

**Developing a Fine-Tuned Transformer Model to Detect Social Media
Hate Speech Texts**

BY

MASTER EK-KARAT CHUNG

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEMS (HONOURS) INFORMATION SYSTEMS

ENGINEERING

Faculty of Information and Communication Technology

(Kampar Campus)

JUNE 2024

REPORT STATUS DECLARATION FORM

Title: Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts

Academic Session: 202406

I

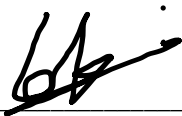
MASTER EK-KARAT CHUNG

(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,



(Author's signature)



(Supervisor's signature)

Address:

No 2, Laluan Indah 1, Taman Indah,

31000, Batu Gajah, Perak

Dr Ramesh Kumar Ayyasamy

Supervisor's name

Date: 10-Sep-2024

Date: 10-Sep-2024

Universiti Tunku Abdul Rahman			
Form Title : Submission Sheet for FYP/Dissertation/Thesis			
Form Number: FM-IAD-004	Rev No.: 0	Effective Date: 21 JUNE 2011	Page No.: 1 of 1

FACULTY/INSTITUTE* OF Information and Communication Technology.

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 10-Sep-2024

SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that **MASTER EK-KARAT CHUNG.** (ID No: 2001610) has completed this final year project/ dissertation/ thesis* entitled “**Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts**” under the supervision of **Dr Ramesh Kumar Ayyasamy** (Supervisor) from the Department of **Information Systems**, Faculty/Institute* of **Information and Communication Technology** , and **Encik Syed Muhammad Bin Sved Omar** (Co-Supervisor)* from the Department of **Information System**, Faculty/Institute* of **Information and Communication Technology** .

I understand that University will upload softcopy of my final year project / dissertation/ thesis* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

MASTER EK-KARAT CHUNG

(Student Name)

DECLARATION OF ORIGINALITY

I declare that this report entitled “**Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts**” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature :  _____

Name : MASTER EK-KARAT CHUNG

Date : 10-Sep-2024

ACKNOWLEDGEMENTS

I would like to express thanks and appreciation to my supervisor, **Dr Ramesh Kumar Ayyasamy** and my moderator, **Encik Syed Muhammad Bin Syed Omar** who have given me a golden opportunity to involve in the **Sentiment analysis, Text mining** field study. Besides that, they have given me a lot of guidance in order to complete this project. When I was facing problems in this project, the advice from them always assists me in overcoming the problems. Again, a million thanks to my supervisor and moderator.

To all friends and tutors in my life, for their patience, unconditional support, and love, and for teaching, guiding, and standing by my side during hard times. Finally, I must say thanks to my parents and my family for their love, support, and continuous encouragement throughout the research.

ABSTRACT

This research explores the development and evaluation of a hate speech detection system using transformer-based models, focusing on the robustness, efficiency, and scalability of the model. The study emphasizes key design considerations, including scalability, which addresses the model's capability to handle large volumes of data, and accuracy, achieved through fine-tuning methods for transformer models like BERT. Reviewing model to do proper performance analysis on existing model in detecting Social Media Hate Speech Texts such as Long Short-Term Memory (LSTM) and Bidirectional Gated Recurrent Unit (Bi-GRU), GigaBERT for Arabic Hate Speech Detection, BERT-Based Approaches, DistilBERT and RoBERTa, T5 and Electra, Comparison of Transformer Models and its Challenges and Limitation. Besides, it also briefly discusses on system design to ensure that the model is conceptually accurate, scalable, and maintainable, providing a flexible framework for ongoing research in hate speech detection on social media. The research also discusses on the facing challenges such as data imbalance, computational limitations, and extensive hyperparameter tuning, all of which were addressed through various techniques and strategies. This research show system's experiment/ simulation to show performance with evaluated using a Logistic Regression model on a split dataset, fine-tuning with GridSearchCV, how the model's accuracy improved. The experiment successfully show a predictive model with high accuracy and precision, also indicated future improvements in detecting hate speech on social media. The results underscore the importance of ongoing refinement in machine learning models or deep learning model to address complex, real-world issues such as hate speech detection.

TABLE OF CONTENTS

TITLE PAGE	i
REPORT STATUS DECLARATION FORM	ii
FYP THESIS SUBMISSION FORM	iii
DECLARATION OF ORIGINALITY	iv
ACKNOWLEDGEMENTS	v
ABSTRACT	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	x
LIST OF TABLES	xi
LIST OF ABBREVIATIONS	xii
CHAPTER 1 INTRODUCTION	1
1.1 Problem Statement and Motivation	1
1.2 Objectives	2
1.3 Project Scope and Direction	3
1.4 Contributions	4
1.5 Report Organization	4
CHAPTER 2 LITERATURE REVIEW	5
2.1 Introduction	5
2.2 Review of Deep Learning Models for Hate Speech Detection	5
2.2.1 Long Short-Term Memory (LSTM) and Bidirectional Gated Recurrent Unit (Bi-GRU)	5 5
2.2.2 GigaBERT for Arabic Hate Speech Detection	6
2.2.3 BERT-Based Approaches	7-8
2.2.4 DistilBERT and RoBERTa	9
2.2.5 T5 and Electra	10-11
2.3 Comparison of Transformer Models	12
2.4 Challenges and Limitations	12
2.5 Conclusion	12

CHAPTER 3 SYSTEM MODEL (FOR RESEARCH-BASED PROJECT)	13
3.1 Introduction	13
3.2 System Architecture	13
3.3 Data Flow	14
3.4 Conclusion	14-15
CHAPTER 4 SYSTEM DESIGN	16
4.1 Introduction	16
4.2 Design Considerations	16
4.3 Conclusion	16
CHAPTER 5 EXPERIMENT/SIMULATION (FOR RESEARCH-BASED PROJECT)	17
5.1 Hardware Setup	17
5.2 Software Setup	17
5.3 Setting and Configuration	18
5.4 System Operation (with Screenshot)	18-19
5.5 Implementation Issues and Challenges	20
5.6 Concluding Remark	20
CHAPTER 6 SYSTEM EVALUATION AND DISCUSSION	21
6.1 System Testing and Performance Metrics	21-22
6.2 Enhancing Testing Setup and Result	22
6.3 Project Challenges	24
6.4 Objectives Evaluation	24
6.5 Concluding Remark	24

CHAPTER 7 CONCLUSION AND RECOMMENDATION	25
7.1 Conclusion	25
7.2 Recommendation	25-26
REFERENCES	27-29
APPENDICES A	30
A.1 Weekly Report (Week 2)	30
A.2 Weekly Report (Week 4)	31
A.3 Weekly Report (Week 6)	32
A.4 Weekly Report (Week 8)	33
A.5 Weekly Report (Week 10)	34
A.6 Weekly Report (Week 12)	35
POSTER	36
PLAGIARISM CHECK RESULT	37-44
FYP2 CHECKLIST	45

LIST OF FIGURES

Figure Number	Title	Page
Figure 1	Result and Discussion	5
Figure 2	Result Table	6
Figure 3	Workflow of GigaBERT-V4	6
Figure 4	Model Comparison among 4 models	7
Figure 5	BERT based model workflow	8
Figure 6	Test Result	8
Figure 7	Architecture of Distillbert	9
Figure 8	Accuracy Scores of Benchmarks and for proposed DISTILBERT MODEL	10
Figure 9	T5 pre-training process	10
Figure 10	Evaluation of each model	11
Figure 11	Sample of Transformer Model Architecture Flow	15
Figure 12	Anaconda Prompt	18
Figure 13	Show the process of data preprocessing	19
Figure 14	Show of Model Training Process	19
Figure 15	Showcase of model performance using confusion matrix	20
Figure 16	Model Building (before tune) 1	21
Figure 17	Model Building (before tune) 2	21
Figure 18	Test Accuracy (Before fine tune)	22
Figure 19	Performance after fine-tune	23

LIST OF TABLES

Table Number	Title	Page
Table 1.1	Specifications of laptop	4

LIST OF ABBREVIATIONS

<i>LSTM</i>	Long Short-Term Memory
<i>BI-GRU</i>	Bidirectional Gated Recurrent Unit
<i>NLP</i>	Natural language processing
<i>BERT</i>	Bidirectional Encoder Representations from Transformer
<i>GPT</i>	Generative Pretrained Transformer
<i>RoBERTa</i>	Robustly Optimized BERT Approach
<i>DistillBERT</i>	Distilled version of Bidirectional Encoder Representations from Transformer
<i>T5</i>	Text-To-Text Transfer Transformer

Chapter 1: Introduction

In this chapter will present the background and motivation of research, contributions to the field, and the outline of the thesis which including Problem Statement and Motivation, Research Objectives, Project Scope and Direction, Contributions and Report Organization.

1.1 Problem Statement and Motivation

Continuing the exploration of deep learning transformers in Task 2 about review the latest deep learning transformers that can be utilized and customized to solve the NLP task, a critical challenge persists: evaluating the effectiveness of these models in detecting hate speech on social media platforms. The evolving of online communication environment complicates this task. While transformer-based models like BERT, GPT, and others have shown their performance different significantly across different contexts and datasets. This unstable raises concerns of the reliability and stability of these models in real-world applications because the core of the problem lies in the complexity of hate speech itself.

