Analyzing Trends and Topics of Sinhala Hate Speech on Twitter: A Time Series Approach.

D.A. Rajapaksha
Faculty of Information Technology
university of Moratuwa
Moratuwa Sri Lanka
diliniayesha1234@gmail.com

Dr. Supunmali Ahangama
Faculty of Information Technology
university of Moratuwa
Moratuwa Sri Lanka
supunmali@uom.lk

M.D. Dushanthi
Faculty of Information Technology
university of Moratuwa
Moratuwa Sri Lanka
dushanthimadhushika3@gmail.com

K.D.G.I Madhurangi
Faculty of Information Technology
university of Moratuwa
Moratuwa Sri Lanka
kpgmadurangi@gmail.com

2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS) | 979-8-3503-2363-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIIS58898.2023.10253497

***Abstract*— With the exponential increase in popularity, social media has become an influential platform for expressing opinions on various topics. However, these conversations may contain both hate-promoting and non-hate-promoting content, and excessive promotion of hate content can significantly impact the masses. Hence, it is critical to identify trending periods of hate-promoting topics to avoid harmful incidents among individuals or society. Considering the previous experiences in Sri Lanka, it would be helpful to detect such discussions early. This paper proposes a deep learning-based approach to identify trending periods of hate topics on Twitter, specifically focusing on hate-promoting Sinhala topics. The proposed model can identify trending periods of hate-promoting content in real-time, providing valuable insights for policymakers and stakeholders to counter the spread of hate speech. This research has significant implications for mitigating the impact of hate speech in the Sinhala-speaking community, contributing towards a more inclusive and tolerant society.**

Keywords—LSTM, LDA, Autoencoder, Timeseries analysis, Hate Contents, Twitter, social media.

I. INTRODUCTION

In the last two decades, the world has seen exponential growth in technology, with social media platforms leading the pack. As of October 2022, two of the world's top five most visited websites are social media platforms [1]. Social media has revolutionized how people communicate and share their thoughts, news, and life events. With the ability to share opinions, news, and life events through messages and posts, social media has become crucial in connecting people worldwide, transcending age, race, orientation, and geography.

In Sri Lanka, popular social media platforms such as Facebook, Twitter, Instagram, and YouTube are commonly used in languages such as Sinhala, English, Singlish, and Tamil. As a result, personal interactions and relationships have become more visible and measurable due to the rapid expansion of social networking sites.

However, a significant amount of content shared on social media is hate-promoting and can perpetuate prejudice and intolerance in society. Social media platforms, especially Twitter, have become breeding grounds for hate speech. As a result, researchers and policymakers are increasingly concerned about the negative impact of hate speech on social media. Twitter has become a popular platform for analyzing hate speech trends and topics because of its vast user base and the availability of public data. Hate speech is any language used to attack, insult, or demean a particular group of people based on their race, ethnicity, religion, gender, or sexual orientation. Hate speech on Twitter can have negative consequences, including increased hostility, polarization, and violence. Early identification of trending hate speech can help prevent conflicts and protect society.

In this research paper, we will use a time series approach to analyze Twitter trends and topics of hate speech. By analyzing the data, we aim to understand the dynamics of hate speech on Twitter and the factors contributing to its prevalence. So that policymakers and social media platforms can develop more effective strategies for detecting and mitigating hate speech online.

Despite the research published on identifying or categorizing hate or offensive content, cyberbullying, or trends in social media, there is a notable research gap bringing all together to monitor or identify trends through shared content. There is an opportunity for improvements, especially in identifying and monitoring hate trends emerging in social media. A few recent incidents in Sri Lanka showed that hate content shared on social media sparked conflicts. Some recent incidents are "Aragalaya" and Conflicts between Muslims and Sinhala racists in "Darga" town. Therefore, this study aims to fill this research gap by analyzing Twitter trends and topics of Sinhala hate speech. The results were verified by comparing findings with actual timeline incidents.

This paper contains a brief introduction in section one, a Literature review of related works in section two, a Methodology used in section three, a Discussion of findings in section four, and a conclusion and future work in section five.

