



**PROJECT REPORT**  
**ANALYSIS FAKE NEWS USING RANDOM FOREST**  
**AND SUPPORT VECTOR MACHINE**

**JULIUS WARIH ANGKASA**  
**16.K1.0059**

**Faculty of Computer Science**  
**Soegijapranata Catholic University**  
**2020**

## APPROVAL AND RATIFICATION PAGE



Final Project Title : Analysis Fake News Using Random Forest And Support Vector Machine

Submitted by : Julius Warih Angkasa

ID : 16.K1.0059

Date approved : July 17, 2020

Approved by

Supervisor : Hironimus Leong, S.Kom., M.Kom.

Examiners 1 : R. Setiawan Aji Nugroho, S.T., MCompIT., Ph.D

Examiners 2 : Hironimus Leong, S.Kom., M.Kom.

Examiners 3 : Rosita Herawati, S.T., M.I.T.

Examiners 4 : Y.b. Dwi Setianto, S.T., M.Cs.

Head of Undergraduate Program : Rosita Herawati, S.T., M.I.T.

Dean of Faculty Computer Science : R. Setiawan Aji Nugroho, S.T., MCompIT., Ph.D

This page is a legitimate page and can be verified via the address below.

[sintak.unika.ac.id/skripsi/verifikasi/?id=16.K1.0059](http://sintak.unika.ac.id/skripsi/verifikasi/?id=16.K1.0059)

## STATEMENT OF ORIGINALITY

I, the undersigned:

Name : JULIUS WARIH ANGKASA

ID : 16.K1.0059

Certify that this project was made by myself and not copy or plagiarize from other people, except that in writing expressed to the other article. If it is proven that this project was plagiarizes or copy the other, I am ready to accept a sanction.



Semarang, July 17, 2020



JULIUS WARIH ANGKASA  
16.K1.0059

**APPROVAL PAGE FOR PUBLICATION OF  
SCIENTIFIC PAPERS FOR ACADEMIC INTEREST**

The undersigned below :

Name : JULIUS WARIH ANGKASA  
Undergraduate Program : TECHNICAL INFORMATION  
Faculty : COMPUTER SCIENCE  
Type of work : SKRIPSI

Approved to give Non-Exclusive Royalty Free Right to Soegijapranata Catholic University Semarang for scientific work entitled “**Analysis Fake News Using Random Forest And Support Vector Machine**” along with the existing tools (if needed). With this Non-Exclusive Royalty Free Right Soegijapranata Catholic University has the right to store, transfer data / format, manage in the form of a database, maintain, and publish this final project as long as I keep my name as a writer / creator and as a Copyright owner.

This statement I made in truth.

Semarang, July 17, 2020

Sincerely



JULIUS WARIH ANGKASA

## ACKNOWLEDGEMENTS

First and foremost, praises and thanks to the God, the Almighty, for His showers of blessings throughout my final project work to complete the final project successfully. The final project is intended as one of the requirements to take the Bachelor of Computer Science exam in the Informatics Engineering Study Program at Soegijapranata Catholic University Semarang.

In the preparation of this final project, the author received assistance from various parties. Therefore, on this occasion the author would like to express his thanks to

1. Father, Mother, Teteh, and Bima who always provide prayer and support to the author while working on the final project.
2. Hironimus Leong, S. Kom., M. Kom as a Supervising Lecturer who has provided guidance and input to the author, so that the final project can be completed properly.
3. Campus friends, especially Tian, Gilbert, Alvido, Reinaldo, Aldus, and Ko Revival who always provide guidance, prayer, and motivation.
4. And other parties which the author cannot mention one by one and have provided support and assistance provided to the author during the preparation of this final project.

Semarang, July 17, 2020



JULIUS WARIH ANGKASA

## ABSTRACT

*Dissemination of information through online media is very fast, online media can influence people's mindsets in consuming information that is being scouted. So that the internet is very easy for humans to spread information. Easily disseminating this information is not uncommon for many individuals or groups not responsible for spreading false news.*

*To develop technology in dealing with deceptive news, it is necessary analysis fake news. The study uses the calculation of the TF-IDF weight value and uses two algorithms, namely Support Vector Machine and Random Forest algorithm. Before calculating the weight value of TF-IDF, it is necessary to do the pre-processing stage. The word results from pre-processing are used to calculate the TF-IDF value, after calculating the TF-IDF value it is necessary to calculate the TF-IDF category. TF-IDF category to determine grouping based on TF-IDF value. The TF-IDF results were trained using two algorithms, namely Support Vector Machine and Random Forest algorithms. In the test carried out 4 scenarios with portions of training data and testing data that is different - different. Training data is used to train machines, while testing data is used for testing in terms of prediction results, accuracy, precision, recall, f1-score values.*

