Hatred Comments Detection in Twitter using Deep Learning

KR Hingurage# and DS Vithanage

Department of ICT, Faculty of Technology, University of Ruhuna, Sri Lanka

#hingurage.kr@gmail.com

Abstract - Social media applications are the most popular web and mobile applications across the globe. In the meantime, however, this has resulted in the emergence of conflict and hatred by making online environments, particularly Twitter, uninviting for users. This issue typically affects individuals, organizations and governments because it can have a reasonable or unreasonable impact on someone's reputation, as well as could trigger discrimination, hostility and violence, which can lead to or include terrorism or atrocious crimes in society. Therefore, an accurate, efficient automatic model to detect and classify hate speech on Twitter is a particularly useful tool for the relevant authorities. This paper describes the detection and classification of hate speech on Twitter by using deep learning. This can fill in the gaps between current models with higher accuracy and reliability. Thus, this research is beneficial in several ways, such as the detection of hate speech in distinct categories, such as in toxic, severe toxic, obscene, threat, insult and identity hate. The developed application used deep-learning algorithms to find the number of occurrences of the words and semantic words. The LSTM model is used to train the data set and to get the probability values. The classes of hate speech were calculated against the training data set and were found to be above 72%. In conclusion, the developed method can help to detect and classify hate speech into six classes on Twitter.

Keywords: social media, Twitter, hate speech

I. INTRODUCTION

Social media is any digital tool that allows users to quickly create and share or exchange ideas with the public by building up virtual networks and communities. Social media applications are the most used web and mobile applications across the globe. There are different types of social media networks that are popular at the present time that can be categorized as social networks (Facebook, Twitter, LinkedIn.), Media Sharing Networks (Instagram, Snapchat, YouTube), Discussion Forums (Reddit, Quora, Digg), Bookmarking and Content Curation Networks (Pinterest, Flipboard), Consumer Review Networks (Yelp, Zomato, TripAdvisor) etc. People are spending more time than ever before on their smartphones. People are able to communicate with other people and share all the media, like messages, images, audio and videos, worldwide at their fingertips through the use of social media. It can assist users in interacting with one another, raising their voices against an unjust act or issue, sharing valuable information, spending free time, and many other identified uses. Getting informed about the latest news around the world, marketing/purchasing products and services, making friends are some advantages of social media. Other disadvantages are failing to classify good and bad, inappropriate postings/content, cyberbullying, spreading of gossip, rumors, publishing, racism, spreading terrorism, hate speech. In Sri Lanka, viral postings and news on social media occasionally contain disgrace and hate speech, frequently targeting political parties, politicians, celebrities, and reputable companies. Twitter has information that is considered important by users. Messages (tweets) from other users who follow them will appear on the home page to read. Users can not retweet or resend messages sent by other users. Twitter, as a microblogging platform, has a large and growing user base on Twitter, users publish short messages using 140 or fewer characters to "tweet" about their opinions on various topics and to share information or to have conversations with their followers. Promote violence against others or personally harass or assault them because of

their race, nationality, national origin, caste, sexual orientation, gender, gender identification, religious belief, age, disability, or serious illness. For example, targeting may be done in a variety of ways. Mentions, including a screenshot of a person, referring to someone by their full name. and so on, are all examples of targeting. It poses serious threats to democratic society's stability, civil rights security, and the rule of law. If left unaddressed, it may escalate to larger-scale incidents of aggression and confrontation. Hate speech is, in this context, an intense expression of intolerance that leads to crimes.

Hate speech has an effect on a number of current United Nations areas of operations, including human rights security, atrocity crime prevention, genocide prevention, and counter-terrorism. The root causes of violent extremism and terrorism; avoiding and responding to gender-based violence; improving civilian protection; refugee protection; combating all types of bigotry and discrimination; minorities must be protected, unity must be maintained, and women, girls, and young people must be engaged. As a result, combating hate speech necessitates a concerted effort that addresses the origins and generators of hate speech, as well as the broader consequences for victims and communities. Therefore, hate speech detection is critical on Twitter. Uther, there are various classes of hate speech in the twitter. *The comments of toxic class* are harsh, insulting, or likely to cause someone to quit a conversation. The severe toxic class means severe stage of the toxic comments. The comments of obscene class are appealing primarily to a voyeuristic interest in sexual activity show or describe sexual action in a plainly objectionable manner and lack genuine literary, aesthetic, political, or scientific. The comments belong to threat class means a statement of the desire to do harm, hurt, or damage. The insult *class* is contemptuous or sneering. The comments of in the identity hate class belongs to someone's sexual orientation or identity, aggressive or antagonistic towards them.