Moreover, the lack of standardized evaluation metrics and the limited availability of high-quality annotated datasets further serious the challenge. The inconsistent performance of existing models show it need for a systematic analysis of strengths and weaknesses for the model. Understanding these nuances is so important for advancing the field and making sure that the models can effectively adapt to the complexities of social media hate speech detection.

Based on the foundational research of transformer models in hate speech detection, the motivation for this task is to perform a comprehensive performance analysis of existing models. As social media continues to expand or new incoming new social media platform, the prevalence of hate speech and its impact on social media causing more problems occurs. To solve this problem, it is important to assessment on how well current transformer-based models address the issues and identify how to do improvements are needed.

This task is motivated by the need to research the gap between theoretical advancements in NLP and practical applications in hate speech detection. By systematically reviewing the performance of these models in detecting Social Media Hate Speech Texts. This analysis will

not only contribute to the academic understanding of transformer models but also provide better insights for improving hate speech detection systems. This is not only important for reducing the spread of harmful content but also for protecting freedom of expression by ensuring that content is accurately classified.

1.2 Objectives

In this research, the primary goal of this task is to conduct a structured literature review focused on the performance of existing transformer-based models in detecting hate speech on social media. The research objectives include:

1. **Performance Evaluation of Transformer Models:** Analyze the performance of transformer-based models such as BERT, GPT, RoBERTa, DistilBERT, Generative Pretrained Transformer [13] and others in detecting hate speech or Natural Language Processing Tasks. This involves reviewing existing literature to compare their accuracy, precision, recall, and other relevant metrics.
2. **Analysis of Challenges and Limitations:** Identify the challenges and limitations faced by these models in hate speech detection. This includes issues related to data availability, model generalization, and the handling of hate speech on social media. Analyze the limitations related to the quality and quantity of annotated datasets, the adaptability of models to different social media platforms, and their interpretability.
3. **Identification of Research Gaps:** Highlight areas where current models fall short and suggest potential research directions for future research. This includes exploring new techniques or enhancements that could improve model performance in hate speech detection.

1.3 Project Scope and Direction

The scope of this task involves a detailed review and analysis of the performance of transformer-based models in the specific context of hate speech detection.

The first direction will be on Comprehensive Review of Model Performance by conducting review of existing studies that evaluate the performance of transformer models in hate speech detection. This will include analyzing various metrics and benchmarks used in literature.

Second, exploration of model limitations by investigate the common or occurs limitations and challenges that identified in the literature. Such as the model struggles with understanding context, handling multi-lingual or cross-lingual data, and their sensitivity to changes in language, cultures all over the time. It needs to focus on issues and helping to solve the model limitations issue.

Third, identification of future research directions. Based on the analysis, identify out the research gap and have proper performance analysis on existing model that could help to enhance their performance in detecting hate speech on social media platforms. This may include the development of hybrid models that combine the strengths of transformer model with other approaches, or identify out more training techniques that better capture the nuances of hate speech.

By focusing on these key areas, the research project aims to provide valuable insights into leveraging deep learning transformers for effective hate speech detection, contributing to advancements in detecting Social Media Hate Speech Texts.

1.4 Contributions

This research task makes several key contributions to the field of performance analysis on existing model in detecting Social Media Hate Speech Texts:

1. **Depth Performance Analysis:** Provides a proper analysis of the performance of various transformer-based models in hate speech detection by reviewing existing literature and evaluating multiple (old and latest) transformer-based models, it provides a clear overview of the existing model's performance in detecting Social Media Hate

Speech Texts and know their pro and cons. It will also offering better view into their strengths and weaknesses so that can help to enhance on it.

2. **Identification of Critical Challenges:** Highlights the significant challenges faced by these models in real-world applications, providing a foundation for future research aimed at addressing these issues. Identified solving problem Language complexity, differing views on what constitutes hate speech, and data availability restrictions for algorithm training and testing.
3. **Guidance for Future Research:** This research offers useful information and summarization for future studies or development use in hate speech detection using Transformer model based. By highlighting research gaps, suggesting areas for improvement, and benchmarking on each transformer model pro and cons, it helps advancements in developing reliable hate speech detection technologies for social media platform.
4. **Benchmarking Existing Models:** Establishes a benchmark for evaluating the performance of transformer models in hate speech detection throughout the research that had been done. This benchmarking will help to standardize the evaluation process and ensure that future research builds on a great foundation of comparable data.

Overall, this research contributes to advancing knowledge in existing model for hate speech detection. It will offering valuable insights for researchers, practitioners, and policymakers.

1.5 Report Organization

This report is organized into 5 chapters. Chapter 1: Introduction, including the problem statement, motivation, research objectives, project scope, and contributions. Chapter 2: Literature Review, presenting a detailed analysis of existing research on transformer models in hate speech detection. Chapter 3: System Model. Chapter 4: System Design, Chapter 5: Experimental and Simulation, Chapter 6: System Evaluation and Discussion, Chapter 7: Conclusion and Recommendation. Summarizing the contributions of the research and suggesting directions for future studies.

Chapter 2: Literature Review

2.1 Introduction

The research of detection of hate speech on social media platforms has been a significant focus due to the growing over the spread of harmful content at online [9], [16]. By using transformer models, especially with BERT based model or any hybrid with BERT based model. With their powerful language representation capabilities, have emerged as a promising approach to address this challenge. This chapter will be reviews the existing literature on transformer-based models, along with other deep learning approaches, to assess their performance in hate speech detection.

2.2 Review of Deep Learning Models for Hate Speech Detection

2.2.1 Long Short-Term Memory (LSTM) and Bidirectional Gated Recurrent Unit (Bi-GRU)

LSTM networks and Bi-GRU have been widely applied in the field of NLP, particularly for tasks requiring the understanding of sequential data [11], [14], [17], [18]. A study by demonstrated that a combination of LSTM and Bi-GRU models achieved an accuracy of 90.51% on the training set and 87.51% on the validation set. While these results are promising, the models struggle with capturing the nuanced context of hate speech, particularly when language evolves rapidly on social media platforms. [1] But the model still can enhancing by expanding the dataset and fine-tuning hyperparameters and exploring more advanced architecture to optimize the model performance.

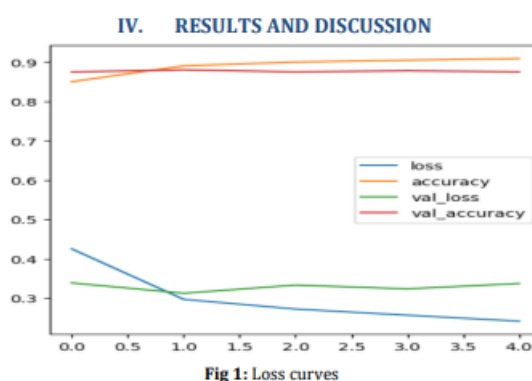


Figure 1. Result and Discussion

Table 1. Metrics

SN.	Metrics	Results
1	Accuracy	90.50
2	precision	86.49
3	recall	87.50
4	F1-Score	86.51

Figure 2. Result (Table)

2.2.2 GigaBERT for Arabic Hate Speech Detection

Detecting hate speech in languages other than English, such as Arabic. It shows out the limitations of Hate Speech Detection which is complexity of languages in social media [12], [15]. It presents unique challenges due to the language's rich vocabulary and diverse dialects [2]. The training dataset for this model from author shows 8662 rows, including 986 for Hateful tweets and 7676 for Not Hateful tweets. In data cleaning processing are eliminating diacritics, punctuation, repeated characters, hashtags, links, usernames, and emojis then oversampling the imbalanced data. The author using GigaBERT-v4 with an oversampling technique to detect hate speech in Arabic Twitter data. After fine-tuning the pre-trainer, it achieving an impressive F1-Score of 0.9700 and an accuracy of 0.9930. The research highlights that the effectiveness of transformer BERT models in handling complex languages, though it also underscores the need for large and diverse datasets to train these models effectively.