II. RELATED WORKS

As a result of social media becoming one of the primary communication methods in the present, monitoring changes, patterns, and trends on these platforms has become a hot topic. Understanding trends on social media can help answer the question "What is happening now?" and can provide significant advantages to market and governing entities.

These platforms have a zero-tolerance policy when it comes to hateful content. Each platform has its own guidelines and community standards to prevent the spread of hate speech. These standards are in place to ensure that the online community remains safe and inclusive. YouTube, as one of the major social media platforms, identify hate as "Content

979-8-3503-2363-4/23/$31.00 ©2023 IEEE

promoting violence or hatred against individuals or groups based on any of the following attributes: Age, caste, disability, ethnicity, gender identity, and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin and veteran status." [2]. Facebook community standards define hate as "Direct attacks against people – rather than concepts or institutions – based on protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust, or dismissal, cursing, and calls for exclusion or segregation." [3].

Previous studies on hate speech detection have used various approaches, including supervised [4]and unsupervised machine learning techniques [5]. Furthermore, natural language processing (NLP) algorithms and CNN are also applied.

Researchers have widely used machine learning classifiers to detect and classify native language-based text. Classifiers such as Logistic regression [6], Maximum Entropy, Support Vector Machine (SVM) [7], Random Forest, Naïve Bayes (NB) [8], XGBoost, K-means, and Decision tree classifier [9] are primarily used in native-language classification. Logistic regression, SVM, and NB are the most used models among these classifiers. However, XGBoost has been shown to perform better than other classification models.

Many researchers have been working on developing effective methods to identify and classify hate speech on different platforms. These studies have identified several categories of hate speech, including abusive, offensive, obsessive, aggressive, cyberbullying, insults, provocation, racism, sexism, threats, toxicity, severe toxicity, identity attack, abuse, and profanity [10] [11]. These categories can help to categorize different types of hate speech, making it easier to detect and respond to this harmful content [12].

Some studies have focused on identifying and classifying hate speech in English, while others have explored hate speech in less resource languages. The study by Guruge et al. [13] focuses on the spread of Sinhala hate content on social media, particularly among Twitter users in Sri Lanka. They have defined hate content as "Direct or Indirect, verbal, or nonverbal (stickers, emojis, symbols) expression which can hurt, feel negative feelings on someone's mind. The hated content distributors' target audience may differ according to the time, situation, and content. Considering intensity, user intent, and hate target when identifying a sentence as hated or not." The study introduced a supervised mechanism for detecting hate content on Twitter using an ensemble method. Which showed an accuracy of 63%, an F1 Score of 58%, a Precision of 61%, and a Recall of 58% when predicting hate content. The past two decades' records in Sri Lanka substantiate evidence of how hate contentment leads to conflicts and injustice toward people. Aliff [14] discuss post-war challenges and policy gaps hindering the transition to sustainable peace in Sri Lanka, particularly the violence against religious sites and members of religious communities.

Twitter has become an essential platform for analyzing and predicting political events worldwide. Various studies have highlighted Twitter's potential to predict election results in advance of official polls, including the 2010 British and Dutch elections by Broersma et al. [15] and the 2020 US Presidential election by Arthur et al. and Aditya et al. [16] [17]. Chew and Eysenbach's research focused on the 2009 H1N1 pandemic and found that specific health-related terms on Twitter indicated the public's concerns and demands, thus enabling health authorities to respond accordingly [18].

In contemplation of identifying how social media trends affect news creation, Madhushika [19] explores how social media, particularly Twitter, affects news dissemination in Sri Lankan contexts. They analyze Twitter trending topics to identify newsworthy content. The study suggests a set of features that add "news value" to a tweet and a Neural Network model with Clustering + Topic modeling to determine the topics shared in Sinhala. The elevated solution achieves an F1 score of 0.73 and an accuracy of 64.41%.In a study by Muhammad et al. [20], Twitter trends were defined as topics that are the subject of many posts within a short period. They used big data analytics and the Hadoop ecosystem with deep learning to perform language identification, specific segmentation, sentence boundary detection, and entity detection. They also used hashtags with the highest frequency to identify events within the trend. This method automatically detects the number of topics and key posts, mainly focusing on event identification rather than monitoring identified event behaviors with time.

