

**ANALYTIC OF IMAGE POSTS ON SOCIAL MEDIA
FOR HATE SPEECH**

Kalhari Shanthika Ahangama Walawage

198106D

Master of Philosophy (MPhil.)

Department of Information Technology

Faculty of Information Technology

University of Moratuwa

Sri Lanka

June 2025

ANALYTIC OF IMAGE POSTS ON SOCIAL MEDIA FOR HATE SPEECH

Kalhari Shanthika Ahangama Walawage

198106D

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Philosophy (MPhil.)

Department of Information Technology
Faculty of Information Technology

University of Moratuwa
Sri Lanka

June 2025

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Name of Student: K.S.A. Walawage

Signature:

Date: 29/07/2025

The above candidate has carried out research for the MPhil thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. L. Ranathunga

Signature of the Supervisor:

Date: 29/07/2025

ACKNOWLEDGEMENTS

This research was supported by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education funded by the World Bank.

I would like to express my sincere gratitude to my supervisor Dr. Lochandaka Ranathunga, for his support, patience, motivation, and immense knowledge. His valuable guidance helped me to succeed in this entire MPhil research and to write this thesis.

I would like to thank my examiners, Prof. Indika Perera and Dr. Amalraj, for their advice, motivation and guidance.

I'm grateful to my parents, Mr. A.W.B. Chandrasena and Ms. B.D.G. Liyanage and family for their continuous support and understanding during my research and thesis writing.

ABSTRACT

Social media platforms have been developing rapidly for the last thirteen years. Billions of people communicate their ideas, information and other expressions with each other through social media. Recently most of the fake news and hate speech have spread by image posts and videos in Facebook, YouTube and Twitter platforms within Sri Lankan community. There is a vital need for obtaining the meaning of the image posts automatically. This research has been supported to the above solution and done to automatically separation and identification of Sinhala and Singlish text in image posts and video thumbnails in social media platforms. This research has three research module components. The first component is the main research component. It is the extraction of Sinhala and English text in social media images. Image posts and video thumbnails were acquired from Facebook and YouTube social media platforms which are on public. Three main algorithms and another three supporting algorithms have been derived to complete this component. They are text detection, text glow removing, text line segmentation, touching character segmentation, Sinhala English character separation and Sinhala English character recognition. The study has achieved the text line segmentation accuracy of 90% for 1,669 text lines and the overall touching character segmentation accuracy of 98% for 58,403 characters and the overall script separation accuracy of 93% for 22,383 script images. A new feature vector was formed in this study and 11,088 Sinhala script images were trained in a sequential model neural network. The accuracy of printed Unicode Sinhala character recognition is higher than the accuracy of printed non-Unicode and handwritten Sinhala character recognition. In the second component, a simple algorithm has been derived to identify the status of social issues in images which is connector module for the main research group work. In the third component, it is tested with three methods to identify key attributes of images to detect redistribution of images within social media and selected the Brute-Force Matcher with ORB descriptors as the highest accuracy obtained by this method.

Keywords:

Social Media Images, Text Detection, Segmentation, Sinhala English Separation, Character Recognition

TABLE OF CONTENTS

DECLARATION	i
ACKNOWLEDGEMENTS	ii
ABSTRACT.....	iii
TABLE OF CONTENTS.....	iv
LIST OF FIGURES	vii
LIST OF TABLES	xi
List of Appendices	xiv
CHAPTER 1	1
INTRODUCTION	1
1.1 Background	1
1.2 Research Problem.....	2
1.3 Aim and Objectives	2
1.3.1 Research Aim.....	2
1.3.2 Objectives	3
1.4 Solution	3
1.5 Chapter Summary.....	4
CHAPTER 2	5
REVIEW OF LITERATURE	5
2.1 Text Extraction.....	5
2.1.1 Critical Engagement with Contemporary Peer-Reviewed Works	6
2.2 Segmentation	8
2.2.1 Segmentation by Projection Profiles	8
2.2.2. Segmentation by Contour Analysis Method.....	8
2.2.3. Segmentation by Water Reservoir-based Method	8
2.2.4. Segmentation by Cluster Detection Technique	8
2.2.5. Segmentation by Self Organizing Feature Map.....	9
2.2.6 Critical Engagement with Contemporary Peer-Reviewed Works	9
2.3 Sinhala English Script Separation	10
2.3.1 Critical Engagement with Contemporary Peer-Reviewed Works	11
2.4 Character Recognition.....	12
2.4.1 Recognition by Feature Extraction	12
2.4.2 Recognition by CNN/LSTM	14

