CyberSuccor: Intelligence Technique for Sinhala Language Cyberbullying Detection on Social Media

Vajith Chamuditha
Computer Science and Engineering
University of Westminster
London, UK
vajithc@gmail.com

K.B.N.Lakmali
Department of Computing
Informatics Institute of Technology
Colombo 06, Sri Lanka
niwarthana.k@iit.ac.lk

***Abstract*—This research explores a deep learning-based approach to detect and categorize various forms of cyber bullying in Sinhala language text comments and text written within images. The primary objective of this research is to enhance the accuracy and effectiveness of cyber bullying detection, recognizing its significant impact on mental health. To process social media data, the system employs essential text preprocessing tasks derived from a comprehensive literature review. The evaluation of this system encompasses both subjective and objective measures, ensuring its quality and effectiveness.**

***Keywords*— Deep Learning, Ensemble Learning, Neural Networks, Natural Language Processing, Text Classification, Cyberbullying Detection**

I. Introduction

Bullying is defined as the behavior of fearing, forcing, or coercing another person to harass, physically dominate or threaten them. Bullying through electronic media is referred to as cyberbullying. This could manifest as the publication of gossip, threats, sexual remarks, personal details about the victim, or derogatory terms [1]. The Sinhalese, the largest ethnic group in Sri Lanka, use Sinhala as their native language. This research project aims to develop a new method for detecting cyberbullying in Sinhala language comments, including those in text and image format, with a focus on identifying negative comments in social media.

II. Problem Domain

A. Sinhala Language

Sinhala is the primary language of the Sinhalese community, which constitutes the largest ethnic group in Sri Lanka. Roughly 16 million people consider Sinhala as their first language, while an additional 3 million consider it as their second language [2].

B. Social Media

Social media refers to a group of web-based platforms that make it easier for users to create and share user-generated content. These platforms are constructed on the conceptual and technical underpinnings of Web 2.0 [3].

C. Cyberbullying

Cyberbullying refers to bullying through digital technology and can involve sending hurtful messages, sharing embarrassing media, or spreading rumors and lies online [4].

D. Natural Language Processing (NLP) in Sinhala Language Cyberbullying Detection and Classification

Cyberbullying cannot be easily manually detected because there are so many different types of social media data that are constantly changing, heterogeneous, and unstructured. Because of this, research into creating automated tools for spotting instances of cyberbullying has exploded [5]. Some researchers have suggested using machine learning and NLP techniques for autonomous detection of cyberbullying content to address the problem of cyberbullying and work towards reducing or eliminating it [6]. The prevention and mitigation of bullying are just two of the objectives these systems are meant to achieve. Additionally, according to Chia et al. (2002), these detection mechanisms can help both victims and perpetrators be identified.

III. Exsisting works

A. Sinhala Langugage Cyberbullying Detection

Numerous studies have been conducted to explore the application of machine learning and deep learning techniques for the detection of cyberbullying across various online platforms, including social media.

[7] Preprocessing involves lexical normalization, transformation of numerical data, removal of outbound vocabulary words, and elimination of repetitive or missing parameters. The BCO-FSS technique is used to extract features from the preprocessed data. SSA is used to optimize hyper-parameters of the DBN model for classification. The FSSDL-CBDC technique was demonstrated to have improved classification performance through simulations. The training process of the DBN model involves the use of a fitness function..

[8] conducted a study on hate speech detection on Twitter. Hateful content from Twitter was used to construct a dataset for this investigation. The dataset was annotated by three neutral annotators to ensure objectivity. Tokenization and stop word removal were used as preprocessing techniques. The primary aim of the research was to evaluate the performance of deep learning models in the context of hate speech. An ensemble of these deep learning models was produced using a majority vote. Along with deep learning techniques, conventional machine learning models were also used.

[9] created a gender-specific cyberbullying detection algorithm for the under-resourced Bangla language. They developed the GenDisc dataset, which has around 2600 data

samples, and used the K-fold cross-validation approach to improve the performance of their model on unreported data. The authors developed a framework using an ensemble method and four separate models to train a text classifier that discriminates based on gender. Various performance indicators, including Accuracy, Mcc Score, Precision, Recall, and F1, were used to analyze the models' performance and gain a deeper understanding of their effectiveness.

B. *Sinhala Language Multi-class text Classification*

[10] In the study on Sentiment Analysis for Sinhala Language, the researchers pre-processed the text by removing undefined characters, tokenizing the words using the Sinhala tokenizer from sinling4, and investigate the impact of word embedding dimension on sentiment analysis using FastText and Word2Vec models with dimensions ranging from 50 to 450. Performed sentiment analysis on the dataset using several neural network models, such as RNN, LSTM, GRU, and BiLSTM.