Another study by Shams et al. [21] introduced an RNN LSTM model to analyze patterns of trends for a particular community, also focusing on hashtag monitoring. The researchers identified trends by analyzing changes in the slope of the time series. They used a batch function to enhance learning speed and a Min-Max scaler to normalize the data. However, due to the small amount of data, the method's accuracy was low, and the mean absolute error was 20.4935. A study by Asif et al. [22] used an ensemble deep learning model to analyze the sentiment of online hate speech towards women and migrants on social media. The study found that women are frequently targeted by hate speech and identified specific types of hate speech, such as objectification, sexual harassment, and sexist language.

Caldera et al. performed a time series-based trend analysis for hate speech on Twitter during the COVID-19 pandemic. Using a crowdsourcing approach, the researchers used tweets in Sinhala and classified them into hate, not hate, or neutral categories. A subsequent analysis was performed to identify the data set's seasonality, trend, cyclicity, or residuals. The researchers identified that hate comments had no specific seasonality and that any hate speech during an epidemic gradually decreased with time. In another study by Jayasekara and Ahangama. [23], a ranking algorithm was introduced for ranking topics on Twitter during the Easter attack period in Sri Lanka. Clustering algorithms were used to identify topics, and the identified topics were ranked based on trendiness, determined by the number of tweets, retweets, and favorites. The researchers evaluated the model's accuracy by comparing identified topics with local newspaper first-page articles, and it was found to have 69.29% accuracy.

Overall, time series analyses are favored for identifying trends or seasonality of a topic. Deldari et al. [24] suggested TS CP2 for time series. This novel self-supervised CPD approach learns an embedded representation from historical samples to forecast a future interval of a time series. Their approach was the first CPD method to use contrastive learning to extract a compact input representation and achieved higher accuracy and robustness than existing CPD methods.

These studies highlight the importance of monitoring trends on social media and the potential benefits of understanding these trends. While each study has its limitations and strengths, they collectively contribute to our understanding of trend analysis and its applications.

III. METHODOLOGY

Our method for analyzing trends in hate speech on social media involves four key stages, with the involvement of Time-Series Analysis.

1. Data Collection
2. Identification of Hate Content
3. Identify Topics followed by Latent Dirichlet Allocation (LDA) [25]
4. Identify Hate Trending Time Frames and Visualization

Fig. 1 illustrates our proposed architecture from classification to visualizing the results. Each step is described in followed sub-sections.

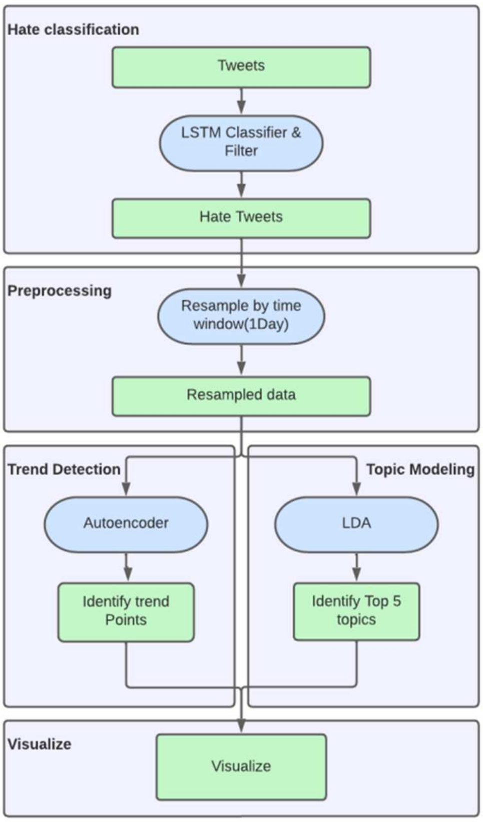

Fig. 1 Architecture diagram

.