II. RELATED WORK AND MOTIVATION

(Ginting, 2019) recommended a method for detecting hate speech on Twitter. This method is built on the automated collection of unigrams and patterns from the training set. These patterns and unigrams are later used as functions of a machine learning algorithm, among other things. (Gitari,2015) employed the paper to develop a model classifier that employs sentiment analysis techniques, specifically subjectivity detection, to not only recognize and rank the polarity of sentiment expressions, but also to detect that a given sentence is subjective. (He,2013) presented a paper on an in-depth case study that uses text mining to analyze social media data in order to teach businesses how to conduct a social media strategic analysis and turn social media data into information for decision makers and e-marketers. Furthermore, this paper highlights that more enterprises are moving to social media sites like Facebook and Twitter to offer services and connect with consumers. (Jiang, 2013) developed a statistical method for mining Twitter data for drug effects. Adverse drug reactions have been one of the leading causes of death. This research has developed a statistical method for mining Twitter data for drug effects. The attempt to create an automatic method to derive possible drug effects from Twitter data is discussed in this article. This research has established a statistical framework for gathering, sorting, and evaluating Twitter data to search for drug effects. (Jianqiang, 2017) presented a study on sentiment analysis, which is the method of automatically identifying emotional or opinionated content in a text fragment, as well as determining the polarity of the text.

The aim of Twitter sentiment classification is to categorize a tweet's polarity as positive, negative, or neutral. (Kurniawan,2016) presented a research study to create a real-time traffic classification system using social network data. Preprocessing, feature extraction, and tweet classification are all stages of Twitter data processing based on text mining that employ three machine learning algorithms: Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT).In recent years, the utility of excluding stop words in the form of Twitter sentiment classification has been called into

question (Saif, 2014).Using pre-compiled stop word lists or more advanced methods for dynamic stop word detection, removing stop words from textual data is a common technique for reducing noise. However, in recent years, the utility of excluding stop words in the form of Twitter sentiment classification has been questioned. One of the major challenges that Twitter sentiment analysis approaches must overcome is the noisy design of the data provided by Twitter. Singh (2016) presented a paper based on text mining. To check the meaning and sentiment translation of the slang terms, the proposed preprocessing approach relies on the binding of slang words to other coexisting words. To find the bindings, this paper used n-grams and conditional random fields to verify the meaning of slang words.(Wakade,2016) showed how to use Weka data mining software to retrieve valuable information for classifying sentiment in Twitter tweets. Introduce a new system for pre-processing tweets to practice decision trees. (Watanabe, 2018) demonstrated how to use a computer to classify elements of hate speech in a text so that the speech can be understood later. Using the form of Multinomial Logistic Regression words. Wakade (2016) showed how to use Weka data mining software to retrieve valuable information for classifying sentiment in Twitter tweets. Introduce a new system for pre-processing tweets to practice decision trees. Watanabe (2018) demonstrated about the usage of a computer to classify elements of hate speech in a text therefore, the speech can be understood later using the form of Multinomial Logistic Regression. According to the existing studies summarized above consider only the detection of the hate speech in the social media. On the other hand, the accuracy of the previous researches are low. These observations on existing research triggered our motivation to detect and identify the hate speech into six categories with high accuracy in Twitter.

III. METHODOLOGY AND EXPERIMENTAL DESIGN.

The proposed model is capable of detecting various types of hate speech, such as toxic, severe toxic, obscene, threats, insults, and identity hate using deep learning algorithms. The labeled data set was gathered from Twitter. Thereafter, a set of deep learning operations has been applied to detect different types of hate speech if any as illustrated in the figure [1].