[11] conducted research on the classification of code-mixed text using capsule networks. The dataset was preprocessed to reduce noise by removing stop words and tokenizing the text. Text classification was performed using a Capsule + biGRU model, where CBOW word embeddings were used as the first layer of the neural network.

[12] study explores multiple machine learning algorithms for multiclass text classification. The dataset underwent preprocessing steps where punctuation, non-Sinhala letters, and stop words directly extracted from the corpus were removed. Various techniques including Logistic Regression, Naive Bayes, eXtreme Gradient Boosting, Random Forests, and Support Vector Machines were employed to categorize the text. To ensure consistency, the identical data from the DRAFT 4 split was used to run each algorithm five times, using random sampling for training data.

IV. DATASET

This research using a dataset, which is a combination of two datasets created using the Twitter API and Facebook comments obtained from the Kaggle website.

To obtain the necessary Facebook data for this research, the Sinhala Unicode Hate Speech dataset [13] which has both hate speech and neutral content was utilized. While Kaggle offers several datasets on Sinhala Cyberbullying (including hate speech and offensive content), this particular dataset was selected due to its high quality. The collection of Twitter data was accomplished through the utilization of the Twitter Standard search API. This involved conducting searches for tweets that contained pre-selected keywords, which were determined by Sinhala language experts. The collected data was labeled as Neutral, Offensive, Racist and Sexism with the assistance of experts proficient in the Sinhala language. After conducting the literature review, it concluded that manual labeling is a more effective method for Sinhala, given the language's extensive range.

V. METHODOLOGY

Two important aspects of this implementation process are dataset preprocessing and the implementation of the base models. Dataset preprocessing involves converting raw data into a clean, organized format suitable for use. Base model implementation, on the other hand, involves developing and training a model to perform the intended tasks using the preprocessed data. These are critical in laying the foundation for the system to operate efficiently and produce accurate results.

A. *Dataset preparation*

Twitter data was retrieved using the standard Twitter API endpoint by specifying relevant search parameters. The standard Twitter API provides an endpoint that enables the retrieval of Twitter data based on specified search parameters. Once the data was retrieved, it was stored in an Excel file for further analysis and processing.

B. *Data Preprocessing*

The first step in the process involved removing stop words. A list of stop words was created, and then each comment in the comment column was iterated through to remove any stop words present. This approach helps to eliminate common words that do not carry much meaning and can result in noisy data. Next, duplicate comments were removed, and the text was cleaned by removing URLs, mentions, retweet states, numbers, punctuation, and emojis. Then Sinhala characters were simplified because social media content has so much of spellings and other mistakes.

C. *Tokenization*

The training data was preprocessed using a subword tokenizer to generate input sequences for the model. The tokenizer was constructed based on the training data. The sentences were tokenized into subwords and padded to a maximum length of MAX_LEN using zeros. This preprocessing step ensured that the input sequences were uniformly long for efficient processing.

D. *Word Embeddings*

A word index dictionary is created to map subwords to unique indices. A pre-trained FastText model is loaded and the dimensionality of the word embeddings is determined. An embeddings matrix is initialized with zeros and dimensions corresponding to the vocabulary size and embedding dimension. For each subword in the word index, the corresponding word vector is retrieved from the FastText model and assigned to the respective position in the embeddings matrix.

E. *Implementation of Base Models*

a) LSTM model: First base model is LSTM model. It is made up of an embedding layer, an LSTM layer, a dense layer, a dropout layer, and an output layer with softmax activation. The model's performance is measured by accuracy and is built using categorical cross-entropy loss and the Adam optimizer.

b) ANN model: Second base model is an artificial neural network model with an input layer that takes in integer sequences, an embedding layer that translates the values of integers to a lower-dimensional vector space, and many dense layers with ReLU activation, dropout regularization, and a softmax activation function. The model is trained using the categorical cross-entropy loss function and optimized using the Adam optimizer.

c) DCNN model: Third base model is Deep Convolutional Neural Network (DCNN). It takes in an embedding matrix, vocabulary size, and other hyper-parameters as input, and returns a compiled Keras model. The model consists of two convolutional layers followed by a concatenation layer, a feedforward neural network layer, a dropout layer, and a softmax output layer. The model is designed to prevent overfitting by applying dropout only during training.

F. *Implementation of the Ensemble Model*

Two ensemble strategies, averaging and stacking, were explored to find the optimal way to acquire the final output from the ensemble model. Stacking was selected as the approach for deriving output from the base models. The outputs of the base models are concatenated and then passed through a final model that learns to combine the predictions of the individual models into a final output. This stacked ensemble is able to capture information from different aspects of the data, leading to a more robust and accurate final prediction.