A. *Data collection and hate classification.*

To collect the Twitter data, we utilized the Twitter API. We gathered a massive dataset of 8 million Sinhala tweets posted between May 2021 and May 2022. These tweets were then classified as either hate or non-hate using an LSTM model [26] [27]. The tweets were then preprocessed to remove stop words, stem the text, and tokenize the text. Out of the total collected tweets, our model successfully identified 1,732,731 tweets as containing hate speech.

B. *Preprocess hate content.*

Next, we resampled the hate speech tweets using a one-day time window, creating a final data frame with the timeframe, the number of hate tweets, and the content of all hate tweets for each day. This step facilitated the analysis of trends and topics in hate speech tweets over time.

C. *Identify topics.*

To identify the topics discussed in the hate speech, we used LDA [28] [29], a widely used topic modeling technique that analyzes co-occurrences of words and phrases to identify latent topics in a text. For each day, we identified five trending hate topics.

D. *Identify hate trending time frames.*

In the hate-trending time frames identification stage, we used 50% of the data as training data and 50% as test data. We considered hate content trending as an anomaly since it is shared daily. We calculated the standard normal to normalize the test data and created a sequence with NumPy. We then built a model with Keras sequential layers, using a convolutional reconstruction autoencoder [30] with Conv1D and Conv1DTranspose layers. The model was trained with 2000 epochs, and Fig. 2 depicts the autoencoder model's graph representation of the layered architecture. The training loss and validation loss graphs are shown in Fig. 3.

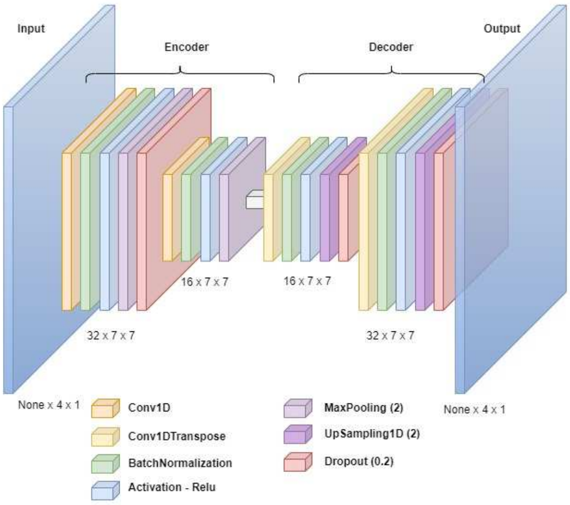

Fig. 2 Autoencoder Architecture

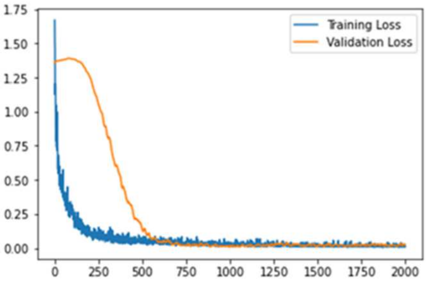

Fig. 3 Validation loss & Training Loss

We used Mean Absolute Error (MAE) loss to identify the hate speech trends. The maximum MAE loss was used as the threshold to identify any anomalies in the series. The maximum MAE loss occurs when the model's performance is at its worst, and we used 2 * maximum MAE loss to identify only the most apparent trend points.

E. *Visualization*

Trending data subsets were identified in this stage. Fifty-three samples with anomalies were detected and plotted with their trending topics. Fig. 4 shows the results of our research. Here in this graph Hate tweet count per day is plotted. Trending points are indicated in red with trending topics identified by LDA.

In data collection and hate classification, tweets were classified as hate or not with the LSTM model [26] [27]. Out of 8,000,000 Sinhala tweets, 1,732,731 tweets were selected as hate content. This dataset was later used for the subsequent stages of our analysis.

