HATE SPEECH DETECTION ON SINHALA SOCIAL MEDIA TEXT USING LSTM AND FASTTEXT

W.A.S.N. Perera[1*], I. Perera[1], S. Ahangama[2]

[1]Department of Computer Science and Engineering, [2]Department of Information Technology, University of Moratuwa, Sri Lanka

INTRODUCTION

Human beings can express and share information on what they think, and others have the freedom to agree/disagree or express their opinions in a peaceful disposition ensuring the freedom of expression which is essential to an individual's liberty. However, freedom of expression is being grossly misused these days. Therefore, people need to take responsibility and enlighten themselves on the consequences retrospectively.

There are several ethnoreligious cases reported in Sri Lanka: the Easter Sunday attack in 2019, Ampara and Digana in 2018, Gintota in 2017, violence in Aluthgama in 2014, Mawanella riots in 2001. Those cases evidenced the consequences such as escalated to the point of property damage, grievous injury, and deaths. The United Nations High Commissioner for Human Rights has identified that hate speech has worsened the societal and racial tensions (OHCHR | Joint Open Letter on Concerns about the Global Increase in Hate Speech, 2019). Thus, it is scientifically evidence-based that there is an increase in ethnoreligious violence in recent years.

Online hate speech has a ramp up violence around the world. This is encouraged with the digital development of social media platforms and cyberspace due to the protection of anonymity. Social media platforms allow for their users to communicate with their native languages such as Sinhala, Tamil, German, and French (Facebook In-Stream Ads, n.d.). In Sri Lanka, online hate speech is more prevalent on social media platforms. Further, this is prominent at the ground level, with racist, sexist remarks at election rallies and canvassing campaigns (Hate Speech, 2020).

The explosion in popularity of social media encouraged online social networking by remarking the fastest assimilation in communication technology and revolutionized the way of communication. The proliferation of social media encourages people to connect socially online and express their ideas with a wider community. However, this has resulted in exposure to highly undesirable phenomena such as exposure to hate content.

The population of Sri Lanka is 21.46 million in January 2021 and 10.90 million uses the internet while 7.90 million of them are social media users (Digital in Sri Lanka, 2021) which is 36.8% of the total population (Social Media Stats Sri Lanka, 2021). Among the usage of social media platforms, the most famous platform is Facebook (FB) which is 57.44%, followed by Twitter 33.99% and YouTube 4.17% respectively. Therefore, from the statistics, it is evident that a vast majority of Sri Lankans use FB as their social media platform for communication. Therefore, it is rational to choose FB for the selection of our corpus for the study of hate speech detection.

It is important to automatise early detection of hate speech to minimize the spread over social media. However, detecting hate speech is challenging with morphologically rich languages such as Sinhala. However, it has become gruelling with the automatic detection of Sinhala hate content. Furthermore, the social media platform owners do not have good linguistic knowledge of the Sinhala language. Therefore, the identification of hate content has become an arduous task which is time-consuming to take actions such as removing content from their platforms. The hate content on social media causes toxic behaviours: harassing behaviour online, harassment directed towards others, purposeful embarrassment, physical threats, and sexual harassment. Therefore, it necessitates the early and urgent detection of hate content in those platforms to suspend or delete such messages to reduce the spread.

There is a paucity of research that has been conducted for Sinhala hate speech detection on social media platforms. A keyword-based approach was used many years back for hate speech detection (MacAvaney et al., 2019). An ontology or dictionary-based approach is used to identify hateful words such as hatebase, which is an online repository of structured, multilingual, usage-based hate speech. However, this approach has several limitations such as the inability to identify hate contents that do not have hate words and slang. The Support Vector Machine was used to classify racist and non-racist comments obtained from FB and however the precision of the results in this model dropped at some point (Dias et al., 2018). According to the review by Dikwatta & Fernando, 2019, the creation of an unbiased set of data is essential and the results depend on it. Further, they have identified that the deep learning methods have outperformed when classifying text.

In this research, we propose an approach of automatic hate speech detection which outperforms the state-of-art Deep Learning techniques. We have used FastText which captures the linguistic context of words or sentences and a special type of Recurrent Neural Networks (RNN) using the deep learning-based model, Long-Short Term Memory (LSTM) for the classification of text to hate and non-hate. The model accuracy was 80% for hate and non-hate speech detection for the 8252 FB corpus which is evenly distributed.

The rest of the paper is organized as follows. The methodology section briefly discusses the used data set and the developed model to detect hate speech. The next sections provide a series of analysis results. We conclude the paper by discussing research contributions along with suggestions for potential future work.

