Login

Topic	**Login**
Description	Users can access the application's functions after the user register and login into their Twitter profile through this application.
Main Actor	User
Preconditions	The application is opened.
Main Flow	1. The use case starts when the user clicks the "Sign In With Twitter" button. 2. The user enters their email and password. 2.1 The user enters the main page of the application.
Postconditions	The application enters the main page.
Alternate Flow	The user enters the wrong email or password 1. The application shows a message that the user's email or password contains a mistake or doesn't exist.

b. Table 3 shows the use case specification for search warning keywords.

Table. 3: Use Case Specification for Search Warning Keywords

Topic	**Search Warning Keywords.**
Description	Users will use the application to search for a warning keyword.
Main Actor	User
Preconditions	Users have already logged in to the application.
Main Flow	1. The use case starts when the user clicks the "Search Word" button on its main page. 2. The users choose and click one of the words in the list of warning words generated. 2.1 The users enter the warning word details page.
Postconditions	The application enters the warning word details page.
Alternate Flow	None

c. Table 4 shows the use case specification for view a list of tweets

Table 4. Use Case Specification for View a List of Tweets

Topic	**View a List of Tweets.**
Description	Users can view a list of Tweets from the user's profile or other user's profile.
Main Actor	User
Preconditions	Users have already logged in to the application.
Main Flow	1. The use case starts when the user clicks the "search timeline" or "search other timelines" button on its main page. 2. If the user clicks the "search timeline" button: 2.1 The application will show a list of tweets from the user's timeline. 3. If the user clicks the "search other timelines" button: 3.1 The application will open an input field to enter a user's profile name. 3.2 The application will show a list of Tweets from the inputted user's timeline.
Post conditions	The application will display a list of user's Tweet
Alternate Flow	User profile name does not exist 1. The application will show a 'No Tweet Detected" message on this page.

d. Table 5 shows the use case specification for view a list of warning keywords

Table 5. Use Case Specification for View a List of Warning Keywords

Topic	**View List of Warning Keywords**
Description	Users can view a list of warning keywords containing cyber bullying implications that are detected from the chosen Tweeter.
Main Actor	User
Preconditions	Users have generated a list of user's Tweets.
Main Flow	1. The use case starts when the user clicks on one of the Tweets from the list of user's Tweeter generated. 2. If the user clicks the "Search" button: 2.1 The application will highlight the word that contains cyber bullying implications from the chosen Tweet. 2.2 The application will list out the words that contain cyber bullying implications in the chosen Tweets.
Post conditions	The application will display a list of warning keywords that contain cyber bullying implications.
Alternate Flow	User profile name does not exist 1. The use case starts when the user clicks one of the Tweets from the list of user's Tweeter that is generated.

e. Table 6 shows the use case specification for view warning keyword details

Table 6. Use Case Specification for View Warning Keyword Details

Topic	**View Warning Keyword Details.**
Description	Users can view the details of words that contain cyberbullying implications.
Main Actor	User
Preconditions	Users have chosen a warning keyword from the list of words.
Main Flow	1. The use case starts when the user chose one word from a list of warning keywords. 2. The users can view the description, category, and advice based on the chosen word from the warning keywords list.
Post conditions	The application will display details of warning keywords.
Alternate Flow	None.

Activity Diagram

The activity diagram for the Cyber guard application is shown in Fig. 3.