Data were pre-processed to analyze trends and topics discussed in hate speech over time. We resampled the hate tweets using a one-day time window. This resulted in a final data frame containing the number of hate tweets, their content, and the time period for each day.

We applied LDA to identify the trending topics in hate speech. We identified five topics for each day by analyzing co-occurrences of words and phrases in hate tweets.

We then used an autoencoder to identify the time periods when hate speech was trending. We trained a convolutional reconstruction autoencoder model with 2000 epochs using Keras sequential layers with Conv1D and Conv1DTranspose layers. We used MAE loss to identify the trend and identified the most apparent trend points by using 2 * Max MAE loss as a threshold.

Using a time series approach in our research allows for a more nuanced understanding of the evolution of hate speech over time. Autoencoder in our study had an Average Reconstruction error of around 0.1392. Further, our results allowed us to identify periods of increased hate speech-related activity and understand the leading discussion topics. Our results identified that the current political struggle initiated a hate content trend among Sinhala social media users around April 2022. The incident timeline was compared with the identified time of hate content sharing trend and topics discovered to validate the results, and they were compatible with each.

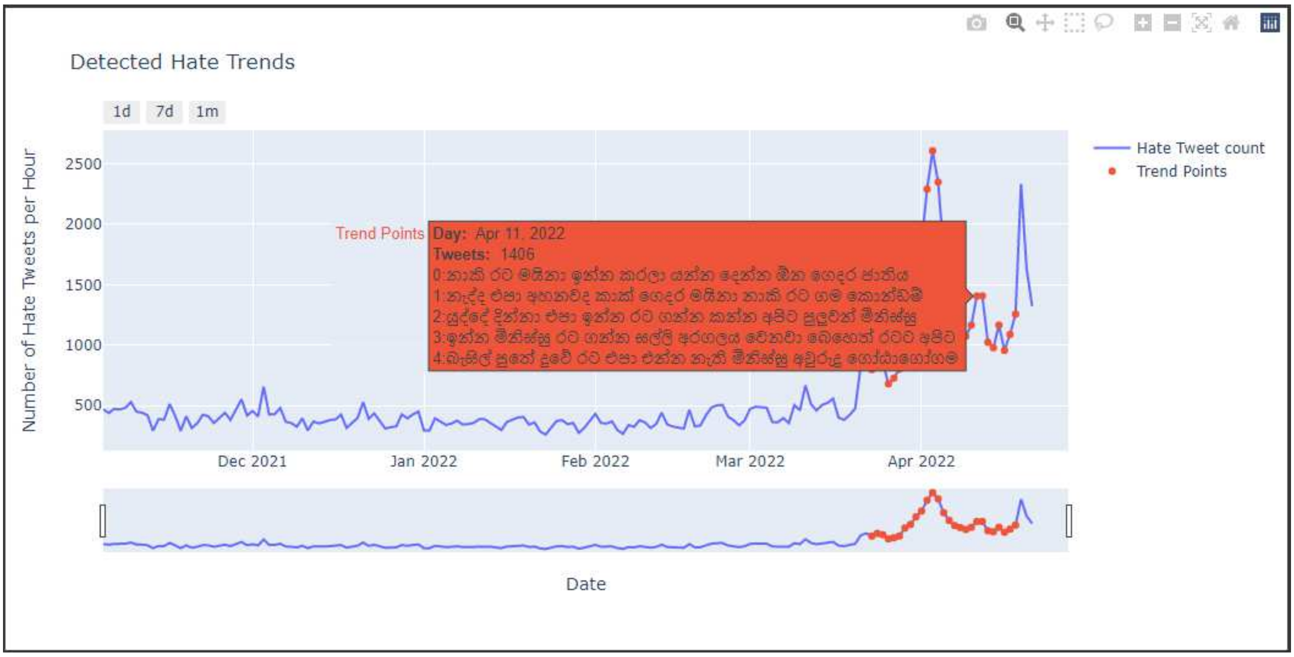

Fig. 4 Trend identification result

IV. Results and Analysis

In this study, we applied a novel approach to identify trends and topics of hate speech on social media using a combination of machine learning techniques. Our method consisted of four main stages: data collection and hate classification, preprocessing of hate content, identification of topics using LDA, and identification of hate-trending time frames using an autoencoder.

