



# PROJECT REPORT

## NEWS TOPIC CLASSIFICATION WITH MACHINE LEARNING USING SEMI-SUPERVISED LEARNING

SAMUEL KURNIAWAN SANTOSO  
18.K1.0019

Faculty of Computer Science  
Soegijapranata Catholic University  
2022



## HALAMAN PENGESAHAN

Judul Tugas Akhir:	: News Topic Classification with Machine Learning using Semi-Supervised Learning
Diajukan oleh	: Samuel Kurniawan Santoso
NIM	: 18.K1.0019
Tanggal disetujui	: 04 Januari 2022
Telah setujui oleh	
Pembimbing	: R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D
Pengaji 1	: Rosita Herawati S.T., M.I.T.
Pengaji 2	: R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D
Pengaji 3	: Hironimus Leong S.Kom., M.Kom.
Pengaji 4	: Y.b. Dwi Setianto S.T., M.Cs.
Pengaji 5	: Yulianto Tejo Putranto S.T., M.T.
Pengaji 6	: Yonathan Purbo Santosa S.Kom., M.Sc
Ketua Program Studi	: Rosita Herawati S.T., M.I.T.
Dekan	: Dr. Bernardinus Harnadi S.T., M.T.

Halaman ini merupakan halaman yang sah dan dapat diverifikasi melalui alamat di bawah ini.

[sintak.unika.ac.id/skripsi/verifikasi/?id=18.K1.0019](http://sintak.unika.ac.id/skripsi/verifikasi/?id=18.K1.0019)

## DECLARATION OF AUTHORSHIP

I, the undersigned:

Name : SAMUEL KURNIAWAN SANTOSO  
ID : 18.K1.0019

declare that this work, titled " News Topic Classification with Machine Learning using Semi-Supervised Learning", and the work presented in it is my own. I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at Soegijapranata Catholic University
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
3. Where I have consulted the published work of others, this is always clearly attributed.
4. Where I have quoted from the work of others, the source is always given.
5. Except for such quotations, this work is entirely my own work.
6. I have acknowledged all main sources of help.
7. Where the work is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Semarang, January, 11, 2022



SAMUEL KURNIAWAN SANTOSO

18.K1.0019

## **HALAMAN PERNYATAAN PUBLIKASI KARYA ILMIAHUNTUK KEPENTINGAN AKADEMIS**

Yang bertanda tangan dibawah ini:

Nama : Samuel Kurniawan Santoso

Program Studi : Teknik Informatika

Fakultas : Ilmu Komputer

Jenis Karya : Tugas Akhir

Menyetujui untuk memberikan kepada Universitas Katolik Soegijapranata Semarang Hak Bebas Royalti Nonekslusif atas karya ilmiah yang berjudul "**NEWS TOPIC CLASSIFICATION WITH MACHINE LEARNING USING SEMI-SUPERVISED LEARNING**" beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Nonekslusif ini Universitas Katolik Soegijapranata berhak menyimpan, mengalihkan media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir ini selama tetap mencantumkan nama saya sebagai penulis / pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya

