

**OPTIMIZATION OF NAÏVE BAYES USING LEVENSHTEIN DISTANCE  
FOR TYPOGRAPHICAL ERROR CORRECTION IN SENTIMENT  
ANALYSIS**

**THESIS**



By:  
**Aziz Yogo Utomo**  
**18.61.0149**

**BACHELOR OF INFORMATICS  
FACULTY OF COMPUTER SCIENCE  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA  
2022**

**OPTIMIZATION OF NAÏVE BAYES USING LEVENSHTTEIN DISTANCE  
FOR TYPOGRAPHICAL ERROR CORRECTION IN SENTIMENT  
ANALYSIS**

**THESIS**

to fulfil the requirements for a Bachelor's degree  
in the Informatics study program



By:  
**Aziz Yogo Utomo**  
**18.61.0149**

**BACHELOR OF INFORMATICS  
FACULTY OF COMPUTER SCIENCE  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA**

**2022**

**APPROVAL**

**THESIS**

**OPTIMIZATION OF NAÏVE BAYES USING LEVENSHTTEIN  
DISTANCE FOR TYPOGRAPHICAL ERROR CORRECTION IN  
SENTIMENT ANALYSIS**

prepared and arranged by:

**Aziz Yogo Utomo**

**18.61.0149**

has been approved by undergraduate thesis supervisor  
on 27 January 2022

**Supervisor,**

**Anna Baita, M.Kom.**  
**NIK. 190302290**

**VALIDATION**

**THESIS**

**OPTIMIZATION OF NAÏVE BAYES USING LEVENSHTTEIN  
DISTANCE FOR TYPOGRAPHICAL ERROR CORRECTION IN  
SENTIMENT ANALYSIS**

prepared and arranged by:

**Aziz Yogo Utomo**

**18.61.0149**

has been maintained by examiners

on 23 February 2022

**The Examiners**

**Examiner**

**Signature**

**Yuli Astuti, M.Kom.**  
**NIK. 190302146**

\_\_\_\_\_

**Supriatin, M.Kom.**  
**NIK. 190302239**

\_\_\_\_\_

**Anna Baita, M.Kom.**  
**NIK. 190302290**

\_\_\_\_\_

This thesis has been accepted as one of  
the requirements for obtaining a Bachelor of Computer degree  
on 26 February 2022

**DEAN OF FACULTY OF COMPUTER SCIENCE**

**Hanif Al Fatta, S.Kom., M.Kom.**  
**NIK. 190302096**

## DECLARATION

I, the undersigned below, state that this thesis is my work (ORIGINAL). The contents of this thesis have never been applied by any other person to receive an academic degree at a particular educational institution. As far as I know, there are no works or thoughts written and/or published by anyone, except those in writing listed in this manuscript and mentioned in the reference list.

Anything that applies to the manuscripts and works that have been made is my responsibility.

