

LB/TH/38/2025
TH5963

**ASSESSING THE QUALITY OF INFORMATION ON
WIKIPEDIA ARTICLES USING DEEP LEARNING**

P.D.S Thisakya Gunathilaka

248035N

Degree of Master of Science

Department of Information Technology
Faculty of Information Technology

University of Moratuwa
Sri Lanka

August 2025

ASSESSING THE QUALITY OF INFORMATION ON WIKIPEDIA ARTICLES USING DEEP LEARNING

P.D.S Thisakya Gunathilaka

248035N

Dissertation submitted in partial fulfilment of the requirements for the degree
Master of Science in Information Technology

Department of Information Technology

Department of Information Technology
Faculty of Information Technology

University of Moratuwa
Sri Lanka

August 2025

DECLARATION OF THE CANDIDATE AND SUPERVISOR

I declare that this is my own work, and this 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.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:
08/19/2025

08/19/2025

Date:

The above candidate has carried out research for the Masters dissertation under my supervision.

I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr.(Mrs) Supunmali Ahangama

Signature of the Supervisor: _____

Date: 20/08/2025

Name of co-supervisor: Dr. Sapumal Ahangama

Signature of the co-supervisor: _____

Date: 21.08.2025

DEDICATION

I dedicate this dissertation to my beloved parents for their unwavering support and encouragement. I am also deeply grateful to my supervisor, Dr.(Mrs) Supunmali Ahangama, for their invaluable guidance, patience, and belief in my abilities throughout this journey. Their support has been instrumental in shaping the research, and I am forever appreciative.

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to all those who helped me make this research successful with their guidance and advice. Firstly, I greatly thank my Principal Supervisor, Dr. (Mrs) Supunmali Ahangama, and Co-supervisor Dr. Sapumal Ahangama for their enormous guidance, advice, and encouragement given to me to conduct this research. Next, I would like to give my heartfelt gratitude to my panel of reviewers, Dr. Thilina Thanthiriwatta, Dr. C.R.J. Amalraj, and Mrs. Indika Karunarathna, for their valuable review comments, suggestions, and advice during the progress reviews. I would also like to thank the University of Moratuwa for offering me the grant SRC/ST/2024/40 for carrying out my research successfully. Lastly, I would like to thank all the academic and non-academic staff of the Department of Information Technology who helped me in diverse ways to make this research a success and a reality.

ABSTRACT

The creation of user-generated content has increased with the modern development of the Internet. Wikipedia stands out as the world's largest open-source digital encyclopedia, offering free access to extensive knowledge. As of March 2025, Wikipedia has over 6.9 million English articles. However, the collaborative aspect brings into question the accuracy and consistency of the information provided. English Wikipedia receives over 2 edits, over 4000 page views every second, with 500 new articles per day. Due to that, maintaining quality standards remains a major challenge. Though traditional encyclopedias rely on expert review, Wikipedia depends on a collaborative editing process, making quality control more complex. While a frosting wealth of information, skepticism persists among academics regarding its credibility as a reliable source. To address these concerns this research proposed a Wikipedia article content-based quality classification model using deep learning. A feed-forward neural network is combined with large language model-based embedding, text embedding Ada 002 for classification. The model is trained on 5810 English Wikipedia articles, which contain 2905 high-quality, featured articles (FA) and 2905 low-quality articles as A, GA, B, C, Start and Stub class randomly. The supervised embedding-based binary classification model achieved an impressive accuracy of 96.56% with precision, recall and F1 scores of 0.9448, 0.9885 and 0.9661 respectively proving the model's effectiveness and robustness in assessing the content quality. The proposed model can help contributors in enhancing article quality and help readers identify reliable, trustworthy information while strengthening Wikipedia's credibility. Future research should explore integrating metadata and extend the approach towards multilingual Wikipedia editions.

Keywords: Wikipedia, Content based approach, Feed Forward Neural Network, Quality classification, GPT 3.5

TABLE OF CONTENTS

Declaration of the candidate and Supervisor	iii
Dedication	iv
Acknowledgement.....	v
Abstract	vi
Table of Contents	vii
List of Figures	x
List of Tables.....	xi
List of Abbreviations.....	xii
Chapter 1	1
INTRODUCTION	1
1.1 Background of the study	1
1.2 Problem identification.....	4
1.3 Research Questions	4
1.4 Research Objectives	5
1.5 Contribution of the study	5
Chapter 2	7
LITERITURE REVIEW	7
2.1 Literature review overall structure	7
2.2 Metadata based approach	8
2.3 Content-based approach	8
2.3.1 Wikipedia feature-based approach.....	8
2.3.2 Wikipedia content feature-based machine learning techniques	9
2.3.3 Actual article content-based approaches	10
2.3.4 Only using Deep learning technique	10
2.3.5 Embedding-Based Models	11
2.3.6 Transformer based approaches.....	12
2.4 Hybrid based approach.....	13
Chapter 3	14
RESEARCH METHODOLOGY	14
3.1 Overall Architecture.....	14
3.2 Data collection	15
3.3 Data Preprocessing.....	16