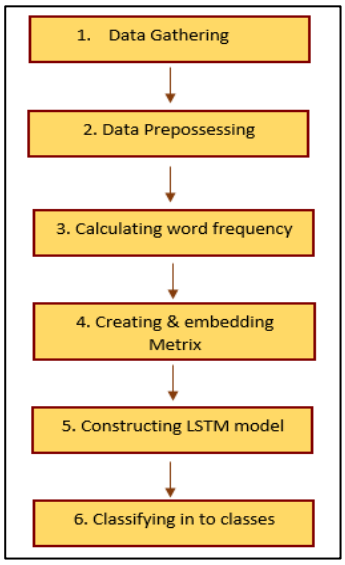

Figure 1. Methodology

The imported data set was preprocessed and tokenized by dropping unnecessary features which are not important for future processes. Following that, the number of clean and unclean entries was computed. Additionally, the distribution of the tag over the data set was counted and plotted on the data by using a bar chart. Furthermore, the frequency of the words was calculated in the clean and unclean data comments. Thereafter, the words frequented in clean and unclean data comments were displayed in a graphical way by appearing as the most frequented words, larger. The embedding matrix was constructed to find semantic words with the help of the Global Vectors for Word Representing (GLOVE) vector. An embedding metrix is a low-dimensional space into which high-dimensional vectors can be translated. Machine learning on big inputs, representing words, is made easier via embedding. An embedding should, in theory, capture some of the input's semantics by clustering semantically related inputs together in the embedding space.Next, the example for each category was displayed for the training purpose of the model.The Long Short-Term Memory (LSTM) model was used to train the proposed model because there might be lags of undetermined duration between critical occurrences in a time

series, LSTM networks are well-suited to categorizing, processing, and making predictions based on time series data. The LSTM model can learn a function that transfers a series of previous observations as input to a new observation as output. As a result, the series of observations must be converted into several examples for the LSTM to learn from. A probability of containing each category of hate speech was calculated for each uncleaned data comment in the dataset and finally taken to its apex.

IV. RESULTS AND DISCUSSION

In this section, the results of the proposed models are discussed. Jupiter 6.0 was used to create the model. About 15,957 data has been gathered from Twitter to detect the different types of hate speech based comments. The collected data was divided into two categories for the training and validation purposes in this research. The data preprocessing techniques were applied to the gathered data to remove hashes, tags, digits, punctuations, and id columns as illustrated in the figures [2] and [3].

```
In [4]: #Dropping the unecessary features
        train_data_or.drop(['id'],axis=1,inplace=True)
        # test_data.drop(['id'],axis=1,inplace=True)
```

Figure 2.Removing id column.

```
In [10]: #defining a function to clean the data
         def clean_text(text):

             text = text.lower ()
             text = re.sub(r"http[s]?://(?:[a-z][0-9]|[$-_@.&+][!*\(\),]|(?:%[0-9a-f][0-9a-f]))+', '', text) # clean url
             text = re.sub(r'#(\w+)', '', text)   # clean hashes
             text = re.sub(r'@(\w+)', '', text)   # clean @
             text = re.sub(r'<[^>]+>', '', text)  # clean tags
             text = re.sub(r'\d+', '', text)      # clean digits
             text = re.sub(r'[\_|@\'\"?\.$%_&#*+-:]', '', text)   # clean punctuation
             text = [APPO[word] if word in APPO else word for word in text.split()]  #

             return text
```

Figure 3. Removing special characters and URLs

The general distribution of the data set was checked and plotted on bar charts as figures [4] and [5]. Firstly, the amount of comments for each category were plotted as in the figure 4 and the percentage of each category were calculated and plotted as in the figure 5.

```
In [14]: plt.figure(figsize=(10,5))
         sns.countplot(x='comment_type', data=train_data)
         plt.xlabel("Type of comment")
         plt.ylabel("Number of comment")
         plt.show()
```

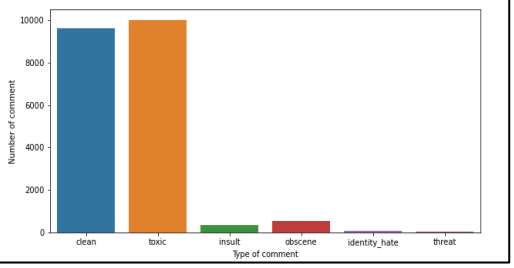

Figure 4. Amount of comments under each class

```
In [16]: plt.figure(figsize=(10,5))

         ax = (train_data.comment_type.value_counts()/len(train_data)*100).sort_index().plot(kind="bar", rot=0)
         ax.set_yticks(np.arange(0, 110, 10))

         for p in ax.patches:
             ax.annotate('{:.2f}%'.format(p.get_height()), (p.get_x()+0.15, p.get_height()+1))

         plt.xlabel("Type of comment")
         plt.ylabel("Number of comment")
         plt.show()
```

