Detection of online hate speech in Sinhala text using machine and deep learning algorithms: A comparative study

F. H. A. Shibly[a]*, U. Sharma[b], H. M. M. Naleer[c]

[a]Department of Arabic Language, South Eastern University of Sri Lanka.
[b]Department of Computer Applications, Assam Don Bosco University, India.
[c]Department of Computer Science, South Eastern University of Sri Lanka.

([a]shiblyfh@seu.ac.lk, [b]uzzal.sharma@dbuniversity.ac.in, [c]drnaleer@seu.ac.lk)

Keywords: Hate detection, machine learning, Sinhala text.

Introduction

The number of Internet users has grown dramatically in recent years, and as a result, communication between two or more people has become a relatively easy thing via the use of the Internet. In contrast to the early phases of social media, the most of them today support a wide range of languages spoken all over the globe thanks to the use of Unicode encoding [1]. As a result, many individuals choose to interact on social media platforms in their own language rather than utilizing international languages such as English or Spanish. As a result of the incorporation of languages such as Sinhala, Tamil, and Hindi onto the Internet, individuals who do not have a strong command of English are more likely to participate in social media and blogs. Users may express themselves freely and anonymously on a wide range of online communication platforms, including social media, which are widely available. However, generating and propagating hatred towards another group is a violation of one's right to free expression [2], which should be respected at all times. There have been numerous research efforts on ways to control such hate speech. The study of how to control hate speech on social media, especially with the help of areas of artificial intelligence such as machine learning and deep learning, is taking place from various angles. However, there has been a lot of research on social media about English language hate speech. As the rate at which individuals share ideas in their mother tongues on social media other than English is high, mechanisms to control hate speech shared in different languages are needed. Based on them, the study of hate speech shared in Sinhala, the most widely spoken language in Sri Lanka, is seen at an early stage. So, this study sets out to effectively detect Sinhala language online hate speeches through machine learning algorithms. The main objective of this research is to analyze how effectively the machine learning algorithm detect hate speech in the Sinhala language. Sub objectives are to find out how each machine learning algorithm reacts to hate speech in terms of accuracy, precision, recall and f1 score and to conclude which algorithm works best in detecting hate speech in Sinhala.

Methodology

The main objective of this research is to compare the performances of machine learning algorithms in detecting Sinhala text hate speech detection process. Therefore, we have taken into consideration a technique that is focused on importing, preprocessing, training, testing, evaluating and comparing the performances of selected algorithms (Figure 1). In Importing phase, we imported the dataset in Google Colab environment and preprocessed the dataset by using remove missing values, numbers and punctuations. We have used the dataset which was uploaded by Sahan Jayasuriya in Kaggle. This dataset consists 6345 facebook comments based on Sinhala Unicode [4]. At the training and testing phases [3], DT, LR, KNN, AB, MNB, GB, RF, SVC and LSTM were trained and tested. There are many methods we can use for this including state of art method. As an initial study, the researchers tried to analyze the performance of ML algorithms in detecting hate speech in Sinhala language. In future, researchers will apply state of art methods further. The data set was a balanced data set since each output class had similar number of input samples. First the text features were preprocessed so as to get a clean data set. Data preprocessing includes Sinhala stop words, symbols, Punctuations, URL and retweet sign removal. For the machine learning models, input text features were tokenized using TfidfVectorizer, where text features were tokenized using texts_to_sequences method and then padded the sentence sequence with pad sequences method. When it comes to data set, it was split into training and validation set with probabilities 0.8 and 0.2 for the ML models. Dataset separated into 3 sets as training, validation and test set with the corresponding probability 0.6, 0.2 and 0.2 for the LSTM model. Validation set was used to tune the LSTM model. For the LSTM model optimum hyper parameters were 1e-3 for learning rate and 1e-8 for the épsilon. Adam optimizar was used to optimize the model were trained and tested. As the evaluation phase, we analyzed F1 score, precision, recall and accuracy of each algorithm over to the selected dataset. Finally, we compared the performances of machine learning algorithms and find out the best algorithm in detecting Sinhala text online hate speech with this research context.

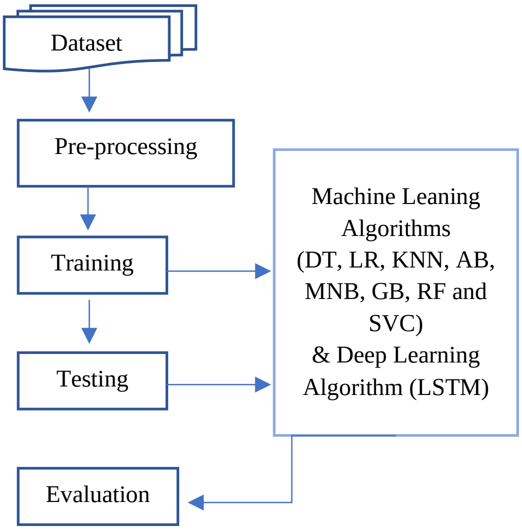
Dataset
Pre-processing
Training
Testing
Evaluation
Machine Leaning Algorithms (DT, LR, KNN, AB, MNB, GB, RF and SVC) & Deep Learning Algorithm (LSTM)

Figure 1. Design of the research.