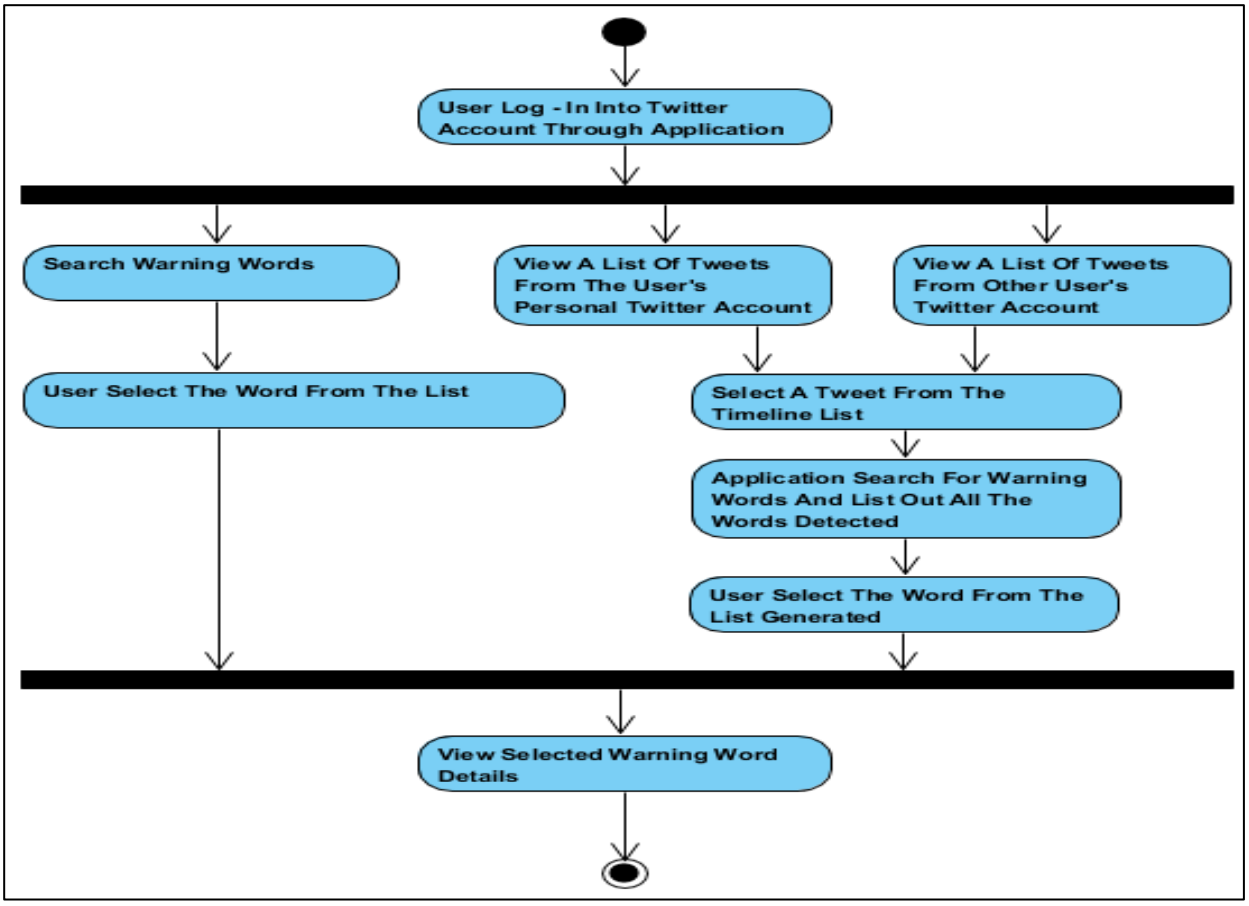

Figure 3. Activity Diagram for Cyber guard Application

The activity diagram shows the activity phase that is happening in the application. To start using the application, the user has to login to their Twitter account before it can be used. The application can then search for warning words or view a list of Tweets from the user's profile or other people's profile. A list of warning words will be displayed and the user can choose one of the words to view the detail of that word in the search warning words function. If the user wants to see warning words from their Twitter social media account, they have to use it to list the Tweets from the user's profile or other people's profile. A list of Tweets will be generated and the user can then choose one of the Tweets from the list. The application can then search and detect warning words from the chosen Tweets. The user can then select the warning word to view the detail of the warning word. The application activity ends once the warning word is displayed to the user.

