Unsupervised Sinhala Cyberbullying Categorization

B.G.M Chandrasena

（Reg. No.: MS19810874）

M.Sc. in IT Specialized Cyber Security

Supervisor: Dr. Lakmal Rupasinghe

December 2021

Faculty of Graduate Studies and Research Sri Lanka Institute of Information Technology

DECLARATION

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, to be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Signature of Candidate: . . . . . Date: ..../..../....

Name of Candidate: B.G.M Chandrasena

KEYWORDS

Cyberbullying, Hate Speech, Machine Learning, NLP (Natural Language Processing), Supervised Learning, Unsupervised Learning , Artificial Neural Network

ABSTRACT

The objective of unsupervised machine learning is to categorize the social media comments into a given number of pre-learned categories. The earlier studies of this domain have used many the dataset for supervised learning & introduced a large number of techniques, methodologies. A major challenge there was training labels. Although words with training comments are easy to find, separating them manually is not an easy task.

Through this research, we hope to find a solution to this using unsupervised machine learning techniques. the proposed technique divides the comments into words and removed special characters, emojis, and links from the comments & categorized each comment using a keyword list of each category and similarity findings. And then this was used to categorize comments for training. The implemented method shows the same performance, by Comparison with other supervised machine learning techniques for cyberbullying.

Therefore, this mechanism can be used in any other places where low-cost cyberbullying identification is needed. This also can be used to create train comments.

ACKNOWLEDGEMENT

To convey my heartfelt thanks to Dr. Lakmal Rupasinghe, my primary research supervisor, for his direction, input, advice, and support during the study process. Also I would like to convey my heartfelt thanks to Dr. Anuradha Jakody too.

Many thanks also to Dr. Darshana Kasthurirathna,, Mr. Amila Senarathne, and Mr. Kavinga Yapa Abeywardena. for their assistance and support.

I would also want to express my gratitude to Mr. Suneth Koggalahewa, who provided me with his unwavering support in order to conquer this obstacle and assistance in writing my thesis.

I would also want to express my gratitude to all of my M.Sc. colleagues, for whom I am eternally thankful for all of the memorable experiences we have had together

TABLE OF CONTENT

DECLARATION ..... **ii**

KEYWORDS ..... **iii**

ABSTRACT **1**

ACKNOWLEDGEMENT **2**

TABLE OF CONTENT **3**

LIST OF FIGURES **6**

LIST OF TABLES **7**

CHAPTER 1: INTRODUCTION **8**

1 Overview 8

1.2 Background and Motivation 11

1.3 Problem Definition 12

1.4 Research Questions 12

1.5 Aim 12

1.6 Objectives 13

1.7 Structure of this thesis 13

Chapter 2: Analysis **14**

2.1 Introduction 14

2.2 Literature Review 14

2.3 Features 20

2.3.1 General Features 20

2.4.1 Supervised Machine Learning 27

2.4.2 Unsupervised Machine Learning 27

2.4.3 Machine Learning Algorithm Tree 28

2.5 Data Managing 28

2.5.1 Numpy 28

2.5.2 Pandas 29

2.5.3 Scikit-Learn 30

2.5.4 Imbalanced Learn 31

2.5.5 GSITK 31

2.6 NLP (Natural Language Processing) 32
2.6.1 Natural Language Tool Kit (NTLK) 33
2.6.2 Genism 33
2.6.3 TextBlob 33
2.6.4 Hate Base Application Programming Interface 34
2.7 Mchine Learning Fundamentals 35
2.7.1 Logistic Regression 36
2.7.2 Support Vector Machine 37
2.7.3 Random Forest 38
2.7.4 Artificial Neural Network 39
2.8 Natural Language Processing Fundamentals 43
2.8.1 Bag of Words (BOW) 44
2.8.2 Term Frequency (TF)-Inverse Document Frequency (IDF) 45
2.8.3 LDA 46
2.8.4 Word Embeddings 47
2.8.5 SIMON (Similarity Based Sentiment Projection) 48
..... 49
CHAPTER 3 50
3.1 Introduction 50
3.2 System Approach 51
3.2.1 Preprocessing 52
3.2.2 Training Word List Creation 55
Category 56
Keywords 56
3.2.3 Feature Selection and Classifier 60
3.3 Summary 62
CHAPTER 4 63
4.1 Introduction 63
4.2 System Architecture 63
4.3 Input and Output 63
4.4 Basic Design 64

4.4.1 Dataset Collection 65

4.4.2 Preprocessing 68

CHAPTER 5 74

5.1 Overview 74

5.2 Datasets 74

CHAPTER 6 79

REFERENCES 80

LIST OF FIGURES

Figure 1 Number of research paper per year 14
Figure 2 Sizes of the datasets utilized in the publications 18
Figure 3 Machine Learning Algorithm Tree 28
Figure 4 HatebaseCommunity 34
Figure 5 API of Hatebase 35
Figure 6 System- Overfitted vs Regularized 36
Figure 7 Logistic Function 37
Figure 8 Hyperplane of Maximum Margin 38
Figure 9 Kernel Trick 38
Figure 10 Effect of Modifying the C Parameter 38
Figure 11 Tree Learned- Titanic Data 39
Figure 12 Human Neuron 40
Figure 13 Artificial Neuron 40
Figure 14 Architecture of Perceptron 42
Figure 15 Perceptron of Multi-Layered 42
Figure 16 Comparison of Activation Functions 43
Figure 17 Patterns of Word Embeddings 48
Figure 18 Word Embedding Model- Cosine Similarity 49
Figure 19 Detailed System Approch 51
Figure 20 Preprocessing 52
Figure 21 Training Word List Creation 55
Figure 22 Cyclic word & sentence similarity calculation 58
Figure 23 Feature Selection & Classifier Module 60
Figure 24 High Level View of Proposed Module 64
Figure 25 Export Comment Website's Interface 65
Figure 26 Google Cloud Platform-New Project 66
Figure 27 Google Cloud Platform- API & Services 66
Figure 28 Google Cloud Platform-YouTube Data API V3 67
Figure 29 YouTube Data API Key 67
Figure 30 Clean Comments-Puntuation Removed 70
Figure 31 Clean Comments- Numbers Removed 70
Figure 32 Clean Comments Emojis Removed 71
Figure 33 Manually Labled comments 75
Figure 34 Automatically System Generated Comments 76

LIST OF TABLES

Table 1 Standard Cyberbullying Definitions 9
Table 2 Types of cyberbullying & examples 10
Table 3 The publications' keywords 15
Table 4 Cyberbullying types analyzed in the researches 18
Table 5 Used algorithms for researches 19
Table 6 Representation of BOW 44
Table 7 Sinhala POS Tags 54
Table 8 Sample Keyword Categories 56
Table 9 Sinhala Stop Words 72
Table 10 Example Sinhala POS Tags 72
Table 11 Manually Labled Comments 74
Table 12 Manually Labled Comments 75
Table 13 Automatically Labeled commen 75
Table 14 Confusion Matrix for Unsupervised Model 76