Yogyakarta, 24 February 2022

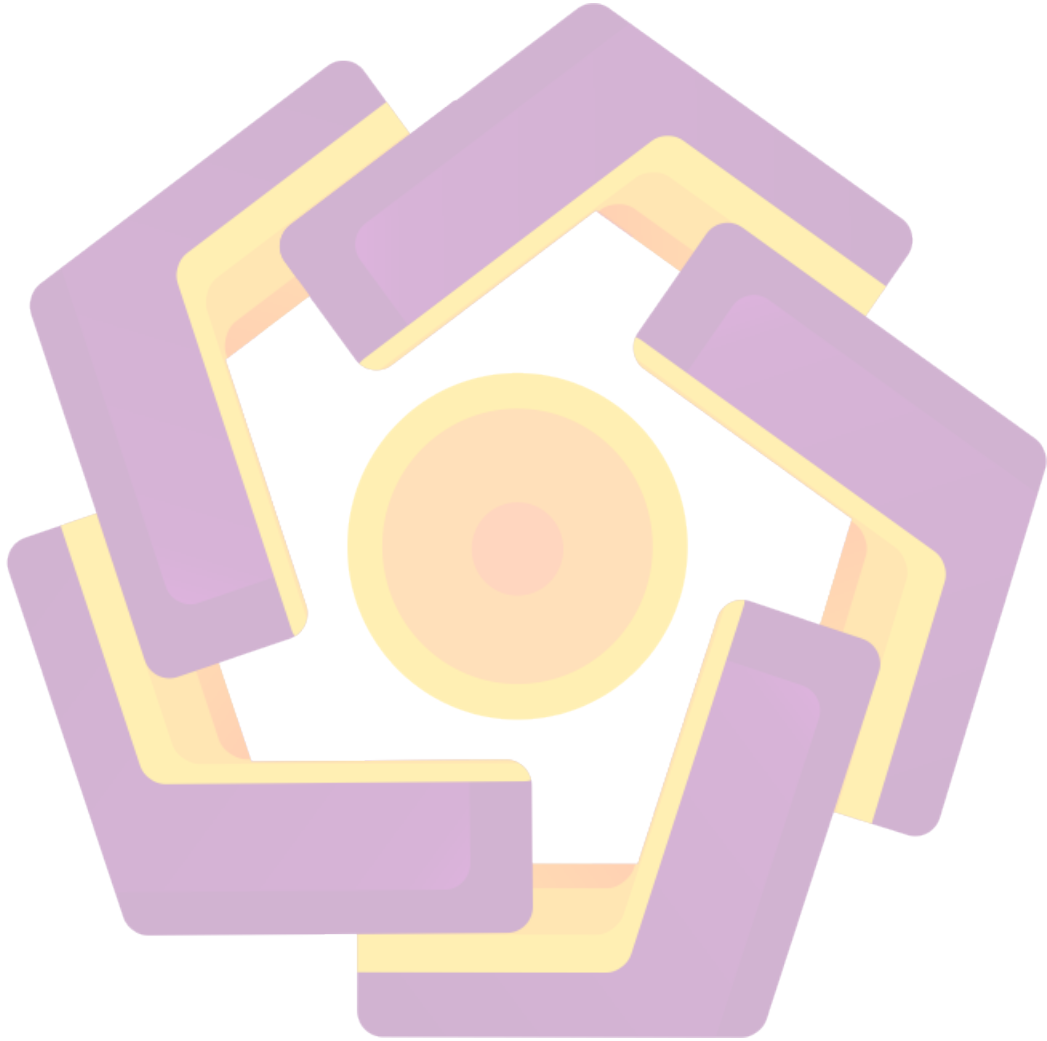


AZIZ YOGO UTOMO

18.61.0149

## MOTTO

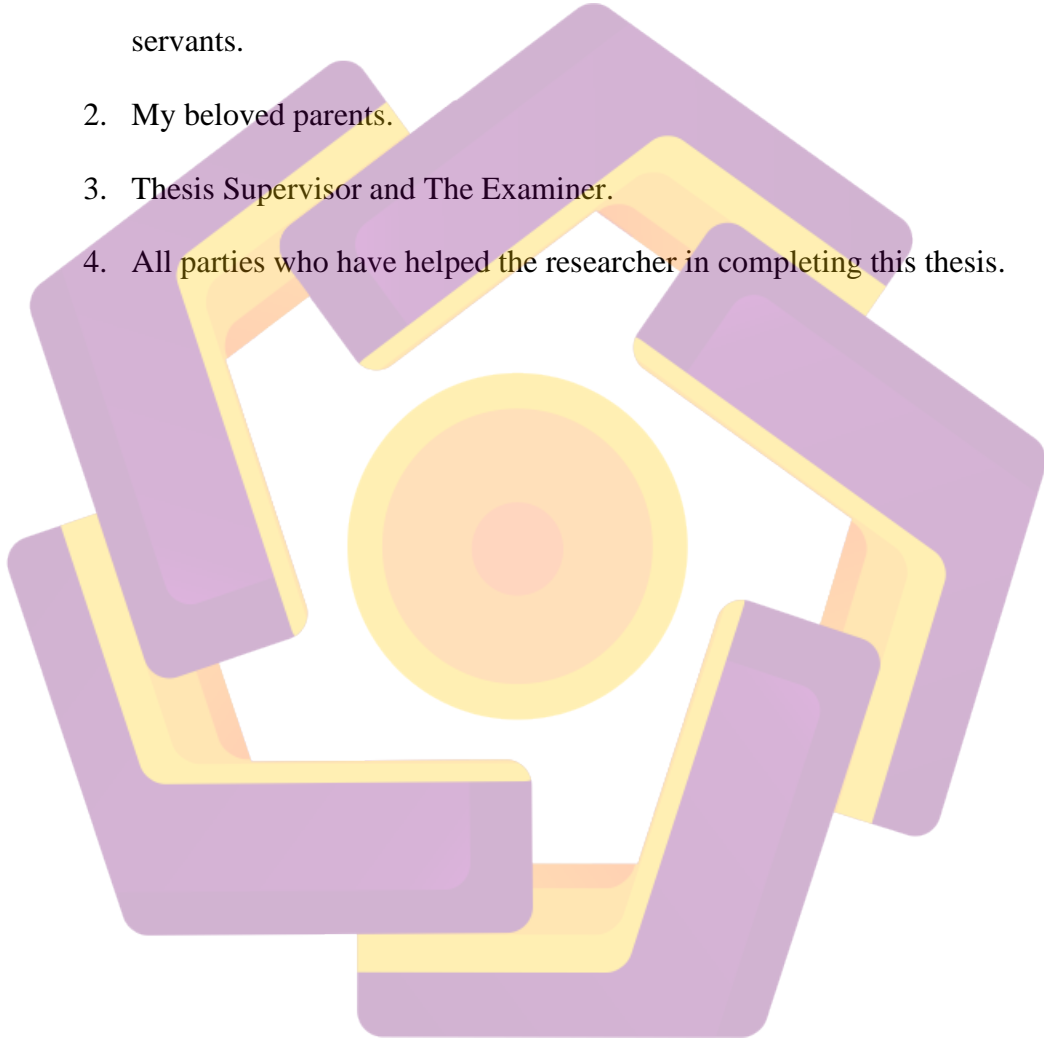
*“Break your limits.”*



## DEDICATION

Praise be to Allah SWT with His blessing; this thesis can be written and completed correctly. With this, the researcher will dedicate this thesis to:

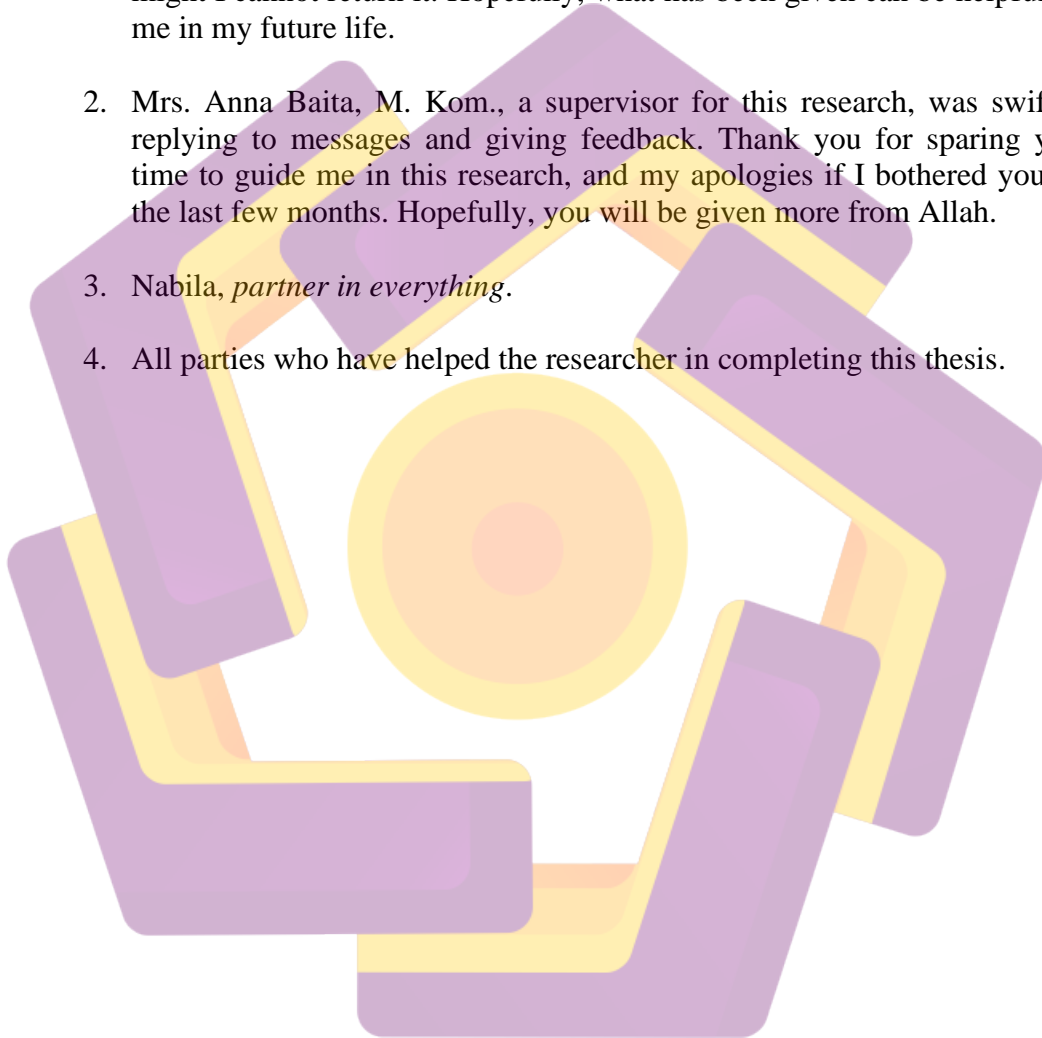
1. Allah S.W.T who always gives love and affection and guidance for His servants.
2. My beloved parents.
3. Thesis Supervisor and The Examiner.
4. All parties who have helped the researcher in completing this thesis.



## ACKNOWLEDGEMENTS

Praise and gratitude I pray to Allah SWT, Lord of the worlds, who controls all things on earth and in the sky, for all the abundance of His graces, favors, and gifts. Only to Him do I surrender and ask for help, and only to Him is the place for me to complain during this research. Not left behind, I also want to say thank you to:

1. Big thanks to my parents, who have provided all kinds of support that might I cannot return it. Hopefully, what has been given can be helpful for me in my future life.
2. Mrs. Anna Baita, M. Kom., a supervisor for this research, was swift in replying to messages and giving feedback. Thank you for sparing your time to guide me in this research, and my apologies if I bothered you for the last few months. Hopefully, you will be given more from Allah.
3. Nabila, *partner in everything*.
4. All parties who have helped the researcher in completing this thesis.