4. Development

Twitter Validation

Before the Cyber guard application can be used with Twitter social media, a developer account for Twitter must be registered on the Twitter Developer website to obtain a unique API key to allow the application to use Twitter social media functions. The application that is being developed must be registered to generate the specific application's unique API key. The unique API key will then be stored in the application resources as a string so that the application can be connected to Twitter Developer. The Twitter account that has logged in using the API key will request authentication from the user to access the user's profile through the Cyberguard application. Fig. 4 shows the display of the API key generated from the Twitter Developer.

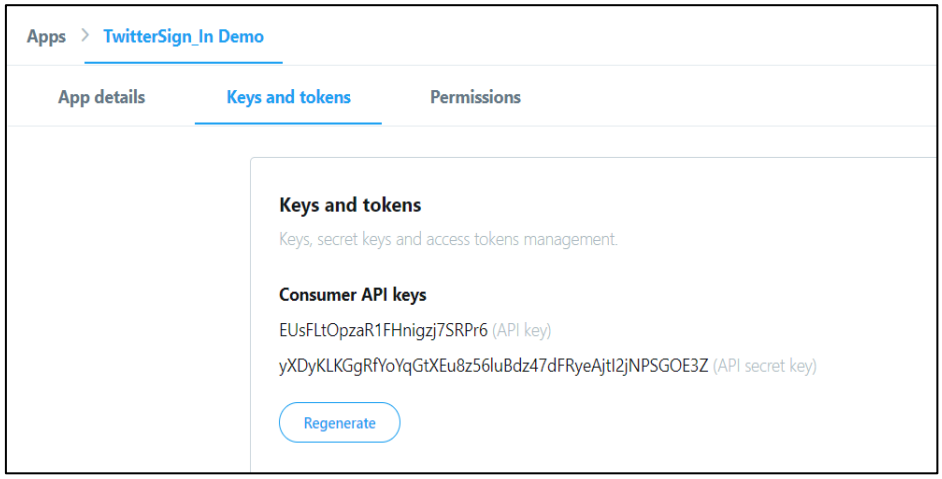

Figure 4. Display of API Key Generated from Twitter Developer

The application needs to get the account details from the user's authenticated account to use Twitter social media details. The function used to complete this action is referred from the Andro hub website (Kumar 2018). Some modifications have been made to the codes to get the user's profile name and the email from the search user's timeline function. The code referred from Androhub also contains a function to obtain the user's profile picture Uniform Resource Locator (URL). The profile picture URL is then used as a reference to display the user's profile image in the application as part of the user interface of the Cyber guard application. The code referred from the Androhub website can only be used to log in to the user's Twitter profile and show their user details such as their user ID, profile name and email. For Cyber guard application usage, the Twitter profile name will be used as an Intent to be sent to the next activity to generate the user's timeline.

Display Real-Time Data using Firebase

Firebase Real time Database is used by this application to store the warning keywords for the Cyber guard application. For this application, every warning word with its keyword will be added to the database. Each word will contain data for the word details which are descriptions, category, sub-category and advice.

The description for each keyword in the database is taken from a web dictionary to give the full word meanings (Thinkmap 2016). The warning keywords will also be categorized based on research done to categorized words with cyber bullying implications (Golbeck et al. 2017). In the study about cyber bullying categorization, the words are divided into several categories based on whether those words are harassment, potentially offensive and not offensive. The warning word will also be sub-categorized based on whether the word is direct harassment or indirect harassment. For instance, the word "die" is categorized as harassment as if it is used for cyber bullying, which is usually used as a slur directed at someone. The database will also store advice for each word shown to the user through the application. A sample display of the data stored in the database is shown in Fig. 5.

