



THE REPUBLIC OF INDONESIA DEFENSE UNIVERSITY

**SENTIMENT ANALYSIS MODEL USING BI-LSTM AND
LATENT DIRICHLET ALLOCATION ON YOUTUBE
COMMENTS TO SUPPORT INTELLIGENCE DATA**

**AGUNG NUGROHO
120220405003**

This Thesis was Written for the Fulfillment of the Requirement
to Earn Master's Degree in Defense Science

**SCIENCE AND DEFENSE TECHNOLOGY FACULTY
CYBER DEFENSE ENGINEERING STUDY PROGRAM**

**BOGOR
2024**



THE REPUBLIC OF INDONESIA DEFENSE UNIVERSITY

**SENTIMENT ANALYSIS MODEL USING BI-LSTM AND
LATENT DIRICHLET ALLOCATION ON YOUTUBE
COMMENTS TO SUPPORT INTELLIGENCE DATA**

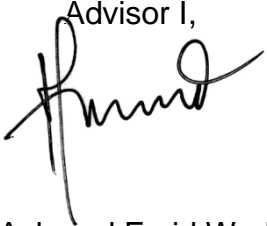


**AGUNG NUGROHO
120220405003**

This Thesis was Written for the Fulfillment of the Requirement
to Earn Master's Degree in Defense Science

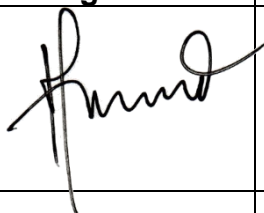




**SCIENCE AND DEFENSE TECHNOLOGY FACULTY
CYBER DEFENSE ENGINEERING STUDY PROGRAM**

**BOGOR
2024**

THESIS APPROVAL SHEET

<p>Student Name : Agung Nugroho Student Identity Number : 120220405003 Study Program : Cyber Defense Engineering Faculty : Science dan Defense Technology Thesis Title : Sentiment Analysis Model using Bi-LSTM and Latent Dirichlet Allocation on Youtube Comments to Support Intelligence Data</p>	
<p>Advisor I,  Dr. Ir. H. Achmad Farid Wajdi, M.M. Date :</p>	<p>Advisor II,  Prof. Ir. Teddy Mantoro, MSc., Ph.D., SMIEEE. Date :</p>
<p>Acknowledged by, Dean of The Faculty of Defense Science and Technology,  Prof. Dr. Ir. Muhamad Asvial, M.Eng. First Class Administrator Date :</p>	

THESIS VERIFICATION SHEET

Student Name : Agung Nugroho Student Identity Number : 120220405003 Study Program : Cyber Defense Engineering Faculty : Science dan Defense Technology Thesis Title : Sentiment Analysis Model using Bi-LSTM and Latent Dirichlet Allocation on Youtube Comments to Support Intelligence Data			
No.	Name	Signature	Date
1.	Advisor I: Dr. Ir. H. Achmad Farid Wajdi, M.M.		26/01/2024
2.	Advisor II: Prof. Ir. Teddy Mantoro, MSc., Ph.D., SMIEEE.		26/01/2024
3.	Reviewer I: Syachrul Arief, S.Si., Ph.D.		26/01/2024
4.	Reviewer II: Dr. Yosef Prihanto, S.Si., M.Si.		26/01/2024
5.	Reviewer III: Ruby Alamsyah, M.Tr.Opsla., M.Han., MCE., CIPA., CIT., CIIQA. Kolonel Laut (P) NRP. 10342/P		26/01/2024

DECLARATION OF ORIGINALITY

I hereby declare that in this thesis there is no work or part of work that has been given to obtain a bachelor's degree at any level at a university; and to the best of my knowledge, no term, phrase, sentence, paragraph, subchapter or chapter of any work has ever been written or published; except as written in this text and mentioned in the Reference List.

If in the future it is proven that there is plagiarism in this thesis, I am willing to accept sanctions in accordance with the provisions of the applicable regulations/laws.

Bogor, January , 2024



Agung Nugroho

FOREWORDS

I would like to thank the presence of Allah SWT, the completion of the preparation of the thesis entitled Sentiment Analysis Model using Bi-LSTM and Latent Dirichlet Allocation on YouTube Comments to Support Intelligence Data can be complete.

The preparation of this thesis is intended as one of the requirements for obtaining a Master's degree in the Cyber Defense Engineering Study Program, Faculty of Science and Defense Technology, The Republic of Indonesia Defense University.