Semarang, 11 Januari 2022

Yang menyatakan

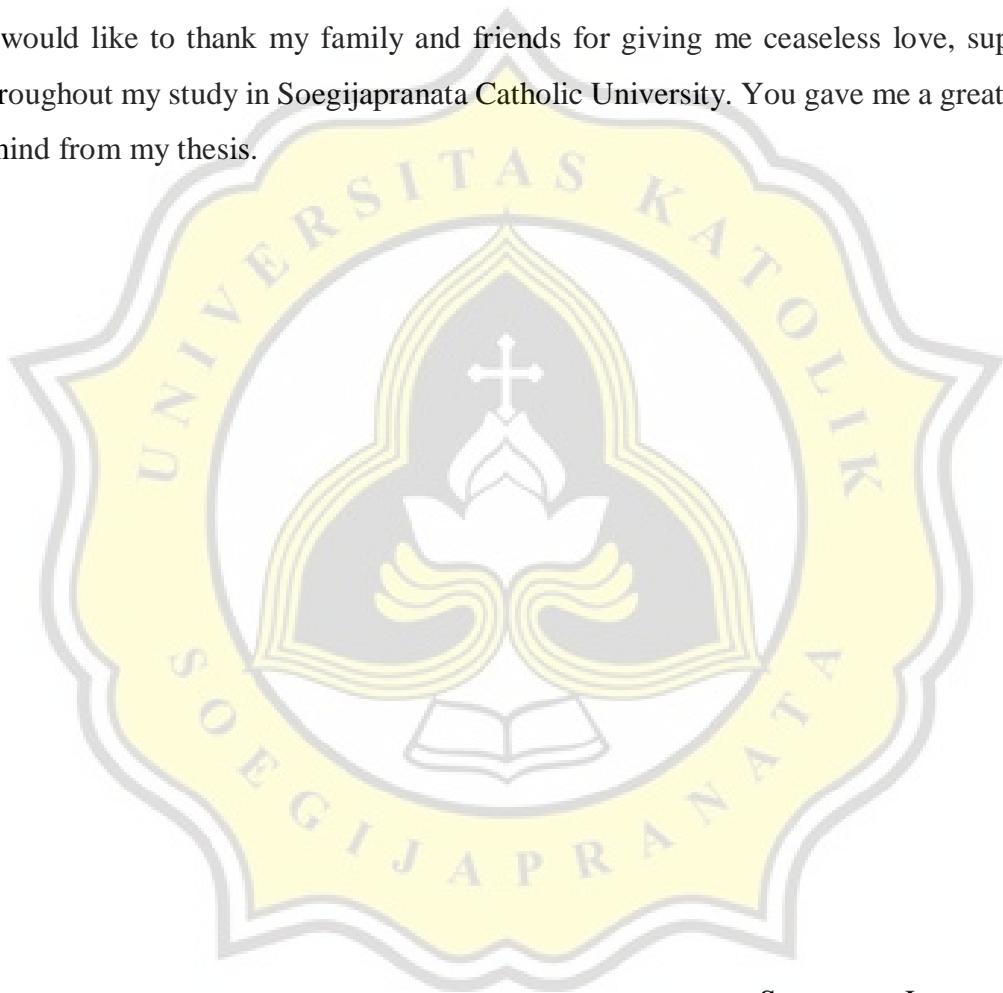


Samuel Kurniawan Santoso

## **ACKNOWLEDGMENT**

I have received a myriad of support, advice, and assistance throughout this document writing. I would like to thank my supervisors Robertus Setiawan Aji Nugroho for formulating this topic. I would also like to thank my friend Davin Chang for annotating my dataset. Also Alexander Jason Lauwren and Stephen Royanmart Patrick for guiding with advice to finish this document.

I would like to thank my family and friends for giving me ceaseless love, support, and advice throughout my study in Soegijapranata Catholic University. You gave me a great escape to rest my mind from my thesis.



Semarang, January, 11, 2022

A handwritten signature in black ink, appearing to read "Samuel Kurniawan S."