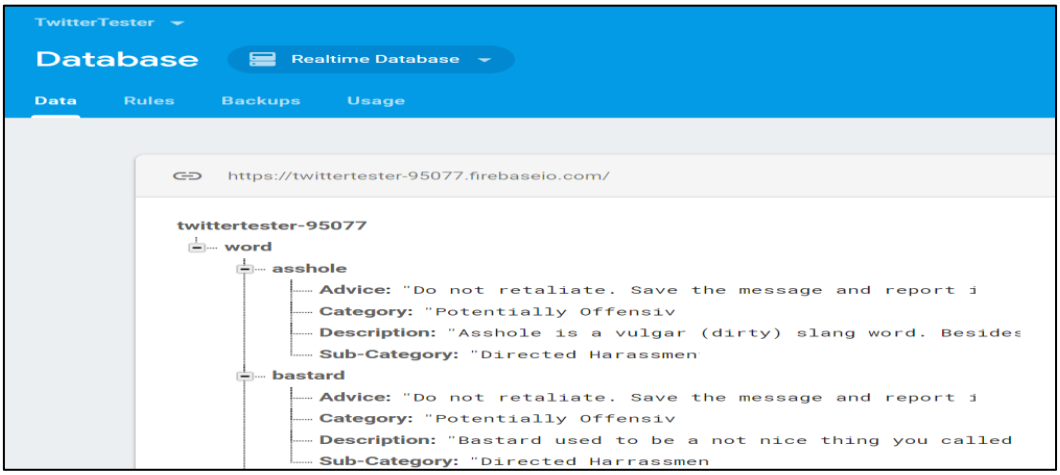

Figure 5. Display of Firebase Real-Time Database

5. Output

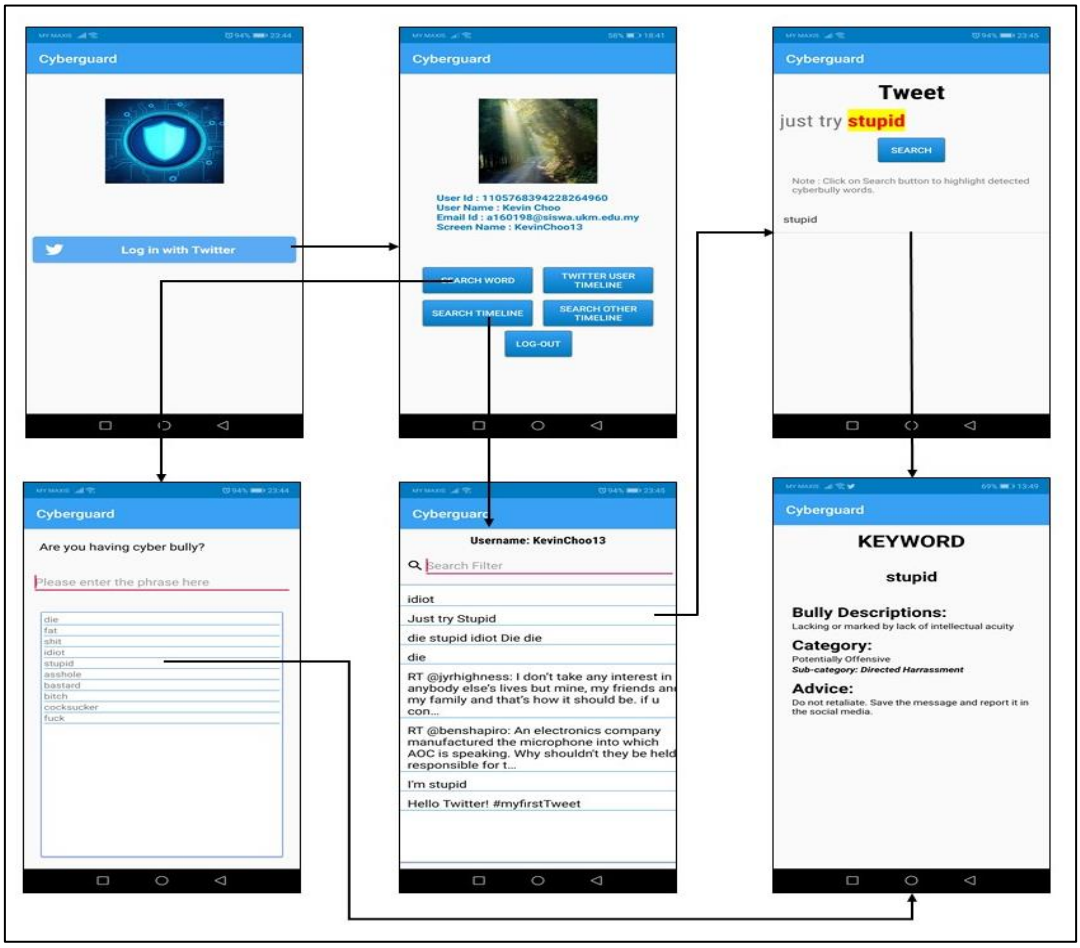

Figure 6. The Cyber guard Application Output

6. Evaluation Procedure

The usability evaluation for the Cyber guard application is conducted based on five factors: usefulness, effectiveness, efficiency, ease of use and aesthetics. For this evaluation, fifteen users with age ranging from 18-30 were given 10 to 20 minutes to test the application with guidance. They evaluate the application through a questionnaire after they have finished testing the application. The questionnaire involves a series of questions based on the five factors using a 5 points Likert Scale rating ranging from (1) Strongly disagree to (5) Strongly Agree. The evaluation results are tabulated to calculate the mean and standard deviation for the usability evaluation results. The usability evaluation result is shown in Table 7.

Table 7. Usability Evaluation Result

Factor	Mean ± SD
Usefulness	**4.28 ± 0.78**
The information and function in the Cyber guard application are useful to me.	4.27 ± 0.68
The application meets my needs.	4.07 ± 0.93
Using the application is a good idea.	4.60 ± 0.71
The Cyber guard application has all the functionality and capabilities that I expect.	4.20 ± 0.75
Effectiveness	**4.40 ± 0.65**
In my view, the information displayed by the Cyber guard application is accurate and complete.	4.27 ± 0.68
In my view, the functions and data in the application produce accurate output.	4.60 ± 0.49
The terms on the screen are clear and not confusing	4.33 ± 0.70
Efficiency	**4.33 ± 0.70**
The applications save my time when I use it.	4.33 ± 0.70