# TABLE OF CONTENT

<b>TITLE</b> .....	<b>I</b>
<b>COVER</b> .....	<b>I</b>
<b>APPROVAL</b> .....	<b>II</b>
<b>VALIDATION</b> .....	<b>III</b>
<b>DECLARATION</b> .....	<b>IV</b>
<b>MOTTO</b> .....	<b>V</b>
<b>DEDICATION</b> .....	<b>VI</b>
<b>ACKNOWLEDGEMENTS</b> .....	<b>VII</b>
<b>LIST OF TABLES</b> .....	<b>X</b>
<b>LIST OF FIGURES</b> .....	<b>XI</b>
<b>ABSTRACT</b> .....	<b>XII</b>
<b>CHAPTER I INTRODUCTION</b> .....	<b>1</b>
1.1 BACKGROUND .....	1
1.2 PROBLEMS .....	3
1.3 PROBLEMS LIMITATION .....	3
1.4 RESEARCH PURPOSE .....	4
1.5 RESEARCH BENEFITS .....	4
1.6 RESEARCH METHODOLOGY .....	4
1.6.1 <i>Literature Review</i> .....	4
1.6.2 <i>Problems Analysis</i> .....	5
1.6.3 <i>System Design</i> .....	5
1.6.4 <i>Implementation</i> .....	5
1.6.5 <i>Testing</i> .....	5
1.6.6 <i>Report Formulation</i> .....	6
1.7 SYSTEMATICS WRITING .....	7
<b>CHAPTER II RELATED WORKS</b> .....	<b>8</b>
2.1 LITERATURE REVIEW .....	8
2.2 THEORETICAL BASIS .....	12

2.2.1	<i>Text Mining</i> .....	12
2.2.2	<i>Sentiment Analysis</i> .....	12
2.2.3	<i>Preprocessing</i> .....	12
2.2.4	<i>Levenshtein Distance</i> .....	13
2.2.5	<i>Bag of Words</i> .....	17
2.2.6	<i>Naïve Bayes</i> .....	18
2.2.7	<i>Laplace Smoothing</i> .....	19
2.2.8	<i>Cross Validation</i> .....	19
2.2.9	<i>Confusion Matrix</i> .....	20
2.2.10	<i>Python</i> .....	21
<b>CHAPTER III RESEARCH METHODOLOGY .....</b>		<b>23</b>
3.1	<b>RESEARCH TOOLS AND MATERIALS</b> .....	23
3.2	<b>RESEARCH FLOW</b> .....	24
3.2.1	<i>Data Collection</i> .....	24
3.2.2	<i>Proposed Model or Method</i> .....	24
3.2.3	<i>Experiment and Testing</i> .....	28
3.2.4	<i>Evaluation</i> .....	29
<b>CHAPTER IV IMPLEMENTATION AND DISCUSSION .....</b>		<b>30</b>
4.1	<b>IMPLEMENTATION</b> .....	30
4.1.1	<i>Data Collection</i> .....	30
4.1.2	<i>Preprocessing</i> .....	35
4.1.3	<i>Modelling</i> .....	39
4.1.4	<i>Evaluation</i> .....	39
4.2	<b>DISCUSSION</b> .....	41
<b>CHAPTER V CONCLUSION .....</b>		<b>45</b>
5.1	<b>CONCLUSION</b> .....	45
5.2	<b>FUTURE WORKS</b> .....	45
<b>REFERENCES.....</b>		<b>46</b>