The preparation of this thesis was completed thanks to the help and support from various parties, both directly and indirectly. For this reason, on this occasion the researcher would like to thank:

1. Dr. Ir. H. Achmad Farid Wajdi, M.M. as advisor I and Prof. Ir. Teddy Mantoro, MSc., Ph.D., SMIEEE II as advisor II for their support and guidance so far and for providing direction to researchers so that this thesis can be completed.
2. The reviewer has provided criticism and suggestions in improving this report.
3. Kolonel Laut (E) Dr. H.A. Danang Rimbawa, S.Si., M.T., M.Tr.Opsla., CEH, CSBA as Head of the Cyber Defense Engineering Study Program and all staff, lecturers and students in the Cyber Defense Engineering Study Program, as well as the entire Defense University community who have helped smooth lectures.
4. Kolonel Laut (P) Ruby Alamsyah, M.Tr.Opsla., M.Han., CIPA., CIT., CIIQA and Kolonel Laut (E) Suginta Ginting, S.Kom., MMSI., M.Tr.Hanla who has fully supported the thesis writing and lecture activities at Defense University.
5. The parties who have helped the researcher a lot during the process of collecting data and writing this thesis. Thank you for taking the time to

discuss, and for sharing your knowledge with researchers so that this scientific work can be completed.

6. All my beloved family, especially father, mother, wife and children, who always pray for the smooth progress of the thesis.
7. National Cyber and Crypto Agency for its support in assignments during the lecture process and thesis work.
8. All fellow researchers whose names I cannot mention one by one due to limited space.

May Allah SWT repay the kindness of various parties for their assistance.

The researcher realizes that this thesis is still imperfect, therefore the researcher humbly hopes for constructive criticism and suggestions to support the perfection of this research.

Finally, we hope that this thesis can provide benefits to the development of defense science and be useful for stakeholders in efforts to improve national security and defense in the cyber sector.

Bogor, January , 2024

A handwritten signature in black ink, consisting of several overlapping loops and a long horizontal stroke extending to the left.

Agung Nugroho

ABSTRACT

SENTIMENT ANALYSIS MODEL USING BI-LSTM AND LATENT DIRICHLET ALLOCATION ON YOUTUBE COMMENTS TO SUPPORT INTELLIGENCE DATA

AGUNG NUGROHO

YouTube has the potential to become a propaganda medium that can disrupt national security stability. So comments on YouTube videos can be used as a medium for assessing public opinion which is formed from acts of propaganda. The presence of Artificial Intelligence technology makes it easier to assess public opinion but is hampered by accuracy and datasets. This research proposes modeling using Bi-Directional Long Short-Term Memory (Bi-LSTM) and Latent Dirichlet Allocation (LDA) using quantitative methods to assess sentiment categories and topics from comments on YouTube videos. This research aims to design a model that is novel in the form of combining a sentiment analysis model with the addition of topic category output to support intelligence data collection on social media. The research results show that the Bi-LSTM model with Word2Vec has the highest performance compared to other models, such as Logistic Regression (LR), Multinomial Naïve Bayes (MNB), K-Nearest Neighbors (KNN), and Random Forest (RF), with an average value -average accuracy, precision, recall and F1-Score reaches 98%. The results of this research contribute to the assessment of public opinion and categorization of topics currently being widely discussed by the public on YouTube as an open source of intelligence information to support defense strategies.

Keywords: Bi-LSTM, Deep Learning, LDA, Pancagatra, Youtube

TABLE OF CONTENTS

COVER PAGE	i
TITLE PAGE	ii
THESIS APPROVAL SHEET	iii
THESIS VERIFICATION SHEET.....	iv
DECLARATION OF ORIGINALITY.....	v
FOREWORDS	vi
ABSTRACT	viii
TABLE OF CONTENTS	ix
LIST OF FIGURES.....	xi
LIST OF TABLES	xiii
LIST OF CHARTS	xiv
LIST OF ABBREVIATIONS	xv
CHAPTER I INTRODUCTION	1
1.1 Background.....	1
1.2 Problems Identification.....	8
1.3 Problem Formulation.....	9
1.4 Research Scope and Limitation	10
1.5 Research Objectives	11
1.6 Research Benefits.....	11
CHAPTER II LITERATURE REVIEW	13
2.1.Theoretical Framework	13
2.1.1. National Defense Concept: Information Warfare.....	13
2.1.2. Sentiment Analysis	16
2.1.3. Youtube Social Media.....	23
2.1.4. Data Scraping on Youtube Comments	25
2.1.5. Public Dataset	26
2.1.6. Word2Vec Embedding.....	28
2.1.7. Bi-Directional Long Short-Term Memory.....	32