*Based on the results of research in analyzing news titles using a dataset of 1100 data, the Random Forest algorithm obtained the highest accuracy value of 72.04301075268818% with 1007 training data. And the lowest accuracy value is 64.61538461538461% with 760 training data. Then the test results using the Support Vector Machine algorithm obtained the highest accuracy value of 65.59139784946237% with 1007 training data and the lowest accuracy value was 49.67032967032967% with the number of training data 644 data. While the results of research in analyzing news content using the Random Forest algorithm, obtained the highest accuracy value of 63.44086021505376% with 1007 training data, and the lowest accuracy value of 58.24175824175825% with the number of training data 644 data. Then the results of the analysis of the news content using the Support Vector Machine algorithm obtained the highest accuracy value of 57.692307692307686 %% with 760 training data and the lowest accuracy value was 40.18691588785047% with the amount of training data 886 data.*

*Keyword: Analysis Fake News, TF-IDF, Support Vector Machine, Random Forest.*

## TABLE OF CONTENTS

Cover.....	i
APPROVAL AND RATIFICATION PAGE.....	ii
STATEMENT OF ORIGINALITY.....	iii
APPROVAL PAGE FOR PUBLICATION OF.....	iv
SCIENTIFIC PAPERS FOR ACADEMIC INTEREST.....	iv
ACKNOWLEDGEMENTS.....	v
ABSTRACT.....	vi
TABLE OF CONTENTS.....	vii
ILLUSTRATION INDEX.....	viii
INDEX OF TABLES.....	ix
<b>CHAPTER 1 INTRODUCTION.....</b>	<b>1</b>
1.1 Background.....	1
1.2 Problem Formulation.....	3
1.3 Scope.....	3
1.4 Objective.....	3
<b>CHAPTER 2 LITERATURE STUDY.....</b>	<b>4</b>
<b>CHAPTER 3 RESEARCH METHODOLOGY.....</b>	<b>9</b>
<b>CHAPTER 4 ANALYSIS AND DESIGN.....</b>	<b>16</b>
4.1 Analysis.....	16
4.2 Desain.....	65
<b>CHAPTER 5 IMPLEMENTATION AND TESTING.....</b>	<b>68</b>
5.1 Implementation.....	68
5.2 Testing.....	87
<b>CHAPTER 6 CONCLUSION.....</b>	<b>91</b>
REFERENCES.....	
APPENDIX.....	A

## ILLUSTRATION INDEX

Illustration 1: SVM images with maximum hyperplane margins.....	57
Illustration 2: Illustration Confusion Matrix.....	63
Illustration 3: Flow Chart.....	65
Illustration 4: Use Case Diagram.....	67





## INDEX OF TABLES

Table 1: Table Training Data.....	16
Table 2: Table Testing Data.....	18
Table 3: Training Data Results that have been changed to ASCII type.....	18
Table 4: Results of Testing Data that have been changed to ASCII type.....	21
Table 5: Results of Training Data that have been removed '[' .....	22
Table 6: Results of Testing Data that have been removed '[' .....	25
Table 7: Results of Training Data that have been through the Folding process.....	25
Table 8: Testing Data Results that have been through the Folding process.....	27
Table 9: Results of Training Data that have gone through the Stemming process.....	28
Table 10: Results of Testing Data that have gone through the Stemming process.....	30
Table 11: Results of Training Data that have been through the filtering process.....	31
Table 12: Results of Testing Data that have been through the filtering process.....	33
Table 13: Split Data Results (Title).....	34
Table 14: Split Data Results (Content).....	35
Table 15: Term Frequency Results (Title).....	36
Table 16: Term Frequency Results (Content).....	39
Table 17: Document Frequency Result (Title).....	43
Table 18: Document Frequency Result (Content).....	44
Table 19: Inverse Document Frequency Result (Title).....	46
Table 20: Inverse Document Frequency Result (Content).....	46
Table 21: Term Frequency Result - Inverse Document Frequency (Title).....	48
Table 22: Term Frequency Result - Inverse Document Frequency (Content).....	49
Table 23: Results for the Term Frequency - Inverse Document Frequency category (Title) .....	51
Table 24: Results for the Term Frequency - Inverse Document Frequency category (Content).....	51
Table 25: Training Data Results after going through Pre-processing,TF-IDF and TF-IDF Category.....	52
Table 26: Testing Data Results after going through Pre-processing,TF-IDF and TF-IDF Category.....	53
Table 27: Document tables with labels.....	60
Table 28: Results of the calculation table for each Training data.....	60
Table 29: Table Prediction calculation results.....	61
Table 30: Document tables with labels.....	62
Table 31: The results of the calculation table for each Training data.....	62
Table 32: Table Prediction calculation results.....	62
Table 33: Table Confusion Matrix Results on Support Vector Machine algorithm (Title and Content).....	64
Table 34: Table Confusion Matrix Results on Random Forest algorithm (Title and Content).....	64
Table 35: Result Performance (Title and Content ).....	64
Table 36: Training Data and Testing Data Tables.....	87

Table 37: Analysis Table of the News Title section.....87  
Table 38: Analysis Table of the News Content section.....89