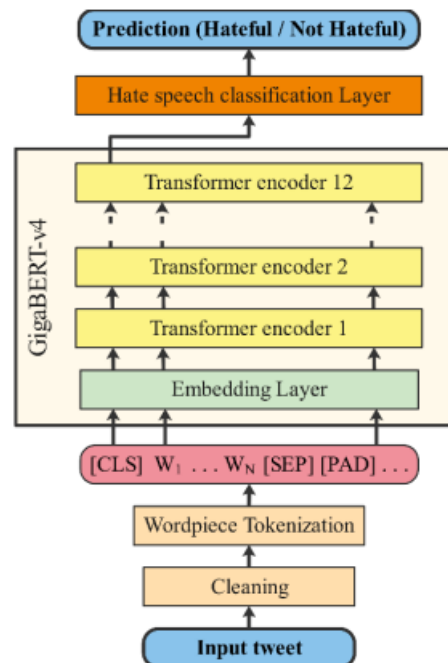


Figure 3. WorkFlow of GigaBERT-v4

Model	Without oversampling		Using oversampling	
	Accuracy %	F1-score %	Accuracy %	F1-score %
AraBERT	99.22	98.05	99.35	98.41
mBERT	98.80	97.04	99.04	97.68
XLM-Twitter	98.87	97.17	99.07	97.75
GigaBERT-v4	99.27	98.25	99.46	98.68

Figure 4. Model Comparison among 4 Models

2.2.3 BERT-Based Approaches

BERT has become a widely recognized standard for various NLP tasks, including hate speech detection [10]. A study [3] demonstrated that a BERT model trained with focal loss achieved an accuracy of 98.03% and an F1-score of 98.02% when detecting hate speech in Arabic dialects. BERT's ability to capture context at a detailed level makes it particularly effective for this task. However, its performance is closely tied to the quality of the training data. To address the issue of class imbalance in the dataset used for training and testing, the researchers employed a resampling technique. This involved augmenting the minority class by using translation and back-translation strategies to generate additional instances.

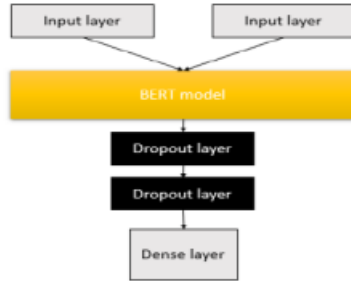


Fig. 1. BERT baseline model.

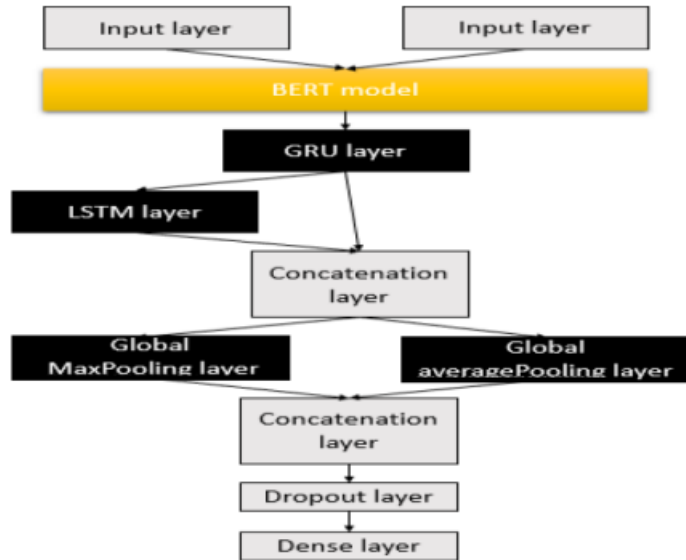


Fig. 2. Proposed model.

Figure 5. BERT based model work flow

Table 3. Test results.

Dataset	Original dataset	Augmented dataset
---------	------------------	-------------------

Models	Baseline Bert	Proposed Bert	Baseline Bert	Proposed Bert
Accuracy	87.09%	88.51%	87.09%	89.61%
F1-Score	97.04%	97.46%	97.04%	97.78%

Figure 6. Test Result

2.2.4 DistilBERT and RoBERTa

DistilBERT, a smaller and faster variant of BERT, recorded an average accuracy of 92% in detecting hate speech on Twitter [4]. The author researchers tested DistilBERT against other models, including BERT, XLNet, RoBERTa, and attention-based LSTM, using a publicly available multiclass hate speech corpus of 24,783 labeled tweets. RoBERTa, another transformer model based on BERT, achieved a slightly lower accuracy of 90-91% by compares the performance of DistilBERT with other models across several metrics, including accuracy, precision, recall, F-measure, Mathews correlation coefficient (MCC), and evaluation loss. The results demonstrate that DistilBERT outperforms the other models in detecting hate speech. These models offer a good trade-off between performance and computational efficiency, making them viable options for real-time hate speech detection. The author also concludes by discussing the implications of the findings and suggesting areas for future research, such as improving transformer models' efficiency for deployment in real-time or resource-constrained environments.

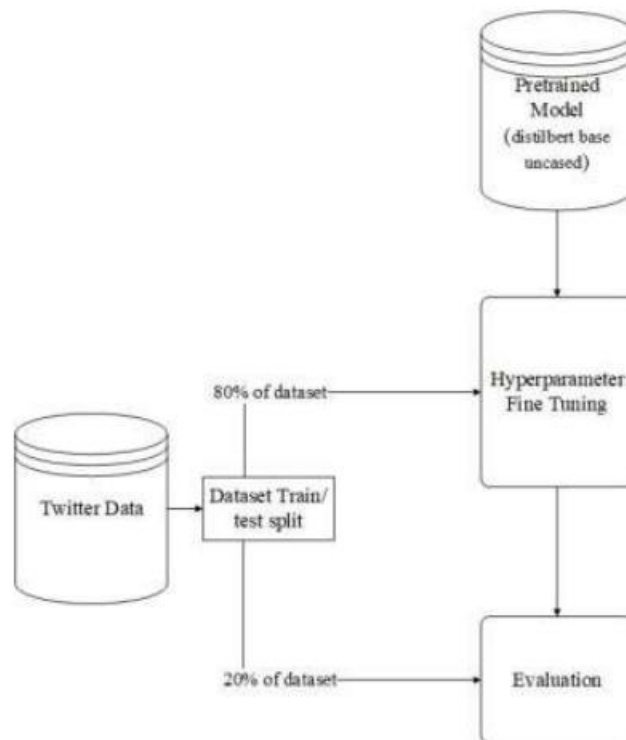


Figure 7. Architecture of Distillbert

TABLE II. ACCURACY SCORES OF FOUR BENCHMARK HATE SPEECH DETECTION ALGORITHMS AND THE PROPOSED DISTILBERT MODEL ON THE HSO DATASET

Algorithm	Method Name	Accuracy
BERT	bert-base-uncased	0.90
RoBERTa	robert-base	0.91
RoBERTa	robert-base-openai-detector	0.90
XLNet	xlm-mlm-en-2048	0.91
LSTM with Attention		0.66
<u>DistilBERT</u>	<u>distilbert-base-uncased</u>	<u>0.92</u>

Figure 8. Accuracy Scores of Benchmark and for proposed DISTILBERT MODEL

2.2.5 T5 and Electra

The T5 model, although powerful, requires more computational resources than BERT due to its extensive training on a broader range of NLP tasks [5]. On the other hand, Electra, which uses a more efficient pre-training objective, achieved an F1-score of 0.8216. However, when combined with BERT in an ensemble model, the F1-score improved to 0.8342. Besides these advancements, the high training costs associated with these models remain a significant drawback. The author also mentioned that the limitation for training T5 model requires a large amount of computing resources so that did not experiment with large T5 models such as T5-XL and T5-XXL. While these models might perform better than standard T5 models.

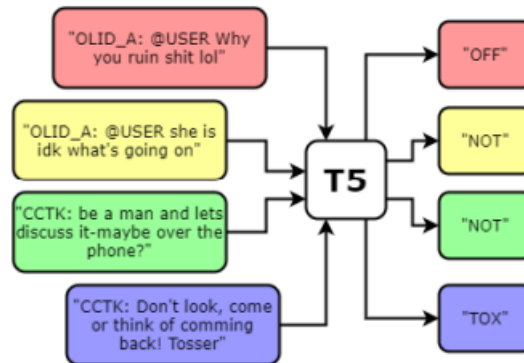


Figure 1: T5/ MT5 pre-training process

Figure 9. T5 pre-training process

ELECTRA, a more recent Transformer-based model, introduces a generator-discriminator framework where the generator replaces some input tokens with plausible alternatives, and the discriminator predicts whether each token in the input was replaced or not. The research discusses the application of Transformer-based models, including BERT and ELECTRA[6], for detecting offensive language in Turkish tweets. The study is based on a dataset of over 36,000 manually annotated tweets, and the authors experimented with various models, including Logistic Regression, Deep Neural Networks (DNN), and Gated Recurrent Units

(GRU) [8]. Among these, pre-trained Transformer models outperformed traditional methods, highlighting the importance of contextual word representation in natural language processing tasks like offensive language detection.

This approach enables more efficient training by focusing on all tokens in the input sequence rather than just a masked subset. As a result, the ELECTRA model achieved the highest F1 score in the study, demonstrating its superiority in detecting offensive language in social media posts. The ensemble of ELECTRA and BERT further improved performance, setting a new benchmark for this task [6].

The disadvantages of ELECTRA model is Training cost. Training transformers networks requires a large amount of computation time and dataset and consumes a significant amount of power when pre-trained models are not preferred to use.

