

**DEVELOPMENT OF LEADS QUALIFICATION MODEL
FOR THE CHATBOT OF SWISS GERMAN UNIVERSITY**

By

Sakinah Arifin
12302005

BACHELOR'S DEGREE
in

INFORMATION TECHNOLOGY
ENGINEERING AND INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

Revision after the Thesis Defense on 8 July 2024

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.



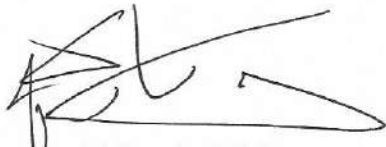
Sakinah Arifin

Student

22-7-2024

Date

Approved by:



Dr. Maulahikmah Galinium S.Kom M.Sc

Thesis Advisor

22.7.2024.

Date



Alva Erwin, ST., M.Sc., MTI

Thesis Co-Advisor

22-7-2024

Date



Dr. Maulahikmah Galinium S.Kom M.Sc

Dean

22.7.2024.

Date

Sakinah Arifin

ABSTRACT

DEVELOPMENT OF LEADS QUALIFICATION MODEL FOR THE CHATBOT OF SWISS GERMAN UNIVERSITY

By

Sakinah Arifin

Dr. Maulahikmah Galinium S.Kom M.Sc, Advisor

Alva Erwin, ST., M.Sc., MTI, Co-Advisor

SWISS GERMAN UNIVERSITY

This study investigated the potential of deep learning models to automate lead qualification in the context of university admissions. The focus was on Swiss German University's (SGU) current manual system for processing WhatsApp conversations with prospective students. While manual review ensures quality control, it suffers from limitations such as time constraints, subjectivity leading to inconsistencies and the risk of human error. This research explored developing and evaluating deep learning models for lead qualification to address these limitations. Convolutional Neural Networks (CNNs) and a hybrid CNN-LSTM architecture were investigated for their effectiveness in classifying leads based on their level of interest and stage in the enrollment process. The models were trained and evaluated on a dataset of WhatsApp conversations in Bahasa Indonesia. The result of this study is that a CNN-LSTM model achieved a promising accuracy of 80.85%. This suggests significant potential for automation and improved efficiency compared to the current manual approach. Furthermore, by automating lead classification, the model minimizes human error and ensures consistent lead categorization, aiming to streamline the university's student recruitment process.

Keywords: Deep Learning, CNN, CNN-LSTM, Leads Qualification, WhatsApp Chat Conversation.

© Copyright 2024
by Sakinah Arifin
All rights reserved

DEDICATION

I dedicate this work for my parents, my husband and my beloved son.

ACKNOWLEDGEMENTS

I would like to thank Allah SWT for the health, ability and opportunity full of blessings to complete my thesis and obtain this degree.

My gratitude to my parents for their prayers, blessings, love and motivation who always accompanied me in writing this research.

My beloved husband, Muhammad Kemal Ichwandi, thank you for your endless support and sacrifices during the process of completing this thesis. Your love and understanding helped me get through difficult times and focus on achieving my goals. To my son Ibrahim Khairullah Ichwandi, thank you for your patience and understanding during the process of completing this thesis.

I wish to thanks Dr. Maulahikmah Galinium, S.Kom, M.Sc as my advisor and Alva Erwin, ST., M.Sc., MTI as my co-advisors who have given their time, energy and thoughts to guide me in completing this thesis. They are an extraordinary and inspiring lecturer for me.

Lastly, I would also like to thank my dearest sister who has given her time and energy to me in completing this thesis. The greatest gratitude in my life has you as a friend, sister, and funny onty for Ibrahim.

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGEMENTS.....	6
TABLE OF CONTENTS.....	7
LIST OF FIGURES	10
LIST OF TABLES	12
CHAPTER 1 - INTRODUCTION.....	13
1.1 Background.....	13
1.2 Research Problems	14
1.3 Research Objectives	14
1.4 Scope of Study.....	15
1.5 Significance of Study.....	17
1.6 Research Questions.....	18
1.7 Hypothesis	18
1.8 Research Limitation.....	18
CHAPTER 2 - LITERATURE REVIEW	19
2.1 Chatbot.....	19
2.2 Lead Qualification	19
2.3 Deep Learning for Text Analysis	20
2.3.1 Convolutional Neural Networks (CNN)	20
2.3.2 Long Short-Term Memory (LSTM)	22
2.3.3 Deep Learning Algorithm Comparison	22
2.4 Related Works	24
2.4.1 Convolutional Neural Networks for Indonesian Aspect-Based Sentiment Analysis Tourism Review	24
2.4.2 Sentiment Analysis Based on Deep Learning – A Comparative Study.....	25