SAMUEL KURNIAWAN S

18.K1.0019

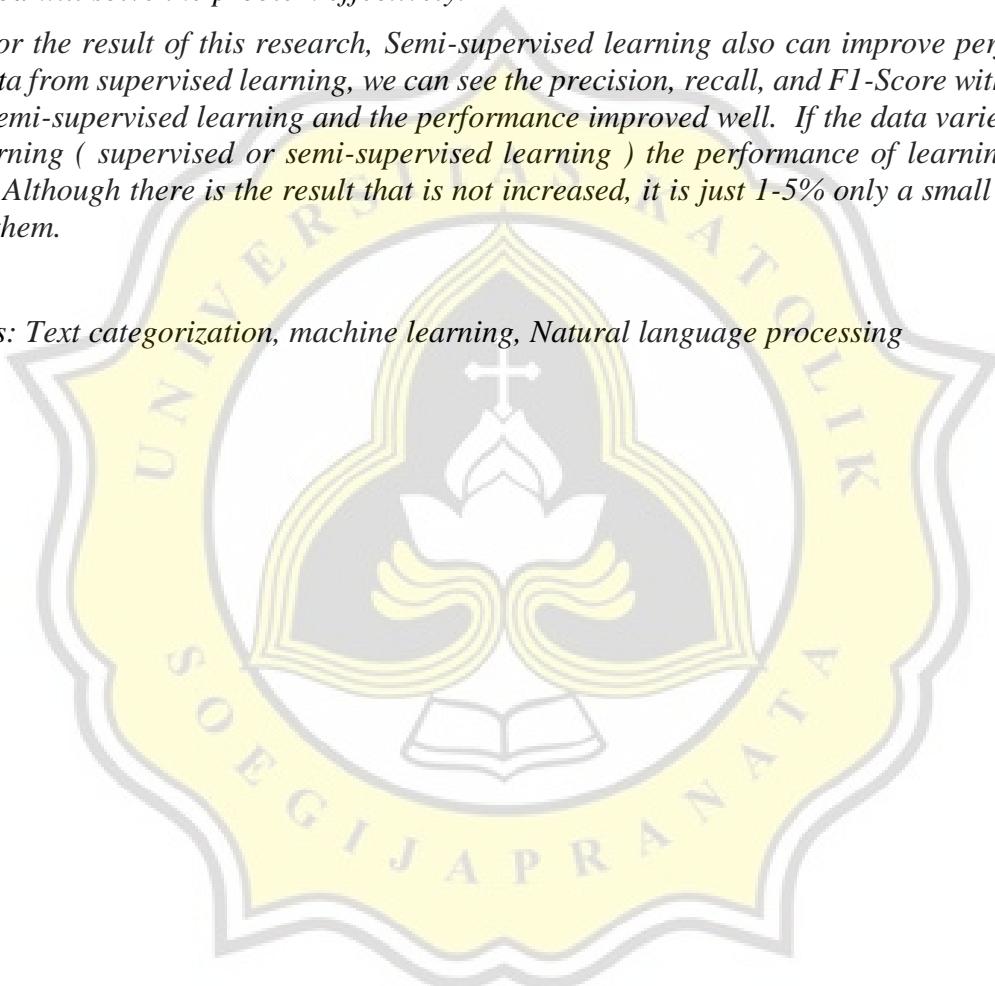
## ABSTRACT

Nowadays, the amount of news has increased every day. They also grow rapidly. This situation makes it difficult for the editor to categorize manually and this manual categorization makes the categories incorrect for the news. They also wasted the time choosing the category for the news which were published.

For that problem, we proposed news categories classification using machine learning with semi-supervised learning with pseudo labeling method. We also proposed an ensemble learning algorithm called Adaboost that can boost the data from misclassified. Hopefully, this algorithm and method will solve the problem effectively.

For the result of this research, Semi-supervised learning also can improve performance. By the data from supervised learning, we can see the precision, recall, and F1-Score with the same ratio in semi-supervised learning and the performance improved well. If the data varied with the same learning ( supervised or semi-supervised learning ) the performance of learning also increased. Although there is the result that is not increased, it is just 1-5% only a small difference between them.

*Keywords:* Text categorization, machine learning, Natural language processing



## TABLE OF CONTENTS

<b>COVER .....</b>	<b>i</b>
<b>HALAMAN PENGESAHAN</b>	
<b>N PENGESAHAN.....</b>	<b>ii</b>
<b>DECLARATION OF AUTHORITY.....</b>	<b>iii</b>
<b>HALAMAN PERNYATAAN PUBLIKASI KARYA ILMIAH .....</b>	<b>iv</b>
<b>ACKNOWLEDGMENT.....</b>	<b>v</b>
<b>TABLE OF CONTENTS.....</b>	<b>vii</b>
<b>LIST OF FIGURE .....</b>	<b>ix</b>
<b>LIST OF TABLE .....</b>	<b>xi</b>
<b>CHAPTER 1 INTRODUCTION.....</b>	<b>12</b>
1.1.    Background .....	12
1.2.    Problem Formulation .....	13
1.3.    Scope .....	13
1.4.    Objective .....	13
<b>CHAPTER 2 LITERATURE STUDY .....</b>	<b>14</b>
2.1.    Background .....	14
2.2.    News .....	14
2.3.    Natural Language Processing (NLP) Classification .....	14
2.4.    Machine learning .....	15
2.5.    Supervised Learning .....	16
2.6.    Semi-Supervised Learning Method .....	16
2.6.1.    Self-Training .....	17
2.6.2.    Co-training.....	17
2.7.    News Categorization.....	18