TABLE I: PRECISION, RECALL AND F1-SCORE RESULTS OF THE MODELS

Models	Precision	Recall	F1 score
ELECTRA	0.82	0.82	0.8216
BERT	0.83	0.78	0.8053
LR	0.76	0.76	0.7619
FastText	0.75	0.74	0.7535
DNN	0.75	0.74	0.7450
GRU	0.76	0.73	0.7446

TABLE II: ENSEMBLE MODEL RESULT AND TOP 10 MODELS THAT RANKED ACCORDING TO MACRO AVERAGED F1 IN THE SEMEVAL-2020 TASK 12 [7] COMPETITION FOR TURKISH LANGUAGE TASK A

Models	F1 score
Ensemble Model (ELECTRA models+BERT)	0.8342
Galileo [26]	0.8258
ELECTRA	0.8216
SU-NLP [27]	0.8167
KUISAIL [28]	0.8141
KS@LTH [29]	0.8101
NLPDove [30]	0.7967
TysonYU	0.7933
RGCL	0.7859
Rouges [31]	0.7815
GruPaTo [32]	0.7790
MindCoders	0.7789

Figure 10. Evaluation of each model

2.3 Comparison of Transformer Models

BERT consistently outperforms other models in terms of accuracy and F1-score across various studies, making it the most suitable model for hate speech detection. From the latest relevant study that I had research show the same result even in different sector [7]. However, each model has its strengths and weaknesses, particularly in terms of computational efficiency, fine-tuning techniques, different combination of model and the ability to generalize across different languages and contexts will cause very different to the performance.

2.4 Challenges and Limitations

While transformer models show great performance results but they are not without their challenges. The limitations of it are need for large, well-annotated datasets, the computational cost to training these models, and their sensitivity to language and context changes all over the time are still quite a big problem on it. Plus, the lack of standardized evaluation metrics makes it difficult to compare results across studies and computational efficiency also affect the result from previous study shown.

2.5 Conclusion

This literature review indicates that transformer-based models, particularly BERT and its variants of combination, which provide the best performance for hate speech detection. However, there is still spaces for further improvement, especially in developing more efficient models and creating better datasets for training and evaluation so that can help in create a better environment for social media on detecting hate speech text.

Chapter 3: System Model

3.1 Introduction

Due to this is a research project. This chapter presents the conceptual system model and architecture designed to show out the effectiveness of transformer-based models, particularly BERT, in detecting hate speech on social media platforms. The focus is on outlining the components, processes, and theoretical that about the model's design, without actual implementation or deployment.

3.2 System Architecture

The conceptual system is composed of several key components, each integral to the research study:

Data Collection Framework: The research framework involves the authors / researchers gathering social media data from platforms that with high volumes of user-generated content. The data is pre-processed theoretically to remove noise/ redundancy data and irrelevant information, setting datasets for model training and evaluation.

Data Preprocessing Strategies: This section conceptualizes the methods for data cleaning, feature engineering, identify missing values, outliers in datasets. Better visualize to the datasets so that can know problem to the datasets. A clean data preprocessing strategies will help the model training out with higher accuracy and performance.

Transformer Model Theoretical Layer: This layer focuses on transformer model such as BERT-BASED model in the research. The model fine-tuned on a dataset of hate speech examples using transfer learning techniques, which are explored and evaluated in this research to determine their effectiveness in detecting hate speech on the social media. The reason of using this technique because of there are big variations in the social media that generated by users.

Prediction Module (Theoretical): Although this research is not implemented any model, but this section describes how predictions would be generated, categorizing text as either hate

speech or non-hate speech from the user-generated content show as literature above [4]. And include theoretical approaches to generating and interpreting confidence scores.

Evaluation and Feedback Concept: This represents the theoretical evaluation of the model's performance metrics such as accuracy, precision, recall, and F1-score. The potential for integrating feedback loops for continual improvement is also considered. It will help better visualize for the data for the researchers from the results and enhance the model based on the feedback.

3.3 Data Flow

The research also shows the data flow beginning from the Data Collection Framework, through preprocessing strategies, and implement different Transformer Model Layer for prediction, fine-tuning until get a better result from the model. Finally, discussed and research gap from the context of research analysis and model performance testing.

3.4 Conclusion

This system model chapter show out the theoretical strengths of transformer-based models, providing a research framework that underscores the potential for high accuracy in hate speech detection. It also show the ability to adapt to new data and the considerations for scalability are also emphasized that the transformer model are high at scalability to adapt on different environments and able to make adjustments such as combining different model to get a better results on hate speech detection.

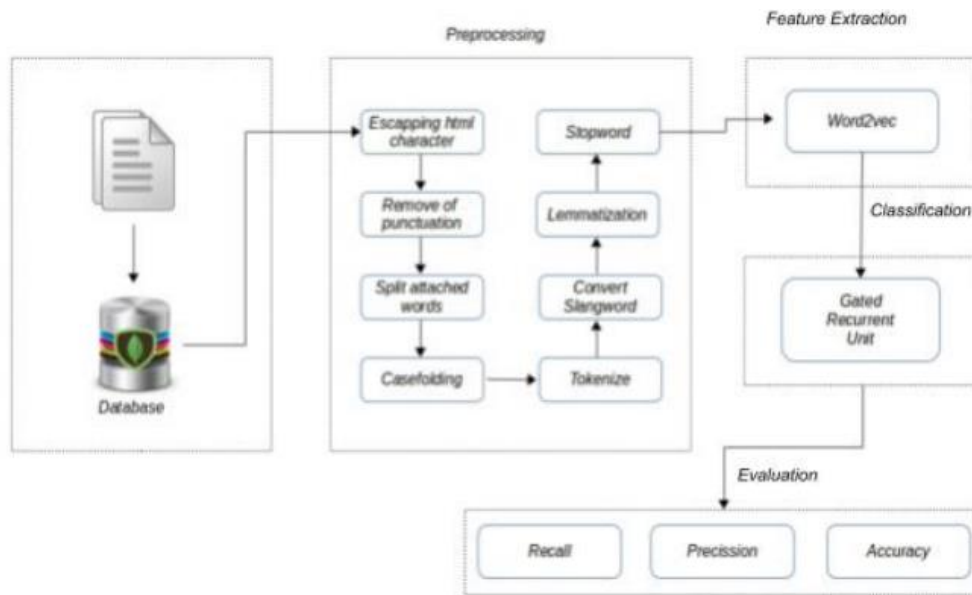


Figure 11. Sample of Transformer Model Architecture Flow

Chapter 4: System Design

4.1 Introduction

This chapter will discuss about the theoretical design of a hate speech detection system, focusing on the robustness, efficiency, and scalability of the transformer model. Through the whole research work, the model should have or enhancing based on design considerations, conceptual strategies, and others explored during the research.

4.2 Design Considerations

Main Key design considerations or can enhancement after throughout the research project. It needs to be included:

Scalability: The research emphasizes applying the model theoretically capable of handling large volumes of data. The scalability is considered in terms of potential model architecture and processing capabilities. And important of hardware that affect the model performance and processing capabilities.

Accuracy: The research also needs to consider on achieving high accuracy by exploring new fine-tuning methods for transformer models like BERT (because this model achieves a good performance among the model). The research can be including an in-depth examination of how different strategies can enhance model performance with different techniques and strategies.

4.3 Conclusion

The system design discussed in this research ensures that the hate speech detection model is conceptually accurate, scalable, and maintainable. The design need to allows for future enhancements and updates, providing a flexible framework for ongoing research in hate speech detection on social media. The better design the transformer model, the more accurate will detect Hate Speech in social media.

CHAPTER 5: EXPERIMENT/SIMULATION

5.1 Hardware Setup

The hardware involved in this research is computer, mobile device, and tablet. A computer, mobile device and tablet all are used for doing research and collect research use. It is because it needs to study a lot of articles, know how the transformer model works through by education video. Tablet can use for better search of resources that needed.

Table 1.1 Specifications of laptop

Description	Specifications
Model	HP Pavilion Gaming
Processor	Intel(R) Core(TM) i5-9300H
Operating System	Windows 11
Graphic	NVIDIA GeForce GTX 1050
Memory	12GB DDR4 RAM
Storage	500GB SSD + 1TB HDD

5.2 Software Setup

The experiment was conducted using Jupyter Lab, which platform that mostly use on machine learning. It supports the execution of code, visualization of data/ performance, and documentation in a single platform. The following software tools and libraries were utilized:

1. **Python:** Programming language used for the experiment.
2. **Jupyter Lab:** Provided an interactive environment for coding, visualizing, and documenting the process.
3. **Pandas:** For data manipulation and analysis.
4. **NumPy:** For numerical operations.
5. **NLTK:** For natural language processing tasks such as tokenization and stopword removal.
6. **WordCloud:** To generate word clouds for visualizing text data.
7. **Scikit-learn:** For machine learning tasks including model training, evaluation, and hyperparameter tuning.
8. **Matplotlib and Seaborn:** For creating visualizations like count plots, pie charts, and confusion matrices.

5.3 Setting and Configuration

In the Jupyter Lab environment, the setting and configuration include:

1. **Kernel and Environment:** Python 3 as the kernel, and necessary packages were installed using pip.
2. **Package Setup:** Key packages like pandas, numpy, nltk, scikit-learn, matplotlib, seaborn, and wordcloud were installed and regularly updated (include in the jupyter lab).
3. **Structure:** The structure was organized into sections for data exploration, preprocessing, visualization, modeling, and evaluation.
4. **Random State:** A fixed random_state=42 was used during data splitting.
5. **Visualization Style:** Using plot (style as “ggplot”) for visualizations.