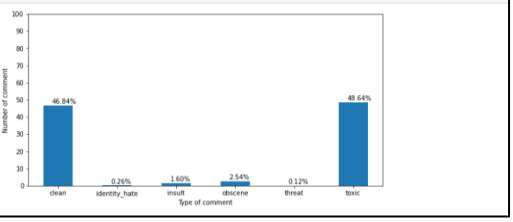

Figure 5. Percentage over the database

The frequency of word count for each category was calculated for the collected data set by making the most frequently occurring words larger, as shown in Figure [6]. Furthermore, as shown in Figure [7], the frequency of clean data counts was calculated.

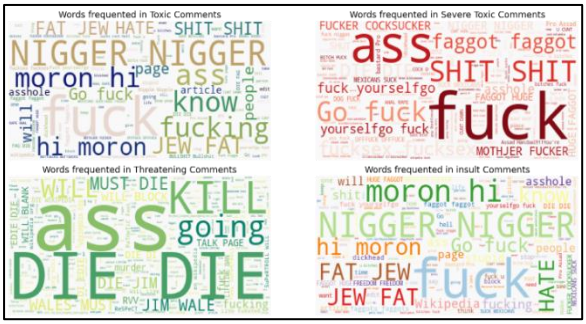

Figure 6. Wordcloud for comments based on hate

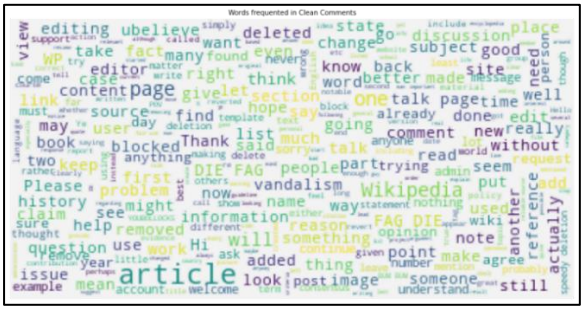

Figure 7. Wordcloud for clean comments

Furthermore, the model was trained by displaying examples of each type of hate speech comment, as shown in figure 8, and this is a critical operation in building the proposed model.

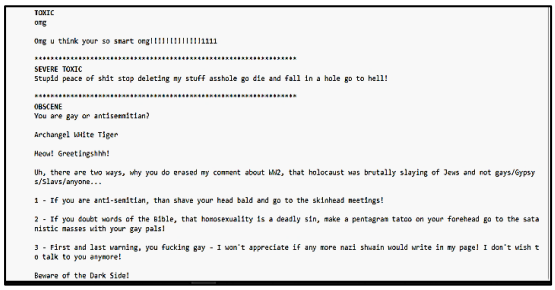

Figure 8. Training of the prediction model.

The probability of each class of hate speech was determined with the help of the LSTM model, as illustrated in the figure [9] at last.