<b>CHAPTER 3 RESEARCH METHODOLOGY.....</b>	<b>21</b>
3.1.    Overview .....	21
3.2.    Dataset .....	21
3.2.1.    Data Processing .....	25
3.2.2.    Imbalanced Class .....	26
3.2.3.    Outliers.....	27
3.2.4.    Analysis .....	27
3.3.    Algorithm.....	29
3.3.1.    Adaboost Algorithm.....	30
3.3.2.    Supervised Learning.....	31
3.3.3.    Semi-Supervised Learning.....	32
3.3.4.    Evaluation Model.....	32
<b>CHAPTER 4 ANALYSIS AND DESIGN .....</b>	<b>33</b>
4.1.    Overview .....	33
4.2.    Analysis .....	33
4.3.    Design.....	35
<b>CHAPTER 5 IMPLEMENTATION AND RESULTS.....</b>	<b>39</b>
5.1.    Overview .....	39
5.2.    Implementation.....	39
5.3.    Results.....	45
<b>CHAPTER 6 CONCLUSION.....</b>	<b>56</b>
<b>REFERENCES .....</b>	<b>a</b>

## LIST OF FIGURE

<b>Figure 3.1</b> Flowchart of This Research.....	21
<b>Figure 3.2</b> Dataset After Stopwords Removal .....	22
<b>Figure 3.3</b> Dataset Before Stopwords Removal .....	22
<b>Figure 3.4</b> Dataset After Stemming .....	23
<b>Figure 3.5</b> Dataset Before Stemming.....	23
<b>Figure 3.6</b> Dataset Before Tokenization .....	24
<b>Figure 3.7</b> Dataset After Tokenization .....	24
<b>Figure 3.8</b> TF-IDF Formula .....	25
<b>Figure 3.9</b> Imbalanced Dataset .....	26
<b>Figure 3.10</b> Boxplot to Detect Outliers .....	27
<b>Figure 3.11</b> Word Count without Stopwords Filtering .....	28
<b>Figure 3.12</b> Ekonomi class without Stopwords Filtering .....	28
<b>Figure 3.13</b> Ekonomi class with Stopwords Filtering .....	29
<b>Figure 3.14</b> Adaboost Multiclass Formula.....	30
<b>Figure 3.15</b> The Alpha Value SAMME Formula.....	31
<b>Figure 3.16</b> The New Weight Value SAMME Formula.....	31
<b>Figure 3.17</b> SAMME Predicted Formula.....	31
<b>Figure 3.18</b> Pseudolabelling in Semi-Supervised Learning .....	32
<b>Figure 4.1</b> Precision General Formula .....	33
<b>Figure 4.2</b> Precision Micro-Averaged .....	34
<b>Figure 4.3</b> Recall General Formula .....	34

<b>Figure 4.4</b> Recall Micro-Averaged .....	34
<b>Figure 4.5</b> Macro-Averaged F1 Score .....	35
<b>Figure 4.6</b> Micro-Averaged F1 Score .....	35
<b>Figure 4.7</b> Flowchart for News Categorization.....	36
<b>Figure 4.8</b> Labeled Dataset .....	36
<b>Figure 4.9</b> Adaboost Multiclass Formula.....	37
<b>Figure 4.10</b> The Alpha Value SAMME Formula.....	37
<b>Figure 4.11</b> The New Weight Value SAMME Formula.....	37
<b>Figure 4.12</b> SAMME Predicted Formula.....	38
<b>Figure 5.1</b> Dataset Info .....	39
<b>Figure 5.2</b> The Top 20 Word Count of Whole Data.....	42
<b>Figure 5.3</b> Boxplot to Detect Outliers.....	42
<b>Figure 5.4</b> Precision Formula .....	48
<b>Figure 5.5</b> Recall Formula .....	48

## LIST OF TABLE

<b>Table 5.1</b> Imbalanced Class Supervised Learning Table .....	47
<b>Table 5.2</b> Micro-averaged with Same Value.....	48
<b>Table 5.3</b> Imbalanced Class Semi-Supervised Learning Table.....	49
<b>Table 5.4</b> Balanced Class Supervised Learning Table.....	50
<b>Table 5.5</b> Balanced Class Semi Supervised Learning Table.....	51
<b>Table 5.6</b> Imbalanced Class Supervised Learning 780 rows without SVD Algorithm ....	52
<b>Table 5.7</b> Balanced Class with 780 Rows Data Supervised Learning Table with SVD Algorithm.....	53