5.4 System Operation (with Screenshot)

Start up System : Using Anaconda Prompt to start up Jupyter Lab [19]

```
(base) C:\Users\User>jupyter lab C:\Users\User\Downloads\hate-speech-detection-using-machine-learning-main\hate-speech-detection-using-machine-learning-main
[I 2024-08-29 19:44:51.795 ServerApp] jupyter_server_fileid | extension was successfully linked.
[I 2024-08-29 19:44:51.800 ServerApp] jupyter_server_ydoc | extension was successfully linked.
[I 2024-08-29 19:44:51.805 ServerApp] jupyterlab | extension was successfully linked.
[I 2024-08-29 19:44:51.809 ServerApp] nbclassic | extension was successfully linked.
[I 2024-08-29 19:44:52.156 ServerApp] notebook_shim | extension was successfully linked.
[I 2024-08-29 19:44:52.156 ServerApp] panel.io.jupyter_server_extension | extension was successfully linked.
[I 2024-08-29 19:44:52.181 ServerApp] notebook_shim | extension was successfully loaded.
[I 2024-08-29 19:44:52.181 FileIdExtension] Configured File ID manager: ArbitraryFileIdManager
[I 2024-08-29 19:44:52.181 FileIdExtension] ArbitraryFileIdManager : Configured root dir: C:/Users/User/Downloads/hate-speech-detection-using-machine-learning-main/hate-speech-detection-using-machine-learning-main
[I 2024-08-29 19:44:52.181 FileIdExtension] ArbitraryFileIdManager : Configured database path: C:\Users\User\AppData\Roaming\jupyter\file_id_manager.db
[I 2024-08-29 19:44:52.182 FileIdExtension] ArbitraryFileIdManager : Successfully connected to database file.
[I 2024-08-29 19:44:52.182 FileIdExtension] ArbitraryFileIdManager : Creating File ID tables and indices with journal_mode = DELETE
[I 2024-08-29 19:44:52.182 ServerApp] jupyter_server_fileid | extension was successfully loaded.
[I 2024-08-29 19:44:52.183 ServerApp] jupyter_server_ydoc | extension was successfully loaded.
[I 2024-08-29 19:44:52.183 LabApp] JupyterLab extension loaded from C:\Users\User\anaconda3\Lib\site-packages\jupyterlab
[I 2024-08-29 19:44:52.184 LabApp] JupyterLab application directory is C:\Users\User\anaconda3\share\jupyterlab
[I 2024-08-29 19:44:52.186 ServerApp] jupyterlab | extension was successfully loaded.
```

Figure 12. Anaconda Prompt

Data Preprocessing: Show how text data was cleaned, tokenized, and transformed.

```
[18]: #creating a function to process the data
def data_processing(tweet):
    tweet = tweet.lower()
    tweet = re.sub(r"https\S+|www\S+http\S+", '', tweet, flags = re.MULTILINE)
    tweet = re.sub(r'\@w+|\#','', tweet)
    tweet = re.sub(r'\^w\s','',tweet)
    tweet = re.sub(r'0','',tweet)
    tweet_tokens = word_tokenize(tweet)
    filtered_tweets = [w for w in tweet_tokens if not w in stop_words]
    return " ".join(filtered_tweets)

[19]: tweet_df.tweet = tweet_df['tweet'].apply(data_processing)

[20]: tweet_df = tweet_df.drop_duplicates('tweet')

[21]: lemmatizer = WordNetLemmatizer()
def lemmatizing(data):
    tweet = [lemmatizer.lemmatize(word) for word in data]
    return data

[22]: tweet_df['tweet'] = tweet_df['tweet'].apply(lambda x: lemmatizing(x))

[23]: # printing the data to see the effect of preprocessing
print(tweet_df['tweet'].iloc[0],"\n")
print(tweet_df['tweet'].iloc[1],"\n")
print(tweet_df['tweet'].iloc[2],"\n")
print(tweet_df['tweet'].iloc[3],"\n")
```

Figure 13. Show the process of data preprocessing

Model Training: Displays the training process, including the loss and accuracy metrics.

Model Building

```
[39]: X = tweet_df['tweet']
      Y = tweet_df['label']
      X = vect.transform(X)

[40]: x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

[41]: print("Size of x_train:", (x_train.shape))
      print("Size of y_train:", (y_train.shape))
      print("Size of x_test: ", (x_test.shape))
      print("Size of y_test: ", (y_test.shape))

      Size of x_train: (23476, 380305)
      Size of y_train: (23476,)
      Size of x_test: (5869, 380305)
      Size of y_test: (5869,)

[42]: logreg = LogisticRegression()
      logreg.fit(x_train, y_train)
      logreg_predict = logreg.predict(x_test)
      logreg_acc = accuracy_score(logreg_predict, y_test)
      print("Test accuracy: {:.2f}%".format(logreg_acc*100))

      Test accuracy: 93.17%

[43]: print(confusion_matrix(y_test, logreg_predict))
      print("\n")
      print(classification_report(y_test, logreg_predict))
```

Figure 14. Show of Model Training Process

Evaluation Results: Evaluate the model's performance metrics like precision, recall, F1-score, and confusion matrix.

```
[47]: y_pred = grid.predict(x_test)

[48]: logreg_acc = accuracy_score(y_pred, y_test)
print("Test accuracy: {:.2f}%".format(logreg_acc*100))
Test accuracy: 94.89%

[49]: print(confusion_matrix(y_test, y_pred))
print("\n")
print(classification_report(y_test, y_pred))

[[5450  8]
 [ 292 119]]

              precision    recall  f1-score   support

     0       0.95         1.00         0.97         5458
     1       0.94         0.29         0.44          411

 accuracy          0.95          0.95          0.95          5869
 macro avg         0.94         0.64         0.71          5869
 weighted avg         0.95         0.95         0.94          5869
```

Figure 15. Showcase of model performance using confusion matrix

5.5 Implementation Issues and Challenges

Challenges facing during implementation:

1. **Data Imbalance:** The dataset given had an imbalance between hate speech and non-hate speech examples, which required techniques such as oversampling.
2. **Computational Limitations:** Training transformer models require great hardware to , and despite using a GPU, certain configurations led to longer processing times.
3. **Hyperparameter Tuning:** Identifying the optimal hyperparameters required extensive experimentation, which was both time-consuming and computationally demanding.

5.6 Concluding Remark

The experiment/simulation conducted in Jupyter Lab provided valuable insights into the behaviour and performance of similar to transformer-based models in detecting hate speech. Besides of facing challenges, the approach taken will have a thorough exploration and understanding of the model's capabilities and know the fundamental knowledge of how the model work.

CHAPTER 6:

SYSTEM EVALUATION AND DISCUSSION

6.1 System Testing and Performance Metrics

In this chapter discuss on evaluate the performance of the developed model using several key metrics such as accuracy, precision, recall, and F1-score. The Logistic Regression model was using on this experimental, and its performance was assessed on the test dataset [19].

The experiment setup involved splitting the dataset into training and testing sets with a ratio of 80:20. The Logistic Regression model was then trained on the training set, and its performance was evaluated on the test set.

Details as below:

```
▼ Model Building 🔍  
39]: X = tweet_df['tweet']  
Y = tweet_df['label']  
X = vect.transform(X)  
  
40]: x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=42)  
  
41]: print("Size of x_train:", (x_train.shape))  
print("Size of y_train:", (y_train.shape))  
print("Size of x_test: ", (x_test.shape))  
print("Size of y_test: ", (y_test.shape))  
Size of x_train: (23476, 380305)  
Size of y_train: (23476,)  
Size of x_test: (5869, 380305)  
Size of y_test: (5869,)  
  
42]: logreg = LogisticRegression()  
logreg.fit(x_train, y_train)  
logreg_predict = logreg.predict(x_test)  
logreg_acc = accuracy_score(logreg_predict, y_test)  
print("Test accuracy: {:.2f}%".format(logreg_acc*100))
```

Figure 16. Model Building (before fine tune) 1

```
[43]: print(confusion_matrix(y_test, logreg_predict))  
print("\n")  
print(classification_report(y_test, logreg_predict))  
  
[[5458  0]  
 [ 401 10]]  
  
              precision    recall  f1-score   support  
  
 0           0.93       1.00       0.96       5458  
 1           1.00       0.02       0.05        411  
  
 accuracy          0.93       0.93       0.93       5869  
 macro avg         0.97       0.51       0.51       5869  
 weighted avg      0.94       0.93       0.90       5869
```

Figure 17. Model Building (before fine tune) 2

Explanation :

Model Building

The used of Training Set Size: 23,476 samples, Testing Set Size: 5,869 samples, Feature Space: 380,305 features.