3.3.1 HTML and Styling Cleanup.....	16
3.3.2 Text Refinement.....	16
3.3.3 Text Transformations	17
3.4 Embedding Generation.....	17
3.5 Classification Model.....	18
3.5.1 Model overall architecture explanation.....	18
3.5.2 Model hyperparameter tuning.....	19
3.5.3 Model performance graph description	19
3.6 Experimental Setup and Environment.....	20
3.7 Reasons for the alternative model.....	20
Chapter 4.....	23
ALTERNATIVE MODEL RESEARCH METHODOLOGY.....	23
4.1 Methodology architecture	23
4.2 Selected features from the literature	24
4.3 Pre-processing techniques used in the alternative model.....	25
4.3.1 HTML and Styling Cleanup.....	25
4.3.2 Text Refinement.....	26
4.3.3 Text Processing	26
4.4 Model result representation.....	26
4.5 Explainability Methods	26
4.5.1 SHAP	27
4.5.2 LIME	27
4.6 Experimental Setup and Environment	28
Chapter 5.....	29
RESULTS & EVALUATION	29
5.1 Performance Metrics and Tuned Hyperparameters.....	29
5.1.1 Hyperparameter Selection.....	29
5.1.2 Model Performance Evaluation.....	30
5.2 Model Training Performance	31
5.2.1 Accuracy Over Epochs.....	31
5.2.2 Loss Over Epochs	31
5.2.3 F1 Score Over Epochs.....	32
5.2.4 Precision Over Epoch.....	33
5.3 The alternative model results	34
5.3.1 Performance Metrics	34

5.3.2 Confusion Matrix Analysis	34
5.4 Alternative model performance description for SHAP.....	35
5.5 Alternative model performance description for LIME	36
5.6 Evaluation	37
5.6.1 Comparison Models	37
5.6.2 Comparison model-based performance metrics.....	39
5.6.3 Confusion Matrix Analysis: Evaluating the Proposed Model vs. Experimental and Baseline Approaches.....	40
Chapter 6	44
DISCUSSION	44
6.1 Discussion based on the first research question.	44
6.2 Discussion based on the second research question	45
6.3 Implications of the study	45
6.4 Research contribution.....	46
6.5 Limitations & Future Research	46
Chapter 7	48
CONCLUSION	48
References	49
APPENDIX A	55
Wikipedia quality classification model code	55
Wikipedia alternative model codes ,feature extraction, model training and interpretations.....	59
APPENDIX B	64

LIST OF FIGURES

Figure	Description	Page
Figure 1:	Total article text in English Wikipedia, measured in gigabytes	2
Figure 2:	Literature overview	7
Figure 3:	Embedding-based text classification literature overview	11
Figure 4:	Overall Methodology Architecture diagram	14
Figure 5:	Classification Model Architecture	18
Figure 6:	SHAP Summary Plot for the Proposed Model	22
Figure 7:	SHAP Force Plot for the Proposed Model	22
Figure 8:	Alternative Model Architecture	23
Figure 9:	Proposed Model Confusion Matrix	30
Figure 10:	Proposed Model Training and Validation Accuracy	31
Figure 11:	Proposed Model Loss Over Epochs Chart	32
Figure 12:	Proposed Model F1- Score over Epochs Chart	32
Figure 13:	Proposed Model Precision over Epochs	33
Figure 14:	Alternative Model Confusion Matrix	34
Figure 15:	Alternative Model SHAP output	35
Figure 16:	Alternative Model LIME output	37
Figure 17:	Proposed Model Confusion Matrix	40
Figure 18:	BERT + Feed Forward Confusion Matrix	41
Figure 19:	GPT 2+ Feed Forward Confusion Matrix	42
Figure 20:	Doc2Vec + DNN Confusion Matrix	42
Figure 21:	Doc2Vec +Feed-Forward Confusion Matrix	43
Figure 22:	GPT 3.5 +CNN Confusion Matrix	43

LIST OF TABLES

Table	Description	Page
Table 1:	Wikipedia content quality grading scheme	3
Table 2:	Wikipedia Label Justification	15
Table 3:	Alternative model features based on literature	24
Table 4:	Hyperparameter tuning chart	29
Table 5:	Evaluation metrics	30
Table 6:	Alternative Model Performance Metrics	34
Table 7:	SHAP Weightages for feature Contribution	36
Table 8:	Comparison Model evaluation chart	40

LIST OF ABBREVIATIONS

Abbreviation	Description
UGC	user-generated content
IEEE	Institute of Electrical and Electronic Engineers
IQ	Information Quality
SERP	Search Engine Result page
API	Application Programming Interface
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
KNN	K nearest neighbor
EAT	Expertise, authoritativeness, and trustworthiness.
HTML	Hypertext Markup Language
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
TFIDF	Term Frequency-Inverse Document Frequency
GRU	Gated Recurrent Unit
Bi-LSTM	Bidirectional Long Short Term Memory
DNN	Deep Neural Network
LLAMA	Large Language Model Meta AI
GPT	Generative Pre-trained Transformer
	Bidirectional Encoder Representations from
BERT	Transformers