2.4.3 Recognition using Tesseract OCR Engine	16
2.5 Image Matching.....	18
2.5.1 Critical Engagement with Contemporary Peer-Reviewed Works	19
2.6 Chapter Summary	19
CHAPTER 3	20
RESEARCH METHODOLOGY.....	20
3.1 Extract Sinhala and Singlish Text in image posts and Video thumbnails.....	20
3.1.1 Data Collection	20
3.1.2 Text Detection	21
3.1.3 Preprocessing	23
3.1.4 Segmentation	25
3.1.5 Sinhala English Character Separation	37
3.1.6 Character Recognition	43
3.2 Identify key attributes of images to detect redistribution of images within social media	55
3.2.1 Brute-Force Matcher with ORB descriptors and Ratio Test	56
3.2.2 Brute-Force Matcher with SIFT descriptors and Ratio Test	58
3.2.3 FLANN based Matcher with SIFT descriptors and Ratio Test	60
3.3 Chapter Summary	63
CHAPTER 4	64
RESULTS AND DISCUSSION	64
4.1 Results of Text Detection Algorithm	64
4.2 Results of Glow Removing Algorithm.....	66
4.3 Results of Text Line Segmentation Algorithm	67
4.4 Results of Word Segmentation Algorithm	68
4.5 Results of Isolated and Overlapping Character Segmentation Algorithm	68
4.6 Results of Touching Character Segmentation Algorithm	70
4.6.1 Printed Characters.....	71
4.6.2 Handwritten Characters	72
4.6.3 Overall Characters	73
4.7 Results of Sinhala and English Script Separation Algorithm	75
4.7.1 Printed Characters.....	75
4.7.2 Handwritten Characters	76
4.7.3 Overall Characters	77
4.8 Results of Sinhala Character Recognition Algorithm.....	79

4.8.1 Character Dataset for training, validation and testing	79
4.8.2 Training and Testing of Novel Features	80
4.9 Results of English Character Recognition Algorithm.....	80
4.9.1 Character Dataset for training, validation and testing	80
4.9.2 Training and testing	82
4.10 Results of BF Matcher with ORB descriptors and Ratio Test Algorithm.....	84
4.11 Results of BF Matcher with SIFT descriptors and Ratio Test Algorithm.....	86
4.12 Results of FLANN based Matcher with SIFT descriptors and Ratio Test Algorithm	87
4.13 Chapter Summary	88
CHAPTER 5	89
CONCLUSION AND FUTURE WORKS	89
5.1 Conclusion.....	89
5.2 Future Works.....	90
5.3 Chapter Summary.....	90
REFERENCES	92
APPENDICES	100

LIST OF FIGURES

Figure 2.1: Main research areas related to this research.	5
Figure 3.1: Flow diagram to extract Sinhala and Singlish Text in image posts and Video thumbnails.	20
Figure 3.2: Outputs of step-by-step process of text detection. (a) acquired original image, (b) preprocessed image (c) gradient morphological operation applied image, (d) Threshold image, (e) closing morphological operations applied image, (f) final output of detected text content.	23
Figure 3.3: A light background with a dark foreground image (a) before binarization (b) after binarization.	24
Figure 3.4: A dark background with a light foreground image (a) before binarization (b) after binarization (c) after the opposite background conversion.	24
Figure 3.5: Outputs of each step of text glow removing. (a) An image with text glow. (b) Binary image. (c) Inverted binary image. (d) Connected background pixels. (e) Binary image after identifying the connected pixels. (f) Final output image.	25
Figure 3.6: Three types of text lines (a) isolated (b) overlapping and (c) touching.	26
Figure 3.7: Text lines (a) with the generated horizontal projection profile histogram, (b) last set of lines after the segmentation.	26
Figure 3.8: An isolated text line after the segmentation.	26
Figure 3.9: Horizontal projection profile histogram (a) valley points, (b) height differences of consecutive valley points.	27
Figure 3.10: Valley differences (a) before compressing the differences, (b) after compressing the difference.	28
Figure 3.11: The procedure of word segmentation (a) resized image, (b) filtered image, (c) binarized inverse image, (d) segmented words.	30
Figure 3.12: Character Types (a) isolated characters, (b) overlapping characters, (c) touching characters.	30
Figure 3.13: The four rounds of connected pixel labeling method (a) binary image, (b) first round completed image, (c) second round completed image, (d) fourth round completed image.	31
Figure 3.14: A connected pixel set; those are touching characters.	32