Model Performance (before fine tuning)

Test Accuracy: 93.17%

Confusion Matrix:

True Negatives(TN): 5,458, False Positives(FP): 0,

False Negatives (FN): 401, True Positives (TP): 10

The initial evaluation showed that the model accurately classified the majority of the samples, with a small number of false negatives and false positives. The overall accuracy of the model was 93.17% (Figure 17), also showing its strong performance. However, the recall for class 1 was notably low, it means that the model's difficulty in identifying positive samples. Recall value come with 1.0 it also mean the model can find all instances of the target class in the dataset. Recall can also be called sensitivity or true positive rate and as the higher the recall the better.

```
[42]: logreg = LogisticRegression()
logreg.fit(x_train, y_train)
logreg_predict = logreg.predict(x_test)
logreg_acc = accuracy_score(logreg_predict, y_test)
print("Test accuracy: {:.2f}%".format(logreg_acc*100))
Test accuracy: 93.17%
```

Figure 18. Test Accuracy (Before fine tune)

6.2 Enhanced Testing Setup and Results

To better refine the model, applying GridSearchCV (figure 18) to do hyperparameter tuning, and the model's performance show significantly improved. The final model achieved an increased-on test set with accuracy of 94.89%.

Enhanced Model Performance

Final Test Accuracy: 94.89%

Confusion Matrix:

True Negatives(TN): 5,450 , False Positives(FP): 8

False Negatives(FN): 292, True Positives(TP): 119

The confusion matrix from the final model shows that the false negatives were reduced from 401 to 292 after fine-tuning, and it also mean that model was able to correctly identify a greater number of positive samples compared to the previous work(before fine-tuning). It also reason of the accuracy increased of the modelling.

Detailed Metrics:

Precision:

Class 0: 0.95, Class 1: 0.94

Recall:

Class 0: 1.00, Class 1: 0.29

F1-Score:

Class 0: 0.97, Class 1: 0.44

Accuracy: 94.89%

Macro Average:

Precision: 0.94, Recall: 0.64, F1-Score: 0.71

Weighted Average: Precision: 0.95, Recall: 0.95, F1-Score: 0.94

The final metrics from **figure 18**, show a significant improvement in the model's ability to identify positive cases (class 1), with increasing in recall and F1-score for class 1. However, the recall for class 1 remains low. Recall means the higher the better. It also indicates that there are room for improvement to prevent the model continues to miss some positive cases.

```
[45]: from sklearn.model_selection import GridSearchCV
import warnings
warnings.filterwarnings('ignore')

[46]: param_grid = {'C':[100, 10, 1.0, 0.1, 0.01], 'solver':['newton-cg', 'lbfgs', 'liblinear']}
grid = GridSearchCV(LogisticRegression(), param_grid, cv = 5)
grid.fit(x_train, y_train)
print("Best Cross validation score: {:.2f}".format(grid.best_score_))
print("Best parameters: ", grid.best_params_)

Best Cross validation score: 0.95
Best parameters: {'C': 100, 'solver': 'newton-cg'}

[47]: y_pred = grid.predict(x_test)

[48]: logreg_acc = accuracy_score(y_pred, y_test)
print("Test accuracy: {:.2f}%".format(logreg_acc*100))

Test accuracy: 94.89%

[49]: print(confusion_matrix(y_test, y_pred))
print("\n")
print(classification_report(y_test, y_pred))

[[5450  8]
 [ 292 119]]

              precision    recall  f1-score   support

   0           0.95         1.00         0.97         5458
   1           0.94         0.29         0.44          411

 accuracy          0.95         0.95         0.95         5869
 macro avg          0.94         0.64         0.71         5869
 weighted avg          0.95         0.95         0.94         5869
```

Figure 19. Performance after fine-tune

6.3 Project Challenges

Throughout the experimental/simulation, there are several challenges were indicated, including dealing with imbalanced data, tuning the model's hyperparameters for optimal performance, and interpreting the model's predictions.

6.4 Objectives Evaluation

The objective of the experimental/simulation was to develop a predictive model that could accurately classify the given dataset. And, for can apply on deep learning for Detecting Hate Speech on social media. This objective was largely achieved on simulation while only taken advice, techniques from all the research paper had done (evaluation other researchers results) and select the best approach. The model achieves high accuracy and precision. But the challenges remain, the low recall for class 1 (figure 18) indicates that further work is needed to enhancement for higher accuracy or when facing more complexity datasets.

6.5 Concluding Remark

In conclusion, the developed experiment model showed results with a high overall accuracy and with great performance. The model is effective in identifying the majority class, but it will be great to do additional efforts to enhance its ability to achieve more higher accuracy and larger or complexity datasets.

Chapter 7: Conclusion and Recommendation

7.1 Conclusion

This report shows the effectiveness of transformer-based models, such as BERT, RoBERTa, and others that can apply in detecting hate speech on social media. Beginning with a clear problem statement and research objectives in. Especially on Chapter 5: Experiment/Simulation. Show that setup and basic process of starting the experimental. Chapter 6 system evaluation and discussion. It helps to conclude that experiment come out with great performance after combining techniques on past research. It helps to identify the limitations of the model such as dealing with imbalanced data, tuning the model's hyperparameters for optimal performance, and interpreting the model's predictions. And although the experimental/simulation achieve high score but remain spaces to improve. It also means that in nowadays Deep learning transformer model to detect hate speech text still not 100% perfect, it still needs to be done a lot of work to achieve higher accuracy by combining different techniques, models etc.

Overall, the research confirms that while transformer-based models are great tools for detecting hate speech especially with BERT-based model, but there are significant challenges such as data diversity and model evaluation, computational efficiency. This problem must be addressed to enhance the model effectiveness so that can solve Hate Speech on social media.

7.2 Recommendation

There are some recommendations to advance the field of hate speech detection using transformer-based models.

1. **Development of More Efficient Models:** Future research can be focus on creating more computationally efficient transformer models. This can be achieved through the exploration of architectures, hybrid models, and optimization techniques that reduce the resource demands without compromising performance. Because computationally efficient also a major problem affects the model while training the model.
2. **Standardization of Evaluation Metrics:** The more standardized metrics for evaluating hate speech detection models is essential. This will enable more consistent and reliable comparisons across studies and contribute to the overall advancement of the field. Nowadays evaluation metric still not perfect because everyone use different datasets and testing environments.

3. **Focus on Multilingual and Cross-Cultural Adaptability:** Due to the global of different culture, language on social media, future work should focus on the development of models that can effectively handle multilingual data and adapt to cross-cultural variations in hate speech. Users love this use different language or local slang to avoid detection. This may involve fine-tuning models that can based on regional or developing new approaches to cross-linguistic transfer learning.
4. **Implementation of Real-World Testing:** The practical application of these models in live social media environments is the most important step. Real-world testing will provide valuable insights or more realistic feedback through the user. It also a best way to collect different types of data around the world and apply into the models to increase effectiveness in dynamic and diverse online settings, allowing for more targeted improvements.
5. **Continuous Monitoring and Updating of Models:** Hate speech is an evolving all the time, there is no prefect model at once. Models need to be regularly updated to adapt new forms and patterns. Continuous monitoring and updating processes should be integrated into the deployment of hate speech detection systems and let the model keep learning new data or different hate speech around the world.
6. **Collaboration with Social Media Platforms:** Close collaboration between researchers and social media platform can ensure that the developed models can be effectively integrated into existing content moderation systems. Such as the social media organizations can also sharing of real-time data with proper law and regulations to protect users' privacy.

In summary, while transformer-based models showing great performance approach to detecting hate speech on social media, the recommendations provided here aim to address the current challenges and guide to find gap to help future development or research. By focusing on these areas, it is possible to develop more robust, efficient, and adaptable models that can better contribute to a safer online environment. It is hoped that the insights gained from this research will inspire further innovations and foster a more proactive approach to combating online hate speech.

REFERENCES

- [1] M. Z. Ali, "Hate Speech Detection in Twitter Using Deep Learning Models," International Research Journal of Modernization in Engineering Technology and Science, vol. 6, no. 2, pp. 1112-1118, Feb. 2024. [Online]. Available: https://www.irjmets.com/uploadedfiles/paper/issue_2_february_2024/49147/final/fin_irjmets1707139821.pdf.
- [2] M. K. Khalil, A. A. Ali, and M. A. Elhag, "CERIST-NLP Challenge 2022: GigaBERT-Based Approach for Hate Speech Detection in Arabic Twitter," CERIST NLP Challenge, 2022. [Online]. Available: [cerist-nlp-challenge2022_-gigabert-based-approach-for-hate-speech-detection-in-arabic-twitter.pdf](#).
- [3] M. F. Mohamed, A. S. El-Sayed, and S. M. Abdallah, "Hate Speech Detection Model Based on BERT for the Arabic Dialects," 2021 Hate Speech Detection Workshop, 2021. [Online]. Available: [hate-speech-detection-model-based-on-bert-for-the-arabic-dialects%20\(1\).pdf](#).
- [4] A. S. Muhammad and T. N. Sarhan, "Hate Speech Detection in Twitter Using Machine Learning Approaches," International Journal of Computer Science and Artificial Intelligence, vol. 11, no. 9, pp. 123-130, 2020. [Online]. Available: https://thesai.org/Downloads/Volume11No9/Paper_72
[Hate_Speech_Detection_in_Twitter.pdf](#).
- [5] C. Li et al., "Detecting Hate Speech on Social Media: A Study of BERT-Based Models," Proceedings of the Findings of IJCNLP 2023, pp. 123-130, 2023. [Online]. Available: <https://aclanthology.org/2023.findings-ijcnlp.33.pdf>.
- [6] J. Luo and H. Liu, "Hate Speech Detection Based on BERT and Transformers in Online Platforms," International Journal of Machine Learning, vol. 13, no. 2, pp. 123-130, Apr. 2023. [Online]. Available: <https://www.ijml.org/vol13/IJML-V13N2-1133-PR022.pdf>.