## LIST OF TABLES

<b>Table 1.1 Dataset</b> .....	6
<b>Table 1.2 SpellChecker Word Frequency List</b> .....	6
<b>Table 2.1 Previous research</b> .....	9
<b>Table 2.2 Previous Research Framework</b> .....	11
<b>Table 2.3 Matrix (1,1)</b> .....	14
<b>Table 2.4 input Matrix (1,1)</b> .....	14
<b>Table 2.5 Matrix (2,1)</b> .....	15
<b>Table 2.6 input Matrix (2,1)</b> .....	15
<b>Table 2.7 Matrix (3,1)</b> .....	15
<b>Table 2.8 input Matrix (3,1)</b> .....	16
<b>Table 2.9 Matrix (1,2)</b> .....	16
<b>Table 2.10 input Matrix (1,2)</b> .....	16
<b>Table 2.11 Result Matrix</b> .....	17
<b>Table 2.12 Bag of Word Table</b> .....	18
<b>Table 3.1 Software Version</b> .....	23
<b>Table 3.2 Computer Specificaion</b> .....	24
<b>Table 3.3 Example Dataset</b> .....	25
<b>Table 3.4 Case-folding and Regex</b> .....	26
<b>Table 3.5 Tokenization, Stopword Removal, Stemming</b> .....	26
<b>Table 3.6 Bag of Words Result</b> .....	26
<b>Table 3.7 Prior Probability</b> .....	27
<b>Table 3.8 Likelihood without typo correction</b> .....	27
<b>Table 3.9 Likelihood with typo correction</b> .....	27
<b>Table 3.10 Posterior Probability without Spell Correction</b> .....	27
<b>Table 3.11 Posterior Probability with Spell Correction</b> .....	27
<b>Table 3.12 Likelihood Document 5 with smoothing</b> .....	28
<b>Table 3.13 Posterior Probability with Smoothings</b> .....	28
<b>Table 3.14 Posterior Probability with Smoothing and Spell Correction</b> .....	28
<b>Table 4.1 Confusion Matrix Result</b> .....	42
<b>Table 4.2 Cross Validation Result</b> .....	43
<b>Table 4.3 Word Frequency List Accuracy Average</b> .....	44

## LIST OF FIGURES

<b>Figure 2.1 Text Mining Process</b> .....	12
<b>Figure 2.2 K-Fold Illustration</b> .....	20
<b>Figure 2.3 Confusion matrix</b> .....	20
<b>Figure 3.1 Proposed Model’s Framework</b> .....	25
<b>Figure 4.1 Code Import Dataset</b> .....	30
<b>Figure 4.2 Code Data Collection</b> .....	30
<b>Figure 4.3 Code Randomize Data</b> .....	31
<b>Figure 4.4 Code Swap Columns</b> .....	31
<b>Figure 4.5 Code Convert Class</b> .....	32
<b>Figure 4.6 Code Remove Neutral Class</b> .....	32
<b>Figure 4.7 Code Re-Convert</b> .....	33
<b>Figure 4.8 Code Convert Class</b> .....	33
<b>Figure 4.9 Code Make a new File</b> .....	34
<b>Figure 4.10 Code Top 5 Common Words</b> .....	34
<b>Figure 4.11 Code Sampling</b> .....	34
<b>Figure 4.12 Code Case-folding, Regex</b> .....	35
<b>Figure 4.13 Code Typographic Correction</b> .....	36
<b>Figure 4.14 Code Tokenization</b> .....	36
<b>Figure 4.15 Code Stopwords Removal</b> .....	37
<b>Figure 4.16 Code Stemming</b> .....	38
<b>Figure 4.17 Code Bag of Words</b> .....	39
<b>Figure 4.18 Code Split the Dataset</b> .....	39
<b>Figure 4.19 Code Modeling</b> .....	39
<b>Figure 4.20 Code Confusion Matrix</b> .....	40
<b>Figure 4.21 Code Accuracy Score</b> .....	40
<b>Figure 4.22 Code Classification Report</b> .....	40
<b>Figure 4.23 Code Cross Validation</b> .....	41

## ABSTRACT

Now we live in an era with a tremendous amount of unstructured data such as text data. Naïve Bayes is an algorithm that is well-performed for dealing with text data. In processing text data, there are several problems, such as the vast amount and the data dimension. Text data derived from human fingers allows typographical errors in writing; this typographical error becomes another problem because it will make the data dimension bigger and change the semantics of the word itself. In addition, typographical errors liable the calculation of the Naïve Bayes Likelihood to be 0 and affect the performance of the model. Levenshtein Distance is one method to correct typographical errors. This research indicates that the Levenshtein Distance has succeeded in increasing the performance of the Naïve Bayes model with optimal results using distance 2 of 5,9%.

**Keywords:** *Text Mining, Classification, Naive Bayes Classifier, Levenshtein Distance, Typographical Error Correction*