Tasks on the Cyber guard application were completed within a reasonable time.	4.27 ± 0.68
The Cyber guard application is simple and only necessary data is required.	4.40 ± 0.71

Ease of use	**4.36 ± 0.79**
Content on the application is straightforward and easy to read.	4.53 ± 0.50
It requires the fewest steps possible to accomplish what I want to do with it.	4.67 ± 0.47
It is easy to interpret or understand the displayed information and output on the Cyber guard application.	4.33 ± 0.70
I can use it without written instructions.	3.80 ± 1.16
The input and text entry methods for this application is easy.	4.47 ± 0.62
Aesthetics	**4.33 ± 0.68**
I am happy with the design of the Cyber guard application.	4.40 ± 0.71
I found the user interface of Cyber guard attractive.	4.13 ± 0.72
I think using the application would be pleasant.	4.27 ± 0.68
The design of the graphic symbols, icons and labels on the icons are sufficiently relevant.	4.33 ± 0.60
Generally, I am satisfied with the Cyber guard mobile application.	4.53 ± 0.62

Based on the results in Table 7, the results from the evaluation of the application are generally positive as the mean for the factors evaluated are all more than 4.28 and the majority of the questions answered have a mean above 4. Based on the evaluation, this application is a good idea that implies that the respondents think using this application is useful and beneficial. The only question with a mean lower than 4.0 is the question on if the user can use this application without written instructions implying that this application might not be simple to use without proper instructions. Hence, this application had upgraded with an appropriate instruction menu.