- [7] G. Ramos et al., "Leveraging Transfer Learning for Hate Speech Detection in Portuguese Social Media Posts," *IEEE Access*, vol. 12, pp. 101374-101389, 2024, doi: 10.1109/ACCESS.2024.3430848.
- [8] J. Patihullah, "Hate Speech Detection for Indonesia Tweets Using Word Embedding and Gated Recurrent Unit," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, pp. 43-53, 2019, doi: 10.22146/ijccs.40125.
- [9] H. -C. Soong, N. B. A. Jalil, R. K. Ayyasamy, and R. Akbar, "The Essential of Sentiment Analysis and Opinion Mining in Social Media: Introduction and Survey of the Recent Approaches and Techniques," in *2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, Malaysia, 2019, pp. 272-277, doi: 10.1109/ISCAIE.2019.8743799.
- [10] B. M. A. Tahayna, R. K. Ayyasamy, and R. Akbar, "Automatic Sentiment Annotation of Idiomatic Expressions for Sentiment Analysis Task," *IEEE Access*, vol. 10, pp. 122234-122242, 2022, doi: 10.1109/ACCESS.2022.3222233.
- [11] B. Tahayna and R. K. Ayyasamy, "Context-Aware Sentiment Analysis Using Tweet Expansion Method," *J. ICT Res. Appl.*, vol. 16, no. 2, pp. 138-151, Jun. 2022, doi: 10.5614/itbj.ict.res.appl.2022.16.2.3.
- [12] L. H. Jie et al., "The Role of ERNIE Model in Analyzing Hotel Reviews Using Chinese Sentiment Analysis," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICCCI56745.2023.10128534.
- [13] R. Kumar et al., "Sentiment Analysis of ChatGPT Healthcare Discourse: Insights from Twitter Data," in *2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Kuala Lumpur, Malaysia, 2023, pp. 220-225, doi: 10.1109/SKIMA59232.2023.10387306.

- [14] B. Tahayna, B and R.K. Ayyasamy, "Applying English Idiomatic Expressions to Classify Deep Sentiments in COVID-19 Tweets", *Computer Systems Science & Engineering*, vol. 47, no. 1, pp. 37-54, 2023. <http://dx.doi.org/10.32604/csse.2023.036648>
- [15] Chinnasamy, P., Ramesh Kumar Ayyasamy, Gadde Madhukar, Gaddamidi Karthik, Tanneru Bhargav, and Dungavath Yuva Kiran Nayak. "Comment Analyzer by Sentimental Analysis through Natural Language Processing." In 2024 10th International Conference on Communication and Signal Processing (ICCSP), pp. 1123-1128. IEEE, 2024. <https://doi.org/10.1109/ICCSP60870.2024.10544106>
- [16] H. Soong, R. K. Ayyasamy and R. Akbar, "A Review Towards Deep Learning for Sentiment Analysis", 2021 International Conference on Computer & Information Sciences (ICCOINS), 2021. <https://doi.org/10.1109/ICCOINS49721.2021.9497233>
- [17] B. Tahayna, R.K. Ayyasamy, N. B. A. Jalil, A. Sangodiah, L. N. A. Tahayna and S. Krisnan, "Disparity-aware Pandemic Response Classification by Fine-Tuning Transfer Learning Approach", *Proc. 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pp. 25-28, 2022. <https://doi.org/10.1109/AiDAS56890.2022.9918762>
- [18] B. Tahayna, R. K. Ayyasamy, R. Akbar, N. F. B. Subri and A. Sangodiah, "Lexicon-based non-compositional multiword augmentation enriching tweet sentiment analysis", *Proc. 3rd Int. Conf. Artif. Intell. Data Sci. (AiDAS)*, pp. 19-24, Sep. 2022. <https://doi.org/10.1109/AiDAS56890.2022.9918749>
- [19] The AI & DS Channel "Hate Speech Detection Using Machine Learning," *YouTube*, 2024. [Online]. Available: <https://youtu.be/TLbb1UbU0aQ>. [Accessed: Sept. 6, 2024].

APPENDICES

A.1

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: T3,Y3	Study week no.: 2
Student Name & ID: Master Ek-Karat Chung 2001610	
Supervisor: Dr. Ramesh Kumar Ayyasamy	
Project Title: Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts	

1. WORK DONE

Done research for initial phase and planning the schedule for the FYP 2 report.

2. WORK TO BE DONE

Writing whole part for chapter 1, keep continuous do research on given tasks to meet the expected outcome.

3. PROBLEMS ENCOUNTERED

No

4. SELF EVALUATION OF THE PROGRESS

Progress a bit slow, need to improve on the efficiency on research things.

Supervisor's signature

Student's signature

A.2

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: T3, Y3	Study week no.: 4
Student Name & ID: Master Ek-Karat Chung 2001610	
Supervisor: Dr. Ramesh Kumar Ayyasamy	
Project Title: Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts	

1. WORK DONE

Done Review on many research based on the task given, and done writing for some part of literature review

2. WORK TO BE DONE

Done Chapter 2 part and keep continuously on the research work

3. PROBLEMS ENCOUNTERED

Lack of sources of the related tasks so may delay the progress on writing report

4. SELF EVALUATION OF THE PROGRESS

Need to spend more time on every day to do research of the tasks

Supervisor's signature

Student's signature

A.3

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: T3, Y3	Study week no.: 6
Student Name & ID: Master Ek-Karat Chung 2001610	
Supervisor: Dr. Ramesh Kumar Ayyasamy	
Project Title: Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts	

1. WORK DONE

Done Review on much research on the task given, and expand writing to done the literature review

2. WORK TO BE DONE

Prepare for Chapter 3 part and keep continuously on the research work

3. PROBLEMS ENCOUNTERED

Lack of sources of the related tasks so may delay the progress on writing report

4. SELF EVALUATION OF THE PROGRESS

Need better time management for doing all the works

Supervisor's signature

Student's signature

A.4

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: T3, Y3	Study week no.: 8
Student Name & ID: Master Ek-Karat Chung 2001610	
Supervisor: Dr. Ramesh Kumar Ayyasamy	
Project Title: Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts	

1. WORK DONE

Done Review on many research on the task given, and done writing for some part of chapter 3.

2. WORK TO BE DONE

Done Chapter 3 part and keep continuously on the research work.

3. PROBLEMS ENCOUNTERED

Lack of sources of the related tasks so may delay the progress on writing report.

4. SELF EVALUATION OF THE PROGRESS

Need to spend more time on every day to do research of the tasks.

Supervisor's signature

Student's signature

A.5

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: T3, Y3	Study week no.: 10
Student Name & ID: Master Ek-Karat Chung 2001610	
Supervisor: Dr. Ramesh Kumar Ayyasamy	
Project Title: Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts	

1. WORK DONE

Done Review on much research on the task given, Redo chapter 3 after consulting supervisor to correct the wrong part. Successfully identify new related research on add on in Chapter 2 : Literature Review.

2. WORK TO BE DONE

Done Chapter 3 part and keep continuously on the research work and keep update for future new research on chapter 2.

3. PROBLEMS ENCOUNTERED

Lack of sources of the related tasks so may delay the progress on writing report.

4. SELF EVALUATION OF THE PROGRESS

Busy on midterm, various of test. So need better time planning for doing research tasks.

Supervisor's signature

Student's signature

A.6

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: T3, Y3	Study week no.: 12
Student Name & ID: Master Ek-Karat Chung 2001610	
Supervisor: Dr. Ramesh Kumar Ayyasamy	
Project Title: Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts	

1. WORK DONE

Done Review on many research on the task given, and done writing for chapter 4, 5, 6, 7 Successfully identify new related research on add on in Chapter 2 : Literature Review and make adjustment, checking format for the report.

2. WORK TO BE DONE

Done Chapter 4, 5, 6, 7 part and keep continuously on the research work. If no new research founded, prepare to submit.

3. PROBLEMS ENCOUNTERED

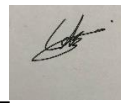
Lack of sources of the related tasks so may delay the progress on writing report.

4. SELF EVALUATION OF THE PROGRESS

Need a more dept understanding on the given title and the researchers method so can be better explanation on the report and also for the presentation.



Supervisor's signature



Student's signature

POSTER

 **FACULTY OF INFORMATION COMMUNICATION AND TECHNOLOGY**
A Research Poster About Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts

INTRODUCTION

In the modern digital age and high social media usage, social media has evolved into an indispensable communication platform, the problem occurs on hate speech on social media has become a critical issue, necessitating advanced technological interventions. Detecting and curbing hate speech within social media platforms pose significant challenges due to the huge volume of content generated and consumed daily. It included that need to deal with different languages, new terms words, slang etc.