Our study has several limitations that should be addressed in future research. First, our approach was limited to Sinhala and may need to be more generalizable to other languages or platforms. It was tailored only to the Sinhala posts written with

Sinhala fonts. There is a requirement for improvement For the Singlish posts (Sinhala posts written with English characters). And it was only focused on identifying trends and topics but needed to explore why hate speech was prevalent in the identified time periods.

In conclusion, our approach provides a valuable tool for identifying and understanding trends and topics in hate speech on social media. Our findings highlight the need for more research and interventions to address hate speech targeting ethnicity, religion, politics, gender, and personal attacks on Sinhala social media.

V. Conclusion and further work

In summary, our study successfully identified trending periods of hate content and provided insights into the topics discussed in such content among Sinhala-speaking users on social media. Our model is capable of real-time monitoring, which has significant implications for understanding and mitigating the impact of hate speech in the community, promoting inclusivity and tolerance.

Despite these promising results, there are still areas for further research and development. First, extending our approach to analyze hate speech in other languages and social media platforms can provide a more comprehensive analysis. Secondly, exploring additional machine learning techniques can improve the accuracy and efficiency of hate speech detection. Lastly, identifying hate speech targeting specific aspects such as ethnicity, religion, politics, gender, and personal attacks on Sinhala social media can be a potential area for future research.

References

[1] "Top Sites in Sri Lanka, " Alexa Internet, Inc., [Online]. Available: https://www.alexa.com/topsites/countries/LK.

[2] YouTube, "YouTube Help," Y.T.O. Services, 2021. [Online]. Available: https://support.google.com/youtube/answer/2801939?hl=en. [Accessed 23 01 2022].

[3] T. center, "Hate Speech, Facebook community standards," Meta, 2021. [Online]. [Accessed 23 01 2022].

[4] E. . Ombui, M. . Karani and L. . Muchemi, "Annotation Framework for Hate Speech Identification in Tweets: Case Study of Tweets During Kenyan Elections,", 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8764868. [Accessed 16 3 2023].

[5] S. Rothe, "Supervised and unsupervised methods for learning representations of linguistic units," , 2017. [Online]. Available: https://edoc.ub.uni-muenchen.de/20938. [Accessed 16 3 2023].

[6] E. F. Unsvag and B. G. ack, "The Effects of User Features on Twitter Hate Speech Detection," in *Association for Computational Linguistics*, 2018.

[7] N. Hettiarachchi, R. Weerasinghe and R. Pushpanda, "Detecting Hate Speech in Social Media Articles in Romanized Sinhala," in *20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2020.

[8] I. Daga, A. Gupt, R. Vardhan and P. Mukherjee, "Prediction of Likes and Retweets Using Text Information Retrieval," in *Procedia Computer Science*, 2020.

[9] B. Karunanayake, U. Munasinghe, P. Demotte, L. Senevirathne and S. Ranathunga, "Sinhala Sentiment Lexicon Generation using Word Similarity," in *20th International Conference on Advances in ICT for Emerging Regions (ICTer 2020)*, 2020.

[10] M. Castelle, "The Linguistic Ideologies of Deep Abusive Language Classification".

[11] P. Fortuna, J. Soler-Company and L. Wanner, "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying?An Empirical Analysis of Hate Speech Datasets," in *12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020.

[12] J. Salminen, M. Hopf and S. Chowdhury, "Developing an online hate classifier for multiple social media platforms," 2020.

[13] M. Guruge, S. Ahangama and D. Amarasinghe, "Analyze Hate Contents on Sinhala Tweets using an Ensemble Method," in *2nd International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka, 2022.

[14] S. . Aliff, "Post-War Conflict in Sri Lanka: Violence against Sri Lankan Muslims and Buddhist Hegemony," *International Letters of Social and Humanistic Sciences,* vol. 59, no. 59, pp. 109-125, 2015.