Softmax activation function was employed to capture the all classes percentages. ReLU was employed in the intermediate layers to characterize complicated interactions between inputs and outputs, creating non-linearity. Adam optimizer was utilized because of its capability to calculate specific learning rates for distinct parameters. This eliminated the need for manual tuning of the learning rate. Categorical cross-entropy was chosen as the optimal loss function because it is widely utilized in multi-class classifications.

G. *Sinhala Optical Character Recognition*

The Tesseract library was used to implement optical character recognition (OCR). OCR's language setting is set to Sinhala. The OCR feature is used to process the image and extract the text from it. The resulting text is then saved and any newline characters have been removed. Finally, the extracted text passed to the classification model.

VI. Results and Discussion

The author performed evaluations on both the individual base models and the ensemble model since the final model is constructed using a technique that involves multiple base models collaborating. Before forming the ensemble model, the author employed different evaluation metrics to assess the selected base models.

A. *Evaluation of Base Models*

After testing base models with self-learned and Fast-Text (300) word embeddings, best approach was selected. Each model has been trained and tested with Self-learned word embeddings and with Fast-Text word embeddings. Base models are evaluated according to their validation accuracy.

B. *Evaluation of the Ensemble model*

a) Accuracy: The ensemble model was put to the test for a few situations while being trained with various parameters to find the model that performed the best in terms of high accuracy and low overfitting. The accuracy and loss variation versus the epoch count is shown in the image below.

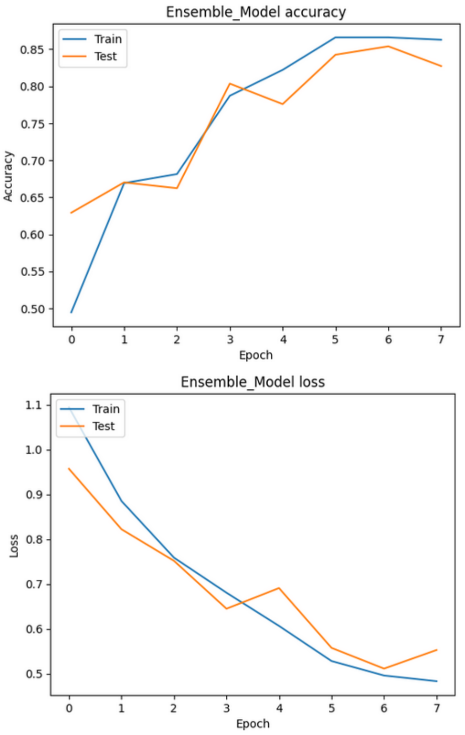
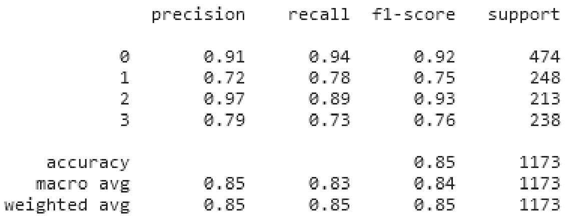


Figure 1: Model Accuracy and Loss against epochs

b) Classification Report: By analyzing the classification report, it can be seen that the model performed particularly well for two of the classes, with high ratio. However, it struggled a bit with the other two classes, which having a moderate ratio.



```
              precision    recall  f1-score   support

           0       0.91      0.94      0.92       474
           1       0.72      0.78      0.75       248
           2       0.97      0.89      0.93       213
           3       0.79      0.73      0.76       238

    accuracy                           0.85      1173
   macro avg       0.85      0.83      0.84      1173
weighted avg       0.85      0.85      0.85      1173
```

Figure 2: Classification report

C. *Performance Comparison*

Below table shows a comparison of accuracies of implemented models.

Evaluation matrix	LSTM model	DCNN model	ANN model	Ensemble model
Training accuracy	83.91%	84.74%	79.80%	90.13%
Testing accuracy	81.28%	80.99%	80.99%	85.38%

Table VI.1: Performance comparison

According to the evaluation metrics, the ensemble model outperformed the other individual models for the multi-class classification problem. Therefore, it can be concluded that the developed model successfully achieved its performance testing objectives.

D. *Benchmarking*

Since the study utilized a customized dataset and there is no prior work conducted on the exact dataset used in the implementation, it was not possible to compare the results against previous efforts. Additionally, due to the unavailability of public Sinhala multi-class datasets, the benchmarking was carried out using the same dataset but with different approaches.

Approach	SVM	CNN + LSTM	GRU	Proposed approach
Accuracy	75.56%	80.54%	80.19%	85.38%

Table VI.2: Benchmarking

The benchmarking mentioned above demonstrates that among the several approaches, the recommended technique was able to attain the best accuracy rate.