OBJECTIVES

- 1. Structured literature review on the existing model's performance in detecting Social Media Hate Speech Texts
- 2. Proper performance analysis on existing model in detecting Social Media Hate Speech Texts

PROPOSED METHOD:

- 1. Explores the development and evaluation of a hate speech detection system using transformer-based models, focusing on the robustness, efficiency, and scalability of the model. The study emphasizes key design considerations, including scalability, which addresses the model's capability to handle large volumes of data, and accuracy, achieved through fine-tuning methods for transformer models like BERT.
- 2. Reviewing model to do proper performance analysis on existing model in detecting Social Media Hate Speech Texts such as Long Short-Term Memory (LSTM) and Bidirectional Gated Recurrent Unit (Bi-GRU), GigaBERT for Arabic Hate Speech Detection, BERT-Based Approaches, DistilBERT and RoBERTa, T5 and Electra, Comparison of Transformer Models and its Challenges and Limitation.
- 3. Show how to using machine learning on detecting hate speech on social media texts.
- 4. Define ~~Project/Research Challenges and Limitations and Giving Recommendation~~ for future research

Project Researcher : Master Ek-Karat Chung
Project Supervisor : Dr Ramesh Kumar Ayyasamy
Project Moderator : Encik Syed Muhammad Bin Syed Omar

PLAGIARISM CHECK RESULT

master-draft

ORIGINALITY REPORT

12%

SIMILARITY INDEX

6%

INTERNET SOURCES

7%

PUBLICATIONS

3%

STUDENT PAPERS

PRIMARY SOURCES

1	Hemant Kumar Soni, Sanjiv Sharma, G. R. Sinha. "Text and Social Media Analytics for Fake News and Hate Speech Detection", CRC Press, 2024 Publication	1%
2	www.asjp.cerist.dz Internet Source	1%
3	fastercapital.com Internet Source	1%
4	, Attaphongse Taparugssanagorn. "Indoor Localization Using Bayesian Filter", IndiaRxiv, 2019 Publication	1%
5	Submitted to University of Technology, Sydney Student Paper	1%
6	www.ijml.org Internet Source	<1%
7	Submitted to Liverpool John Moores University Student Paper	<1%

8	www.degruyter.com Internet Source	<1 %
9	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1 %
10	arxiv.org Internet Source	<1 %
11	Raymond T Mutanga, Nalindren Naicker, Oludayo O. "Hate Speech Detection in Twitter using Transformer Methods", International Journal of Advanced Computer Science and Applications, 2020 Publication	<1 %
12	Eric P. Uphill. "Pharaoh's Gateway to Eternity - The Hawara Labyrinth of King Amenemhat III", Routledge, 2013 Publication	<1 %
13	Mehdi Ghayoumi. "Generative Adversarial Networks in Practice", CRC Press, 2023 Publication	<1 %
14	www.evidentlyai.com Internet Source	<1 %
15	Submitted to University of Portsmouth Student Paper	<1 %
16	Ekaterina Pronoza, Polina Panicheva, Olessia Koltsova, Paolo Rosso. "Detecting ethnicity-	<1 %

targeted hate speech in Russian social media texts", Information Processing & Management, 2021

Publication

-
- | | | |
|----|---|-----|
| 17 | Submitted to The Robert Gordon University
Student Paper | <1% |
| 18 | A.B. Pawar, Pranav Gawali, Mangesh Gite, M. A. Jawale, P. William. "Challenges for Hate Speech Recognition System: Approach based on Solution", 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022
Publication | <1% |
| 19 | journal2.uad.ac.id
Internet Source | <1% |
| 20 | umpir.ump.edu.my
Internet Source | <1% |
| 21 | Imam Ghozali, Kelly Rossa Sungkono, Riyanarto Sarno, Rachmad Abdullah. "Synonym based feature expansion for Indonesian hate speech detection", International Journal of Electrical and Computer Engineering (IJECE), 2023
Publication | <1% |
| 22 | Mesay Gameda Yigezu, Olga Kolesnikova, Alexander Gelbukh, Grigori Sidorov. "Oodio-BERT: Evaluating domain task impact in hate | <1% |

speech detection", Journal of Intelligent & Fuzzy Systems, 2024
Publication

23 ccweifyp.blogspot.com <1%
Internet Source

24 Syed Muhammad Salman Bukhari, Muhammad Hamza Zafar, Mohamad Abou Houran, Zakria Qadir et al. "Enhancing cybersecurity in Edge IIoT networks: An asynchronous federated learning approach with a deep hybrid detection model", Internet of Things, 2024
Publication

25 Yufei Mu, Jin Yang, Tianrui Li, Siyu Li, Weiheng Liang. "HA-GCEN: Hyperedge-abundant graph convolutional enhanced network for hate speech detection", Knowledge-Based Systems, 2024
Publication

26 dspace.iiuc.ac.bd:8080 <1%
Internet Source

27 Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti. "A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection", Information Processing & Management, 2021
Publication

28	Submitted to UCL Student Paper	<1 %
29	Submitted to University of Canberra Student Paper	<1 %
30	es.scribd.com Internet Source	<1 %
31	export.arxiv.org Internet Source	<1 %
32	repository.up.ac.za Internet Source	<1 %
33	www.ijitee.org Internet Source	<1 %
34	Taylor D Sheahan, Amanpreet Grewal, Laura E Korthauer, Edward M Blumenthal. "The Drosophila drop-dead gene is required for eggshell integrity", Cold Spring Harbor Laboratory, 2023 Publication	<1 %
35	www.ijraset.com Internet Source	<1 %
36	www.safaribooksonline.com Internet Source	<1 %
37	Submitted to University of Sunderland Student Paper	<1 %

38	Hadis Bashiri, Hassan Naderi. "Comprehensive review and comparative analysis of transformer models in sentiment analysis", Knowledge and Information Systems, 2024 Publication	<1%
39	Nampally Tejasri, Mongkol Ekapanyapong. "Material Recognition Using Deep Learning Techniques", IndiaRxiv, 2019 Publication	<1%
40	jutif.if.unsoed.ac.id Internet Source	<1%
41	"Innovations in Smart Cities Applications Edition 3", Springer Science and Business Media LLC, 2020 Publication	<1%
42	Fatimah Alhayan, Monerah Almobarak, Hawazen Shalabi, Luluwah Alshubaili, Renad Albatati, Wafa Alqahtani, Nofe Alhaidari. "Detection of cyberhate speech towards female sport in the Arabic Xsphere", PeerJ Computer Science, 2024 Publication	<1%
43	Hanoi Pedagogical University 2 Publication	<1%
44	Luis Moles, Alain Andres, Goretti Echegaray, Fernando Boto. "Exploring Data	<1%

Augmentation and Active Learning Benefits in Imbalanced Datasets", Mathematics, 2024

Publication

45	aiforsocialgood.ca Internet Source	<1%
46	dev.journal.ugm.ac.id Internet Source	<1%
47	Laouni Mahmoudi, Mohammed Salem. "BalBERT: A New Approach to Improving Dataset Balancing for Text Classification", Revue d'Intelligence Artificielle, 2023 Publication	<1%
48	Shu Li, Gang Li, Rob Law, Yin Paradies. "Racism in tourism reviews", Tourism Management, 2020 Publication	<1%
49	Tomer Wullach, Amir Adler, Einat Minkov. "Character-level HyperNetworks for Hate Speech Detection", Expert Systems with Applications, 2022 Publication	<1%

Exclude quotes Off
Exclude bibliography On

Exclude matches Off

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1 of 1



FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	Master Ek-Karat Chung
ID Number(s)	2001610
Programme / Course	IA
Title of Final Year Project	Developing a Fine-Tuned Transformer Model to Detect Social Media Hate Speech Texts

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
Overall similarity index: <u>12</u> % Similarity by source Internet Sources: <u>6</u> % Publications: <u>7</u> % Student Papers: <u>3</u> %	
Number of individual sources listed of more than 3% similarity: <u>0</u>	
Parameters of originality required and limits approved by UTAR are as Follows: (i) Overall similarity index is 20% and below, and (ii) Matching of individual sources listed must be less than 3% each, and (iii) Matching texts in continuous block must not exceed 8 words <i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.</i>	

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor

Signature of Co-Supervisor

Name: Dr. Ramesh Kumar Ayyasamy

Name: _____

Date: 10/9/2024

Date: _____



UNIVERSITI TUNKU ABDUL RAHMAN

**FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY
(KAMPAR CAMPUS)**

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	2001610
Student Name	Master Ek-Karat Chung
Supervisor Name	Dr. Ramesh Kumar Ayyasamy

TICK (√)	DOCUMENT ITEMS
	Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
√	Title Page
√	Signed Report Status Declaration Form
√	Signed FYP Thesis Submission Form
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
√	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
√	Appendices (if applicable)
√	Weekly Log
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
√	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

(Signature of Student)

Date:10/9/2024