Figure 3.15: An example of foreground pixel ranges algorithm (a) Two horizontal pixel lines of a connected pixel set, (b) 4 foreground pixel ranges in first horizontal line, (c) 3 foreground pixel ranges in second horizontal line.....	32
Figure 3.16: An example (a) touching character image, (b) identified upper line and lower line.	33
Figure 3.17: A part of pixels of a binary character image.	33
Figure 3.18: Identified peak points and upper side flat areas on the histogram of upper line pixels.	34
Figure 3.19: Identified valley points and lower side flat area on the histogram of lower line pixels.	35
Figure 3.20: The algorithm to find touching and non-touching points.....	36
Figure 3.21: An example of identified water reservoirs (a) upper and lower reservoirs, (b) peak and valley points and upper and lower side flat areas.	36
Figure 3.22: Detected touching points in a word.	36
Figure 3.23: A character with detected contours (a) with image border contour, (b) without image border contour.	37
Figure 3.24: Examples of the first left foreground pixel lines and last right foreground pixel lines	38
Figure 3.25: An algorithm for finding a vertical line.....	39
Figure 3.26: An algorithm for finding a forward diagonal line.	39
Figure 3.27: An algorithm for finding a backward diagonal line.	40
Figure 3.28: Examples of the first top foreground pixel lines and last bottom foreground pixel lines.	40
Figure 3.29: An algorithm for finding a horizontal line.	41
Figure 3.30: The step-by-step process of the novel algorithm.....	41
Figure 3.31: Characters with four contours (a) first left pixel line is vertical, (b) first left pixel line is forward diagonal.	42
Figure 3.32: The step-by-step procedure of third step of the novel algorithm.	43
Figure 3.33: Characters with two or three contours (a) both left side first pixel line and right side last pixel line are vertical, (b) right side last pixel line is vertical, (c) left side first pixel line and right side last pixel line are forward diagonal and backward diagonal respectively, (d) left side first pixel line and right side last pixel line are backward diagonal and forward diagonal respectively, (e) middle vertical line with top first pixel line is horizontal, (f) middle vertical line with character top reservoir is false, (g) left	

and right reservoirs with nearly equal widths, (h) right reservoir with reservoir width is nearly equal to character height, (i) top reservoir with reservoir width is greater than to half of the character width.	44
Figure 3.34: Sixteen directions in (a) horizontal ellipse, (b) vertical ellipse	44
Figure 3.35: Preprocessing Steps (a) thresholding, (b) skeletonization, (c) cropping, (d) resizing, (e) re-thresholding, (f) border setting.	45
Figure 3.36: The algorithm to decide the character shape.	46
Figure 3.37: Character shape (a) horizontal character, (b) vertical character	46
Figure 3.38: Minimum and maximum distance along the north direction from the center point of the character image	47
Figure 3.39: Number of crossing points along each direction from the center point of the character image	47
Figure 3.40: The algorithm to calculate north direction features.....	48
Figure 3.41: ‘X’ shapes (a) horizontal ‘X’ shape, (b) vertical ‘X’ shape	49
Figure 3.42: The algorithm to calculate minimum and maximum distances for northeast direction.....	49
Figure 3.43: The algorithm to calculate crossing points for northeast direction.	51
Figure 3.44: ‘X’ shapes for secondary intercardinal directions (a) north-northwest, north-northeast, south-southwest, south-southeast directions, (b) west-northwest, west-southwest, east-northeast, east-southeast directions	52
Figure 3.45: Sinhala characters: 8 vowels, 25 consonants, 6 modifiers and 38 compound characters.....	53
Figure 3.46: A sequential neural network model to train the 49 features of Sinhala characters.	53
Figure 3.47: CNN model to train the English characters and Arabic numbers.	55
Figure 3.48: The flow diagram of identify key attributes of images to detect redistribution of images within social media by Brute-Force Matcher with ORB descriptors.	56
Figure 3.49: Images with identified key points using ORB (a) a white-list image with identified key points, (b) a social media image with identified key points.	57
Figure 3.50: Matched features of a white-list image and a social media image using BF Matcher with ORB descriptors and Ratio Test	57