	A	B	C	D	E	F	G	H
1	id	toxic	severe_to	obscene	threat	insult	identity_hate	
2	00001cee341fdb12	0.967946	0.246992	0.850563	0.086653	0.712358	0.212873	
3	0000247867823ef7	0.000117	8.40E-11	1.12E-06	2.84E-08	3.91E-06	9.08E-08	
4	00013b17ad220c46	0.001948	2.39E-08	5.33E-05	2.42E-06	9.74E-05	6.39E-06	
5	00017563c3f7919a	0.000425	4.44E-10	4.64E-06	1.14E-07	1.62E-05	4.29E-07	
6	00017695ad8997eb	0.001735	1.60E-08	3.74E-05	1.72E-06	8.11E-05	4.57E-06	
7	0001ea8717f6de06	0.000938	1.99E-09	8.69E-06	5.35E-07	3.04E-05	1.31E-06	
8	00024115d4cbde0f	0.000159	1.45E-10	1.97E-06	4.04E-08	6.62E-06	1.67E-07	
9	000247e83dcc1211	0.079323	1.11E-05	0.003154	0.000362	0.005844	0.000657	
10	00025358d4737918	0.076081	3.53E-06	0.005358	0.000153	0.007258	0.000491	
11	00026d1092fe71cc	0.000643	1.24E-09	1.06E-05	2.38E-07	2.59E-05	9.13E-07	
12	0002eadc3b301559	0.4506	0.000316	0.078433	0.001822	0.067325	0.004974	
13	0002f87b16116a7f	0.096474	3.05E-06	0.007487	0.000142	0.007025	0.000635	
14	0003806b11932181	0.000354	6.23E-11	3.49E-06	2.87E-08	7.66E-06	1.77E-07	
15	0003e1cccfd5a40a	4.34E-05	8.08E-12	4.01E-07	3.83E-09	1.43E-06	1.93E-08	
16	00059ace3e3e9a53	0.000128	6.14E-11	1.34E-06	2.13E-08	4.05E-06	1.00E-07	
17	000634272d0d44eb	0.008649	2.02E-07	0.000271	1.31E-05	0.000486	3.45E-05	
18	000663aff0fffc80	0.002528	4.04E-08	5.02E-05	4.31E-06	0.000134	8.17E-06	
19	000689dd34e20979	0.00302	1.98E-08	5.08E-05	2.28E-06	0.00014	6.35E-06	
20	000834769115370c	0.000194	1.25E-10	2.03E-06	4.36E-08	7.22E-06	1.49E-07	
21	000844b52dee5f3f	0.001326	1.13E-08	1.90E-05	1.57E-06	6.15E-05	3.47E-06	
22	00084da5d4ead7aa	0.003058	3.51E-08	6.39E-05	3.41E-06	0.000124	8.57E-06	
23	00091c35fa9d0465	0.419164	0.000719	0.05196	0.006396	0.104	0.01319	
24	000968ce11f5ee34	0.033688	3.02E-06	0.001315	0.000135	0.002851	0.000421	
25	0009734200a85047	0.000179	8.01E-11	1.85E-06	2.91E-08	6.48E-06	1.45E-07	
26	00097b6214686db5	0.207203	4.17E-05	0.010608	0.000894	0.020084	0.002431	
27	0009aef4bd9e1697	0.001067	3.53E-09	2.65E-05	5.77E-07	6.85E-05	1.93E-06	

subm

Figure 9. Probability values of having each class of hate

After the train the model, the actual predicated type of hate comments can be displayed finally as in the below figure [10].

```
Predicted type of comment: ['toxic', 'obscene', 'insult']
Actual type of comment: ['toxic', 'obscene', 'insult']
Comment:  Great, some furfag is undoing my quality opsts. Fuck you!

**********

Predicted type of comment: ['safe comment']
Actual type of comment: ['safe comment']
Comment:  "=Update===
Please consult the user talk pages of Gnan/garra and Shell for more discussion
between them of this block of me.  I really do not understand how this block
is warranted.  I explained that I could not change the archived material in a
AN/I because the instructions say not to alter it.  There are plenty of violations
of WP:CIVIL and Wikipedia:Etiquette made by others in that AN/I directed against me,
yet no one else is being made subject to a block, a ban, other kinds of ""sanctions""
or being asked to delete their comments about me in that archived material.  A
retroactive request to delete the record is not something that I am allowed to
act upon.  It would violate the notice in the archived record.  That is not my talk
page; that is an AN/I record.  To change it would change the record.  If the entire
thing is not expunged from Wikipedia, it seems to have to stay as is.
```

Figure 10. Predicted types of comments.

The developed model's accuracy was computed as shown in figure 11 using the validation data set, which was split at the beginning.

```
Train on 16447 samples, validate on 4112 samples
Epoch 1/2
16447/16447 [=======] - 27894s 2s/sample - loss 0.3349 - accuracy: 0.8542 -
val_loss: 0.2163 - val_accuracy: 0.9127
Epoch 1/2
16447/16447 [=======] - 58402s 4s/sample - loss 0.2092 - accuracy: 0.9127 -
val_loss: 0.2009 - val_accuracy: 0.9224
```

Figure 11. Accuracy of the model.