[15] M. Broersma and T. Graham, "Social media as beat: Tweets as news source during the 2010 British and Dutch elections," *Journalism Practice,* pp. 403-419, 2012.

[16] A. Kim and P. Kim, "Estimation of the 2020 US Presidential Election Competition and Election Stratagies," in *IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2019.

[17] A. Singh, A. kumar, N. Dua, V. K. Mishra, D. Singh and A. Agrawal, "Predicting Elections Results using Social Media Activity A Case Study: USA Presidential Election 2020," in *7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2021.

[18] C. Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," *PLOS ONE,* 2010.

[19] M. D. Madhushika, S. Ahangama and D. A. Rajapaksha, "Analyzing the Impact of Social Media on Sinhala News Dissemination in Mass Media," in *2nd International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka, 2022.

[20] M. Ali, L. Liu and M. Farid, "Detecting Present Events to Predict Future: Detection and Evolution of Events on Twitter," in *IEEE Symposium on Service-Oriented System Engineering*, 2018.

[21] M. B. Shams, M. J. Hossain and S. R. H. Noori, "A Time Series Analysis of Trends With Twitter Hashtags Using LSTM," in *11th ICCCNT 2020*, Kharagpur, 2020.

[22] A. Hasan, T. Sharma, A. Khan and M. H. A. Al-Abyadh, "Analysing Hate Speech against Migrants and Women through Tweets Using Ensembled Deep Learning Model," *Computational Intelligence and Neuroscience,* 2022.

[23] L. Jayasekara and S. Ahangama, "Trend Detection in Sinhala Tweets Using Clustering and Ranking Algorithms," in *2020 From Innovation to Impact (FITI)*, 2020.

[24] S. Deldari, D. V. Smith, H. Xue and F. D. Salim, "Time Series Change Point Detection with Self-Supervised Contrastive Predictive Coding," in *WWW '21, April 19–23, 2021,*, 2021.

[25] L. . Lei, G. . Qiao, C. . Qimin and L. . Qitao, "LDA boost classification: boosting by topics," *EURASIP Journal on Advances in Signal Processing,* vol. 2012, no. 1, p. 233, 2012.

[26] J. . Lee, S. . Seo and Y. S. Choi, "Semantic Relation Classification via Bidirectional LSTM Networks with Entity-Aware Attention Using Latent Entity Typing," *Symmetry,* vol. 11, no. 6, p. 785, 2019.

[27] K. . Shuang, X. . Ren, J. . Chen, X. . Shan and P. . Xu, "Combining Word Order and CNN-LSTM for Sentence Sentiment Classification," , 2017. [Online]. Available: https://dl.acm.org/citation.cfm?id=3178230. [Accessed 13 3 2023].

[28] Y. . Song, S. . Pan, S. . Liu, M. X. Zhou and W. . Qian, "Topic and keyword re-ranking for LDA-based topic modeling," , 2009. [Online]. Available: https://dl.acm.org/citation.cfm?id=1646223. [Accessed 13 3 2023].

[29] L. . Alsumait, D. . Barbará, J. E. Gentle and C. . Domeniconi, "Topic significance ranking of LDA generative models," , 2009. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-04180-8_22. [Accessed 13 3 2023].

[30] H. . Tran and D. C. Hogg, "Anomaly Detection using a Convolutional Winner-Take-All Autoencoder.," , 2017. [Online]. Available: http://eprints.whiterose.ac.uk/121891. [Accessed 13 3 2023].

[31] s. jayasuriya, "sinhala-unicode-hate-speech," [Online]. Available: https://www.kaggle.com/sahanjayasuriya/sinhala-unicode-hate-speech.

[32] H.M.M.Caldera, G.S.N.Meedin and I. Perera, "Time Series Based Trend Analysis for Hate Speech in Twitter During COVID 19 Pandemic," in *20th International Conference on Advances in ICT for Emerging Regions (ICTer 2020)*, 2020.