Figure 3.51: The flow diagram of identify key attributes of images to detect redistribution of images within social media by Brute-Force Matcher with SIFT descriptors.	58
Figure 3.52: Images with identified key points using SIFT (a) a white-list image with identified key points, (b) a social media image with identified key points.	59
Figure 3.53: Matched features of a white-list image and a social media image using BF Matcher with SIFT descriptors and Ratio Test	60
Figure 3.54: The flow diagram of identify key attributes of images to detect redistribution of images within social media by FLANN based Matcher with SIFT descriptors.	61
Figure 3.55: Matched features of a white-list image and a social media image using FLANN Matcher with SIFT descriptors and Ratio Test.....	63
Figure 4.1: The curves (a) for train and validation accuracies, (b) for train and validation losses	81
Figure A.1: Screenshot of the initial web application interface.....	100
Figure A.2: Screenshot of the activation of Steps 1 and 2 in the web interface.	101
Figure A.3: Screenshot of the activation of Steps 3 in the web interface.	101
Figure A.4: Screenshot of the resulting text file generated through Step 4.	102
Figure B.1: Screenshot of the initial web application interface.....	103
Figure B.2: Screenshot of the activation of Steps 1 and 2 in the web interface.	104
Figure B.3: Screenshot of the activation of Steps 3 in the web interface.	104
Figure B.4: Screenshot of the resulting text file generated through Step 4.	105

LIST OF TABLES

Table 4.1: Results of Text Detection Method	65
Table 4.2: Performance Comparison of the proposed text detection method with other methods	65
Table 4.3: Results of Glow Removing Method	67
Table 4.4: Results of Overlapping and Touching Text Line Segmentation Method .	69
Table 4.5: Performance Comparison of the proposed Text Line Segmentation method with other methods.....	69
Table 4.6: Performance Comparison of Word Segmentation methods	70
Table 4.7: Performance Comparison of Isolated and Overlapping Character Segmentation methods.	70
Table 4.8: Results of Touching Character Identification for Printed Characters.....	71
Table 4.9: Results of Touching Character Identification for Handwritten Characters	72
Table 4.10: Results of Overall Touching Character Identification	74
Table 4.11: Performance Comparison of the proposed Touching Character Segmentation method with other methods.....	74
Table 4.12: Results of Sinhala English Script Separation for Printed Characters	76
Table 4.13: Results of Sinhala English Script Separation for Handwritten Characters	77
Table 4.14: Results of Sinhala English Script Separation	78
Table 4.15: Performance Comparison of the proposed Sinhala English Script Separation method with other methods.....	78
Table 4.16: Confusion Matrix for 8-character classes	81
Table 4.17: The test accuracies of 77 Sinhala scripts	82
Table 4.18: Performance Comparison of the proposed Sinhala Character Recognition method with other methods.....	83
Table 4.19: Performance Comparison of the proposed English Character Recognition method with other methods.....	85
Table 4.20: Results of BF Matcher with ORB descriptors and Ratio Test.....	86
Table 4.21: Results of BF Matcher with SIFT descriptors and Ratio Test.....	87
Table 4.22: Results of FLANN based Matcher with SIFT descriptors and Ratio Test	88

LIST OF ABBREVIATIONS

Abbreviation	Description
BF Matcher	Brute-Force Matcher
BRIEF	Binary Robust Independent Elementary Features
CER	Character Error Rate
CNN	Convolutional Neural Network
DAWG	Directed Acyclic Word Graph
EAST	Efficient and Accurate Scene Text Detector
FAST	Features from Accelerated Segment Test
FITI	From Innovation to Impact
FLANN	Fast Library for Approximate Nearest Neighbors
FN	False Negative
FP	False Positive
IEEE	Institute of Electrical and Electronic Engineers
JPG	Joint Photographic Experts Group
KNN	K Nearest Neighbors
LSTM	Long Short-Term Memory
MDLSTM	Multidimensional Long Short-Term Memory
MLP	Multi-Layer Perceptron
MSER	Maximally Stable Extremal Region
NLP	Natural Language Processing
OCR	Optical Character Recognition
OpenCV	Open Computer Vision Library
ORB	Oriented FAST and Rotated BRIEF
PCA	Principal Component Analysis
SGD	Stochastic Gradient Descent
SIFT	Scale-Invariant Feature Transform
SOM	Self-Organizing Feature Maps

SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UTF-8	Unicode Transformation Format – 8 bits
VPN	Virtual Private Networks
WER	Word Error Rate

List of Appendices

Appendix A: Interfaces of Sinhala-English OCR Web Application	100
Appendix B: Interfaces of Sinhala Handwritten and non-Unicode Character Recognition Web Application.....	103