METHODOLOGY

This study focuses on identifying hate content in the social media platform. We have used the FB corpus extracted from the Sri Lankan community. Then the data set was manually annotated into two labels as hate and non-hate with a crowdsourcing platform using a supervised classification method. The proposed model was developed using a text analytics model with a deep learning technique, LSTM which is a specialized neural network architecture. Further, FastText which is an open-source, free, lightweight library has been used with LSTM to learn text representations and text classifiers created by FB’s AI Research (FAIR) laboratory. The model was trained with the train data and evaluated with the test data set. Figure 1 depicts the high-level workflow of the proposed model and describes all the steps which have been implemented in Keras.

model.

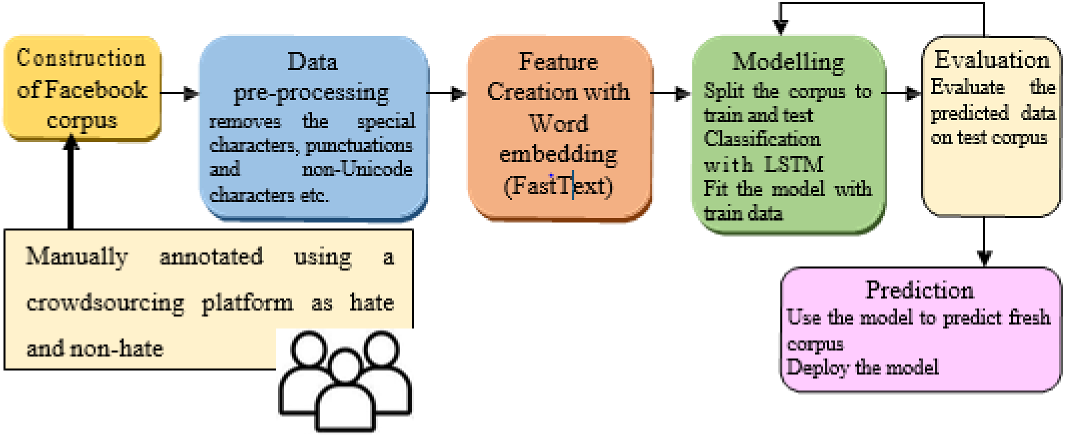
Construction of Facebook corpus
Data pre-processing
removes the special characters, punctuations and non-Unicode characters etc.
Feature Creation with Word embedding (FastText)
Modelling
Split the corpus to train and test
Classification with LSTM
Fit the model with train data
Evaluation
Evaluate the predicted data on test corpus
Manually annotated using a crowdsourcing platform as hate and non-hate
Prediction
Use the model to predict fresh corpus
Deploy the model

Figure 1. The workflow of the model to predict Sinhala corpus as hate and non-hate

Construction of corpus

We have built a corpus of 8252 posts retrieved from the FB public groups and pages chosen with possibly containing hate content. The post written in the Sinhala language was considered and cleaned by removing unwanted data, duplicate values, and invalid data. The model is developed for the automatic detection of hate or non-hate speech as a supervised classification task. Therefore, it is paramount to label the corpora. The data set was annotated as hate and non-hate using a crowdsourcing platform with three annotators ensured by continuous mentoring to get ready for the supervised learning methods to be incorporated for the text analysis. Then posts were saved in a CSV file with UTF-8 formatting.

Data pre-processing

The data set was passed through few preprocessing steps to clean and prepare for the training phase. We removed special characters, punctuations, and non-Unicode characters. TensorFlow has been used for text input preprocessing. Subsequently, each text was transformed into a sequence of integers. In the vectorized representation of data, each sequence is required to have the same length.

The corpus was divided into two sets as training and testing to train and test the model, respectively. The split of the dataset was done by allocating 70% of data for training and the remaining 30% for testing. The performance of the proposed model was evaluated using the test set. In our experiment, we used two classes and the labels used in the data set were hate and non-hate.

Word embedding

The mapping of words from natural language to a real vector space is a prominent task in Natural Language Processing (NLP). According to a study conducted for the word embedding evaluation for Sinhala (Lakmal et al., 2020), it has been identified that the FastText word embeddings with 300 dimensions performed well than the Word2Vec. In this study, the FastText library has been used for word embeddings with its training speed, accuracy and was trained to learn high-quality representations of words by considering the morphology. It consists of multi-lingual word vectors with pre-trained models for 157 different languages including Sinhala by taking advantage of the languages' morphological structure (FastText, n.d.). We have used a pre-trained model for the Sinhala language with 300 dimensions using cc.si.300.vec.gz. Then the created words from our corpus are mapped with the trained corpus in the FastText library.