When compared with the applications stated in the literature review, only the Bark application and Bull Stop application will be compared as only these two applications have public reviews shown on their Google Play Store page. The ratings in Google Play Store uses the same Likert Scale rating which is one as the lowest rating and five as the highest rating. When compared with Bark application, Bark application has 841 reviews on their application page with an average rating of 3.4 which is lower than the mean results for all the factors tested by Cyber guard. However, the number of reviews on the Bark application is significantly higher than the number of respondents who tried Cyber guard's application which could skew the results. In comparison to the Bull Stop application however, the Bull stop application has 13 public reviews on their application page with an average rating of 4.8 which is higher than all the mean ratings for the Cyber guard questions. Despite the fact that the lower number of reviews than Cyber guard, Bull Stop has all their user's ratings at 4 or 5, showing that the users are generally more satisfied with using the Bull Stop application than Cyber guard. Hence, the researcher can conclude that the function “automatically delete any message that has cyber bullying implications” implemented by the Bull Stop application is an essential function in a cyber bullying detector's application.

7. Conclusion and Future Work

In conclusion, the Cyber guard application is developed as an easy-to-use application for users to tackle cyber bullying issues. The study discusses the existing application to learn the strength and weaknesses and as a reference for the design of the Cyber guard application. The system specification and design are also done to ensure that the application development process runs smoothly.

The success of the development of this study is that the system and design have been identified. A prototype of the application has been built with the program design. For future work, a suggestion to improve the application is to have the application work with more social media account such as Facebook and Instagram. This is because the application only works with Twitter, while citizens of this era especially teenagers tend to use more than one social media account. Another suggestion to improve this application is to allow this application to block the account of other users who are sending messages with cyber bullying implications for the user's ease of use. The researcher should also implement the automatically delete function proposed by the Bull Stop application to prevent the cyber bullying implication message from reaching the users.

8. Acknowledgements

This research work was supported by the FRGS/1/2018/ICT01/UKM/02/5 from Ministry of Higher Education Malaysia and the GGPM-2018-011 from Universiti Kebangsaan Malaysia (UKM) Grant Scheme.

References

1. Bark Technologies, I. 2018. Bark - Monitor. Detect. Alert. https://play.google.com/store/apps/details?id=cm.pt.bark parent&hl=en [2 June 2020].

2. Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., Ramachandran, P., Rogers, K. M., Rogers, L., Sarin, M. S., Shahane, G., Thanki, J., Vengataraman, P., Wan, Z. & Wu, D. M. 2017. A Large Labeled Corpus for Online Harassment Research. Proceedings of the 2017 ACM on Web Science Conference. Anjuran ACM. Troy, New York, USA, Huscher, B. 2000. Database Design and Modeling Fundamentals. [16 April 2020].
3. Ipsos. (2018). Malaysian and Global Views on Cyberbullying. Retrieved 4 November 2020, from https://www.ipsos.com/en-my/malaysian-and-global-views-cyberbullying
4. Ismail, N. 2014. Young People's Use of New Media through Communities of Practice.
5. Kumar, S. 2018. Android Twitter Integration. http://www.androhub.com/android-twitter-integration/ [13 March 2020].
6. Lab, C. C. 2018. Bully Alert. https://play.google.com/store/apps/details?id=com.bamboo.bullyalert&hl=en [2 June 2020].
7. Llc, P. 2018. Pocketguardian Parents - Stop Cyberbullying. https://play.google.com/store/apps/details?id=com.gopocketguardian.parentapp&pcampaignid=MKT-Other-global-all-co-prtnr-py-PartBadge-Mar2515-1 [2 June 2020].
8. Peebles, E. 2014. Cyberbullying: Hiding Behind the Screen. Paediatrics & child health 19(10): 527-528.
9. Salawu, S. 2020. BullStop. https://play.google.com/store/apps/details?id=mobile.bullstop.io&hl=en&gl=US [3 November 2020]
10. Thinkmap. 2016. Vocabulary.Com Online Dictionairy. https://www.vocabulary.com/ [16 April 2020].
11. T. Padmapriya & S.V. Manikanthan, “Retracted: Security and Routing protocol for 5G wireless mobile networks”. IJIMT, 2020.
12. Vaillancourt, T., Faris, R. & Mishna, F. 2017. Cyberbullying in Children and Youth: Implications for Health and Clinical Practice. Canadian journal of psychiatry. Revue canadienne de psychiatrie 62(6): 368-373.