2.1.8. Accuracy Evaluation Model	33
2.1.9. Pancagatra News Articles Dataset	34
2.1.10. Latent Dirichlet Allocation	36
2.2. Previous Research.....	39
2.3. Research Framework.....	51
CHAPTER III RESEARCH METHODOLOGY.....	53
3.1 Research Methods and Design	53
3.2 Research Location and Time	57
3.3 Data Collection Techniques	58
3.4 Data Processing Techniques	62
3.5 Data Analysis Techniques.....	63
CHAPTER IV RESULT AND DISCUSSION	68
4.1 Data Description	68
4.2 Data Collection Results.....	71
4.3 Data Processing Results.....	75
4.4 Data Analysis Results	76
4.5 Discussion	83
CHAPTER V CONCLUSION AND RECOMMENDATION	96
5.1 Conclusion.....	96
5.2 Recommendation.....	97
REFERENCES.....	98
ATTACHMENT.....	104

LIST OF FIGURES

Figure 1. Top Websites : Similarweb Ranking.....	1
Figure 2. Data Pre-Processing Step	19
Figure 3. Sentiment Analysis Result Example.....	22
Figure 4. Scraping Comments Process on Youtube.....	25
Figure 5. NLP Semmantic.....	29
Figure 6. Word2Vec Skip-gram Method	31
Figure 7. Bi-LSTM Architecture	32
Figure 8. Research Model Design.....	55
Figure 9. Bi-LSTM Layer Architecture with Word2Vec Embedding	65
Figure 10. Composition Dataset.....	69
Figure 11. Frequency of Comment Sentence Length.....	70
Figure 12. Public Dataset.....	72
Figure 13. Wordcloud Negative Comments on Data Collection.....	73
Figure 14. Wordcloud Positive Comments on Data Collection	73
Figure 15. Topic Category Dataset	74
Figure 16. Wordcloud Negative Comments Data Processing Results	75
Figure 17. Wordcloud Positive Comments Data Processing Results	76
Figure 18. Measurement with 70% Data Training and 30% Data Testing	77
Figure 19. Measurement with 80% Data Training and 20% Data Testing	78
Figure 20. Measurement with 90% Data Training and 10% Data Testing	79
Figure 21. Word Frequency in Public Datasets	80
Figure 22. LDA Model Percentage Data	82
Figure 23. Pancagatra Topic Presentation Diagram.....	83
Figure 24. Youtube Video Thumbnail in Sample 1	86
Figure 25. Live Simulation Sentiment Output on Sample 1	87
Figure 26. Timeseries Number of Users on Youtube Comments Sample 1	87

Figure 27. Topics in Sample Youtube Comments 1	88
Figure 28. Youtube Video Thumbnail in Sample 2	89
Figure 29. Live Simulation Sentiment Output on Sample 2	90
Figure 30. Timeseries Number of Users on Youtube Comments Sample 2	90
Figure 31. Topics in Sample Youtube Comments 2	91
Figure 32. Youtube Video Thumbnail in Sample 3	92
Figure 33. Live Simulation Sentiment Output on Sample 3	93
Figure 34. Timeseries Number of Users on Youtube Comments Sample 3	93
Figure 35. Topics in Sample Youtube Comments 3	94

LIST OF TABLES

Table 1. Public Dataset.....	28
Table 2. Research Environments.....	57
Table 3. Research Timeline.....	58
Table 4. Research Data.....	60
Table 5. Average Model Evaluation Score.....	84
Table 6. Sample Youtube Video Simulation 1.....	86
Table 7. Sample Youtube Video Simulation 2.....	89
Table 8. Sample Youtube Video Simulation 3.....	92

LIST OF CHARTS

Chart 1. Sentiment Analysis Method	17
Chart 2. Bi-LSTM on Sentiment Analysis	18
Chart 3. Data Pre-Processing Step	19
Chart 4. Graphical LDA Model	36
Chart 5. Research Framework	51
Chart 6. Data Pre-processing	62
Chart 7. Data Analysis Model using Word2Vec + Bi-LSTM	63
Chart 8. Analysis Model using LDA.....	66

LIST OF ABBREVIATIONS

AI	: Artificial Intelligence
Bi-LSTM	: Bidirectional Long Short-Term Memory
CBOW	: Continuous Bag of Word
DT	: Decision Trees
Ipoleksosbudhankam	: Ideology, Politics, Economics, Socio-Culture, Defense and Security
Kemhan RI	: Ministry of Defense of the Republic of Indonesia
KNN	: k-Nearest Neighbors
LDA	: Latent Dirichlet Allocation
LR	: Logistic Regression
MNB	: Multinomial Naïve Bayes
NLP	: Natural Language Processing
OSINT	: Open Source Intelligence
RF	: Random Forest