Train the model

The model consists of neural network layers. A sequential model of TensorFlow with a plain stack of layers has been used. Each layer has exactly one input tensor and one output tensor. The used three layers for the model are Embedding, LSTM, and Dense layers. The Embedding layer turns the positive integers into dense vectors of fixed. For the language modelling, it is needed to predict the next value in sequential data and require the previous behaviours and RNN used for that. However, RNN does not have enough capability to capture long-term dependencies in sequences. LSTM was introduced to resolve that which is a specialized version of RNN used as the second layer. The third layer is the dense layer which is a densely connected neural network layer. Here we have used the Sigmoid activation function which always returns a value between 0 and 1. The model employed a sigmoid activation function and a binary cross-entropy loss function. This was trained for 10 epochs with the corpus.

Prediction with the model

We have used a separate set of fresh corpora to predict the above-developed model. The model is capable enough to predict a given Sinhala text as hate or non-hate.

RESULTS AND DISCUSSION

The performance of the model was measured with accuracy, precision, recall, and F1- score. The accuracy is the percentage of posts that are correctly classified under each class. In this study, we considered the hate posts as the positive class. The precision measures the closeness where it is close to the true value. The recall is a measurement where it indicates the relevant data in the dataset classified by the model. F1 score reaches its best value at 1 and worst score at 0. When checking the above performance measurements on training data sets, we achieved accuracy, precision, recall, and F1-score rates of 80%, 81%, 80%, and 80% respectively. The results obtained for the confusion matrix are given in Table 2.

Table 2. Confusion Matrix

	Hate	Non-Hate
Hate	1048	176
Non-Hate	307	945

CONCLUSIONS/RECOMMENDATIONS

In this study, we examined the automatic detection of hate and non-hate speech in the Sinhala FB corpus. For the binary classification of hate speech using embedding representation of words using FastText and a special type of RNN, which is LSTM. Then the model was evaluated with the performance measurements of accuracy, precision, recall, and F1- score. The accuracy of the model is 80% and it is accurately predicting the given fresh corpus as hate/non-hate. Therefore, the proposed model has performed well for the Sinhala hate speech detection with the FB corpus.

Some of the FB texts may have been misclassified as shown in the confusion matrix due to some reasons such as the absence of hate words, the importance of considering the context, the position or the status of the speaker in society. Finally, there are many features we have not considered when identifying the hate content such as caption of the images and image itself. This study is a part of a larger research and future work will be on a comprehensive analysis of hate speech detection and ranking the hate speech propagators in the social media platforms to minimize the consequences that occurred from hate speech online. As a future direction, it is important to classify hate speech such as racism, sexism and minimize their use in the social media platforms in the Sri Lankan context.

REFERENCES

Dias, D. S., Welikala, M. D., & Dias, N. G. J. (2018). Identifying Racist Social Media Comments in Sinhala Language Using Text Analytics Models with Machine Learning. 2018 18th

International Conference on Advances in ICT for Emerging Regions (ICTer), 1–6. https://doi.org/10.1109/ICTER.2018.8615492

Digital in Sri Lanka: All the Statistics You Need in 2021. (2021). DataReportal – Global Digital Insights. Retrieved May 23, 2021, from https://datareportal.com/reports/digital-2021-sri-lanka

Dikwatta, U., & Fernando, T. G. I. (2019). Violence Detection in Social Media-Review, *Vidyodaya Journal of Science*, vol 22, (pp 7-16)

Facebook In-Stream Ads: Country and Language Availability. (n.d.). Facebook Business Help Center.

Retrieved May 23, 2021, from https://www.facebook.com/business/help/267128784014981 Hate speech, divisive language on the rise on Sri Lanka’s campaign trail: CMEV. (2020, July 28).

EconomyNext. https://economynext.com/hate-speech-divisive-language-on-the-rise-on-sri-lankas-campaign-trail-cmev-72452

Lakmal, D., Ranathunga, S., Peramuna, S., & Herath, I. (2020). Word Embedding Evaluation for Sinhala, *12th Conference on Language Resources and Evaluation (LREC 2020),* (pp. 1874-1881). Marseille

MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O. (2019) *Hate speech detection: Challenges and solutions.* PLoS ONE 14 (8), China

OHCHR | Joint open letter on concerns about the global increase in hate speech. (2019). Retrieved May 22,2021,from https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx? NewsID=25036&LangID=E

Social Media Stats Sri Lanka. (2021). StatCounter Global Stats. Retrieved May 23, 2021, from https://gs.statcounter.com/social-media-stats/all/sri-lanka

Word vectors for 157 languages · fastText. (n.d.). Retrieved May 22, 2021, from https://fasttext.cc/index.html