VII. Conclusion

The research was focused at bridging gaps in current knowledge and making contributions to both the technical and domain areas of the field. The primary objective of this research project was to utilize deep learning techniques for detecting instances of cyberbullying in the Sinhala language. Additionally, the aim was to develop and evaluate a novel system that surpasses the accuracy of existing approaches. The research successfully achieved its objectives, as evidenced by the testing and validation of the system. The proposed deep ensemble learning technique demonstrated superior performance compared to previous efforts in terms of accuracy. The findings of this research project highlight the potential of deep learning methods in addressing the challenge of cyberbullying detection in the Sinhala language, as well as the ability of using the proposed novel approach in Sinhala language multi class text classification. The developed system presents a promising solution that can effectively identify instances of cyberbullying with improved accuracy. These results contribute to the field of cyberbullying detection and emphasize the importance of leveraging deep learning approaches to tackle language-specific challenges.

VIII. future works

The author has identified several key areas for further improvement in the research project. Firstly, creating a large data corpus consisting of 20,000-30,000 records would provide a more comprehensive and diverse dataset for training and evaluation. Secondly, the model's performance can be enhanced by exploring different techniques to improve its accuracy. Additionally, while beyond the scope of the current project, incorporating the ability to detect cyberbullying content in audio and video formats would make the CyberSuccor system more applicable for community use. Lastly, enhancing the system's capability to detect cyberbullying in other commonly used languages in Sri Lanka, such as Tamil, English, and Singlish, would be a valuable future development. These improvements would contribute to a more robust and effective cyberbullying detection system.

Acknowledgment

The author expresses their gratitude towards the individuals who contributed to the construction of the paper and the development of the proposed system. The author appreciate the support received from the people involved in the project's development, as well as the valuable comments provided by the reviewers, which enriched the paper. The author extends special thanks to Ms. Niwarthana Kariyabaduge (Supervisor) and Mr. Buddhin Saroj (proof reader) for their guidance and advice throughout the project.

References

[1] A. Akhter, U. K. Acharjee, and M. M. A. Polash, “International Journal of Mathematical Sciences and Computing(IJMSC),” *International Journal of Mathematical Sciences and Computing(IJMSC)*, vol. 5, no. 4, p. 1.

[2] N. de Silva, “Sinhala Text Classification: Observations from the Perspective of a Resource Poor Language,” Jun. 2015.

[3] A.-S. T. Olanrewaju, M. A. Hossain, N. Whiteside, and P. Mercieca, “Social media and entrepreneurship research: A literature review,” *International Journal of Information Management*, vol. 50, pp. 90–110, Feb. 2020, doi: 10.1016/j.ijinfomgt.2019.05.011.

[4] G. W. Giumetti and R. M. Kowalski, “Cyberbullying via social media and well-being,” *Current Opinion in Psychology*, vol. 45, p. 101314, Jun. 2022, doi: 10.1016/j.copsyc.2022.101314.

[5] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. De Choudhury, “A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms,” *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, p. 325:1-325:34, Oct. 2021, doi: 10.1145/3476066.

[6] Y. Khang Hsien, Z. Arabee Abdul Salam, and V. Kasinathan, “Cyber Bullying Detection using Natural Language Processing (NLP) and Text Analytics,” in *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, Apr. 2022, pp. 1–4. doi: 10.1109/ICDCECE53908.2022.9792931.

[7] N. S *et al.*, “Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media,” *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–13, Jun. 2022, doi: 10.1155/2022/2163458.

[8] S. Munasinghe and U. Thayasivam, “A Deep Learning Ensemble Hate Speech Detection Approach for Sinhala Tweets,” in *2022 Moratuwa Engineering Research Conference (MERCon)*, Jul. 2022, pp. 1–6. doi: 10.1109/MERCon55799.2022.9906232.

[9] H. K. Rabib *et al.*, “Gender-based Cyberbullying Detection for Under-resourced Bangla Language,” in *2022 12th International Conference on Electrical and Computer Engineering (ICECE)*, Dec. 2022, pp. 104–107. doi: 10.1109/ICECE57408.2022.10088574.

[10] L. Senevirathne, P. Demotte, B. Karunanayake, U. Munasinghe, and S. Ranathunga, “Sentiment Analysis for Sinhala Language using Deep Learning Techniques.” arXiv, Nov. 14, 2020. doi: 10.48550/arXiv.2011.07280.

[11] S. Chaturanga and S. Ranathunga, “Classification of Code-Mixed Text Using Capsule Networks,” in Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications, INCOMA Ltd. Shoumen, BULGARIA, 2021, pp. 256–263. doi: 10.26615/978-954-452-072-4_030.

[12] V. Jayawickrama, A. Ranasinghe, D. C. Attanayake, and Y. Wijeratne, “A Corpus and Machine Learning Models for Fake News Classification in Sinhala”.

[13] S. Jayasooriya, “Sinhala Unicode Hate Speech.” https://www.kaggle.com/datasets/sahanjayasuriya/sinhala-unicode-hate-speech.