Experimental setup

In the disciplines of data mining and data retrieval, evaluating the accuracy of machine learning classifiers is one of the most important phases. Error rate and F-measure are widely used to determine the accuracy of a classifier's ability to locate the proper category or class of unknown cases. The error rate is the instances of the test set that were erroneously categorized. We'll call this set of data "X" and let "m" represent how many occurrences were misclassified by a classification model C. You can calculate the accuracy of C in selecting the correct classes of X instances using the following formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \qquad (1)$$

The error rate approach ignores the cost of inaccurate predictions in machine learning. For the most part, F-measure is used to solve this problem. To determine the value of F-measure, two basic metrics are used: precision and recall. Imagine that some of the data in the test set belong to a certain class or category S. It assigns a category label to each test data. There will be four kinds of forecasts for the test set S: Percentage of accurately forecast data for category S is known as precision. Percentage of correctly forecast real data for category S is known as recall. It is possible to calculate the F-measure based on precision and recall (2-4).

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (3)$$

$$\text{F1} = 2\ X\ \frac{Recall*Precision}{Recall+Precision} \qquad (4)$$

Experiment process has multiple steps such as data preprocessing, feature tokenizing, train test split, model training and prediction evaluation. Data preprocessing and cleaning directly affect accuracy rate of the model, since uncleaned data add noise to the model. Tf-Idf Vectorizer was used to tokenize text features for ML models. Tf- idf Vectorizer first tokenize text, learn vocabulary and calculate the inverse document frequency. This tokenization process allows us to encode new document or sentence using learned vocabulary.

Using train set, all the models were trained and validation set was used to tune the parameters in LSTM model. When it comes to hyper parameter optimization in LSLM model different learning rates, epsilon values were used for Adam optimizer so as to increase the validation accuracy. Optimum learning rate is 1e-3 and epsilon are 1e-08 for the dataset. In order to avoid overfitting to data, dropout layers were used in LSTM model. Binary cross entropy was used to measure the loss of LSTM model, since this is a binary classification. After the training process, models were evaluated with accuracy, precision, recall and f1-score with the test dataset prediction results.

Results and Discussions

This part describes the overall results of eight ML algorithm tests. Tables I shows the precision, recall, F-measure, and accuracy of all eight algorithms to detect hate and offensive language, respectively. The results of different feature representation and classification algorithms used under experimental settings are shown in the following table. (NHS stands for No Hate Speech and HS stands for Hate Speech).

Table 1. Comparison of the performances of ML algorithms.

ML Algorithms		**Precision**	**Recall**	**F – Measure**	**Accuracy (%)**
DT	NHS	0.69	0.74	0.71	73 %
	HS	0.77	0.72	0.75	
LR	NHS	0.78	0.74	0.76	79%
	HS	0.80	0.83	0.81	
KNN	NHS	0.52	0.75	0.62	57%
	HS	0.67	0.43	0.52	
AB	NHS	0.69	0.79	0.74	74%
	HS	0.80	0.70	0.75	
MNB	NHS	0.90	0.58	0.71	78%
	HS	0.73	0.95	0.82	
GB	NHS	0.72	0.77	0.74	76%
	HS	0.80	0.75	0.77	
RF	NHS	0.77	0.78	0.78	80%
	HS	0.82	0.81	0.81	
SVC	NHS	0.82	0.76	0.79	81%
	HS	0.81	0.86	0.83	
LSTM	NHS	0.78	0.88	0.79	84%
	HS	0.89	0.80	**0.83**	

When comparing with all performances, KNN performs lowest performance in terms of all parameters. DT, AB, GB, and LSTM have obtained 73%, 74% ,76%, and 84% in accuracy respectively. These three algorithms have got relatively similar and good performance scores. LR has got good accuracy as well as good precision, recall and F1 scores. RF is also has obtained highest precision and second-best accuracy. MNB has got highest recall and have obtained good scores in accuracy, precision and F1. It was found that MNB is the best algorithm in classifying Romanized Sinhala [5]. When it comes to LSTM model it could achieve higher accuracy, recall and precision for each target classes. So, LSTM has outperformed other eight algorithms in accuracy and F1 measure.

Conclusion

This study was carried out based on the detection of Sinhala text online hate speech. The main focus of this research was on eight machine learning algorithms for identifying hate speech in facebook comments, which were tested. The analysis and findings emphasized that the LSTM algorithm outperformed other eight algorithms in accuracy and F1 measure. KNN has obtained the poorest performances. RF, MNB, LR and SVC have also obtained good scores and these algorithms can be used to detect Sinhala text hate speech in social media. It is important to note that the findings of this research study will be used as a baseline to evaluate future investigations inside different automated text classification algorithms for automatic hate speech detection, thus the findings of this research study are important.

References

[1] Sandaruwan, H. M. S. T., Lorensuhewa, S. A. S., and Kalyani, M. A. L., Sinhala hate speech detection in social media using text mining and machine learning. In 2019 19th International Conference on Advances in ICT for Emerging Regions, 2019. 250: pp. 1-8. IEEE.

[2] MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., and Frieder, O., Hate speech detection: Challenges and solutions. PloS one, 2019. 14(8):.

[3] Hsu, B. M., Comparison of supervised classification models on textual data. Mathematics, 2020. 8(5): p. 851.

[4] Sinhala Unicode Hate Speech. (2020, April 12). [Dataset]. Sahan Jayasuriya. https://www.kaggle.com/sahanjayasuriya/sinhala-unicode-hate-speech.

[5] Hettiarachchi, N., Weerasinghe, R., and Pushpanda, R., Detecting hate speech in social media articles in Romanized Sinhala. In 2020 20th International Conference on Advances in ICT for Emerging Regions, (2020, November). pp. 250-255. IEEE.