2.4.3	A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis	26
2.4.4	Text Based Sentiment Analysis Using LSTM	26
2.4.5	Sentiment Analysis of Restaurant Reviews Using Combined CNN-LSTM	27
CHAPTER 3 – RESEARCH METHODS		31
3.1	Data Gathering	31
3.2	Data Labelling	32
3.3	Data Preprocessing	35
3.3.1	Case folding	35
3.3.2	Normalization	35
3.3.3	Tokenizing	36
3.3.4	Filtering	37
3.3.5	Stemming	37
3.4	Model Designing	38
3.4.1	Dataset Splitting (Train and Test)	38
3.4.2	Tokenizing and One-Hot Encoding	38
3.4.3	Model Building and Training	39
3.5	Testing and Evaluation	43
CHAPTER 4 – RESULTS AND DISCUSSIONS		44
4.1	Result of Data Gathering	44
4.2	Data Labeling	44
4.3	Data Preprocessing	46
4.3.1	Case Folding	46
4.3.2	Normalization	47
4.3.3	Tokenizing	48
4.3.4	Filtering (Stopword Removal)	49
4.3.5	Stemming	50
4.4	CNN Model Designing	51
4.4.1	Split Dataset to Train and Test	51
4.4.2	One Hot Encoding and Tokenizing	52
4.4.3	Model Building	54
4.4.4	Model Training	57
4.5	Testing and Evaluation	58
4.5.1	Scenario I	58
4.5.2	Scenario II	65

CHAPTER 5 – CONCLUSIONS AND RECCOMENDATIONS	75
5.1 Conclusions	75
5.2 Recommendations	75
GLOSSARY	77
REFERENCES	82

LIST OF FIGURES

Figures	Page
Figure 1.1: Lead to Customer Journey in SGU	15
Figure 2.1: Text Classification Using CNN (Soni et al., 2023)	20
Figure 3.1: Development of Lead Qualification Model	31
Figure 3.2: Lead to Customer Journey	32
Figure 3.3: Architecture Diagram for CNN Model 4	40
Figure 3.4: Architecture Diagram for CNN-LSTM Model 7	42
Figure 4.1: History Chat from SGU's WhatsApp	44
Figure 4.2: Chat Classification based on Characteristics	45
Figure 4.3: Chat distribution based on lead qualification	45
Figure 4.4: Case Folding Program Code	47
Figure 4.5: Normalization Program Code	48
Figure 4.6: Tokenizing Program Code	48
Figure 4.7: Filtering Program Code	49
Figure 4.8: Stemming Program Code	50
Figure 4.9: Program Code to Divide Training Data and Test with 60:40 Ratio	51
Figure 4.10: Detail Data Sharing	52
Figure 4.11: Code for One Hot Encoding	53
Figure 4.12: Result from LabelEncoder Function	53
Figure 4.13: Result from One Hot Encoding Function	53
Figure 4.14: Code for Tokenization	54
Figure 4.15: Code for CNN Model	56
Figure 4.16: Code for CNN-LSTM Model	56
Figure 4.17: Code for Model Training	57
Figure 4.18: Loss and Accuracy Graph for Model 1	59
Figure 4.19: Loss and Accuracy Graph for Model 2	59
Figure 4.20: Loss and Accuracy Graph for Model 3	60
Figure 4.21: Loss and Accuracy Graph for Model 4	60
Figure 4.22: Confusion Matrix Model 1	61

Figure 4.23: Confusion Matrix Model 2	61
Figure 4.24: Confusion Matrix Model 3	62
Figure 4.25: Confusion Matrix Model 4	62
Figure 4.26: Loss and Accuracy Graph for Model 9	63
Figure 4.27: Loss and Accuracy Graph for Model 10	63
Figure 4.28: Loss and Accuracy Graph for Model 11	64
Figure 4.29: Confusion Matrix Model 9	64
Figure 4.30: Confusion Matrix Model 10	65
Figure 4.31: Confusion Matrix Model 11	65
Figure 4.32: Loss and Accuracy Graph for Model 5	66
Figure 4.33: Loss and Accuracy Graph for Model 6	67
Figure 4.34: Loss and Accuracy Graph for Model 7	67
Figure 4.35: Loss and Accuracy Graph for Model 8	67
Figure 4.36: Confusion Matrix Model 5	68
Figure 4.37: Confusion Matrix Model 6	68
Figure 4.38: Confusion Matrix Model 7	69
Figure 4.39: Confusion Matrix Model 8	69
Figure 4.40: Loss and Accuracy Graph for Model 12	70
Figure 4.41: Loss and Accuracy Graph for Model 13	71
Figure 4.42: Loss and Accuracy Graph for Model 14	71
Figure 4.43: Confusion Matrix Model 12	71
Figure 4.44: Confusion Matrix Model 13	72
Figure 4.45: Confusion Matrix Model 14	72
Figure 4.46: Live Chat	74

LIST OF TABLES

Table	Page
Table 2.1: Deep Learning Algorithm Comparison	22
Table 2.2: Comparison of Related Works.....	28
Table 3.1: Keyword for Data Labelling	33
Table 3.2: Example of Case Folding.....	35
Table 3.3: Example of Normalization.....	36
Table 3.4: Example of Tokenizing.....	36
Table 3.5: Example of Filtering	37
Table 3.6: Example of Stemming	37
Table 3.7: Models with Different Data Split.....	38
Table 3.8: Models for Lead Qualification with Different Parameter.....	43
Table 4.1: Result of Case Folding.....	47
Table 4.2: Result of Normalization.....	48
Table 4.3: Result of Tokenizing.....	48
Table 4.4: Result of Filtering	49
Table 4.5: Result of Stemming	50
Table 4.6: Results from Scenario I for Model 1-4 (CNN).....	58
Table 4.7: Results from Scenario I for Model 9-11 (CNN).....	63
Table 4.8: Results from Scenario II for Model 5-8 (CNN-LSTM).....	66
Table 4.9: Results from Scenario II for Model 9-11 (CNN).....	70
Table 4.10: Results from Model 1-14	72