Furthermore, the accuracy of the system was calculated at different simulation times and it can be plotted as in figure 12 and, finally, the accuracy of the developed model was calculated at above 72%.

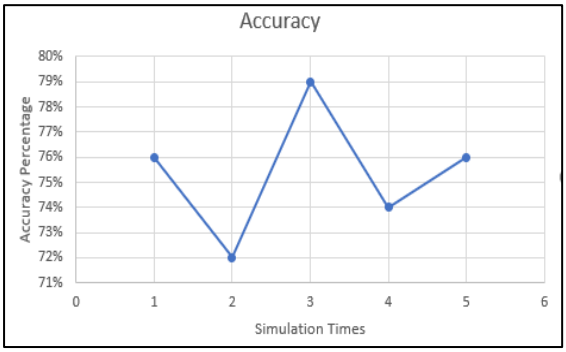

Figure 12. Accuracy of the model in different simulation times.

V. CONCLUSION

People continue to refuse to approve accounts whose primary aim is to harm others based on these categories. Targeting, such as it is, can be accomplished in some ways. Mentions, such as a snapshot of a person or addressing them by their first name, are both forms of targeting. It poses a significant danger to the integrity of a

democratic society, the protection of human liberties, and the rule of law. It can escalate to larger-scale outbreaks of violence and conflict if left unaddressed. In this case, hate speech refers to an outburst of intolerance that contributes to hate crimes. Terrorist networks use social media to carry out large-scale, covert or overt ideological activities. In particular, this will have a negative impact on someone's ability to use and be on Twitter. This has a negative impact on someone's ability to use and be on Twitter. Deep learning algorithms are used to detect six types of hate speech on Twitter. This may help to encourage people to think before making statements that may embarrass other parties or have an impact on their future, personal, or professional lives. There has been a lot of research done on the detection of hate speech using deep learning with low accuracy.This study differs from previous studies in that it focuses on detecting and classifying data into six different classes of hate speech with greater accuracy.Hate speech can be detected and classified into classes with, an accuracy of above 72% by using this model. However, it is necessary to carry out further studies to enhance the accuracy of the system and the usability.

REFERENCES

Ginting, P., Irawan, B. and Setianingsih, C., 2019. Hate Speech Detection on Twitter Using Multinomial Logistic Regression Classification Method. *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*,.

Gitari, N., Zhang, Z., Damien, H. and Long, J., 2015. A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4), pp.215-230.

He, W., Zha, S. and Li, L., 2013. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3), pp.464-472.

Jiang, K. and Zheng, Y., 2013. Mining Twitter Data for Potential Drug Effects. *Springer-Verlag Berlin Heidelberg 2013*, pp.434–443.

Jianqiang, Z. and Xiaolin, G., 2017. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. *IEEE Access*, 5, pp.2870-2879

Kurniawan, D., Wibirama, S. and Setiawan, N., 2016. Real-time Traffic Classification with Twitter Data

Watanabe, H., Bouazizi, M. and Ohtsuki, T., 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. IEEE Access, 6, pp.13825-13835.

ACKNOWLEDGMENT

We would like to express our appreciation to Mr. Shantha Kumara (Consultant) for assisting us in gathering information about the spread of hate speech and for providing valuable advice. Without his help, this research would not be possible.

AUTHOR BIOGRAPHIES

[1] Ms. Kavinday Hingurage is an undergraduate in the Department of Information and Communication Technology, Faculty of Technology, University of Ruhuna, Sri Lanka. Currently, Ms Hingurage works at VizuaMatix (Pvt) Ltd. as an Application Security Engineer (trainee).

[2] Mrs. Dinithi Vithanage is a lecturer in the Department of Information and Communication Technology, Faculty of Technology, University of Ruhuna, Sri Lanka. Mrs Dinithi is currently reading PhD in information technology at the University of Wollongong, Australia. Furthermore, she completed her Masters with Master of Science (Computer Science) from University of Sri Jayewardenepura, Sri Lanka in 2020. She graduated with a first class in the Bachelor of Science (Honours - Information Technology) degree from General Sir John Kotelawala Defence University, Sri Lanka in 2017.