SENTIMENT ANALYSIS AND TEXT CLASSIFICATION OF PEDULI LINDUNGI APPS

THESIS REPORT

SEPTIAN PRATAMA
41518010101

DEPARTMENT OF INFORMATICS
FACULTY OF COMPUTER SCIENCE
UNIVERSITAS MERCU BUANA
JAKARTA
2021

SENTIMENT ANALYSIS AND TEXT CLASSIFICATION OF PEDULI LINDUNGI APPS

THESIS REPORT

Submitted to Complete Terms
Completed a Computer Bachelor Degree

Created By:

SEPTIAN PRATAMA
41518010101

DEPARTMENT OF INFORMATICS
FACULTY OF COMPUTER SCIENCE
UNIVERSITAS MERCU BUANA
JAKARTA
2021

ORIGINALITY STATEMENT SHEET

ORIGINALITY STATEMENT SHEET

The undersigned below:
Student Number : 41518010101
Name : Septian Pratama
Title : Sentiment Analysis and Text Classification of Peduli Lindungi Apps

Stating that my Final Project Report is the work of my own and not a plagiarism.If it is found in my Final Project Report that there is an element of plagiarism, then I am ready to get academic sanctions related to it

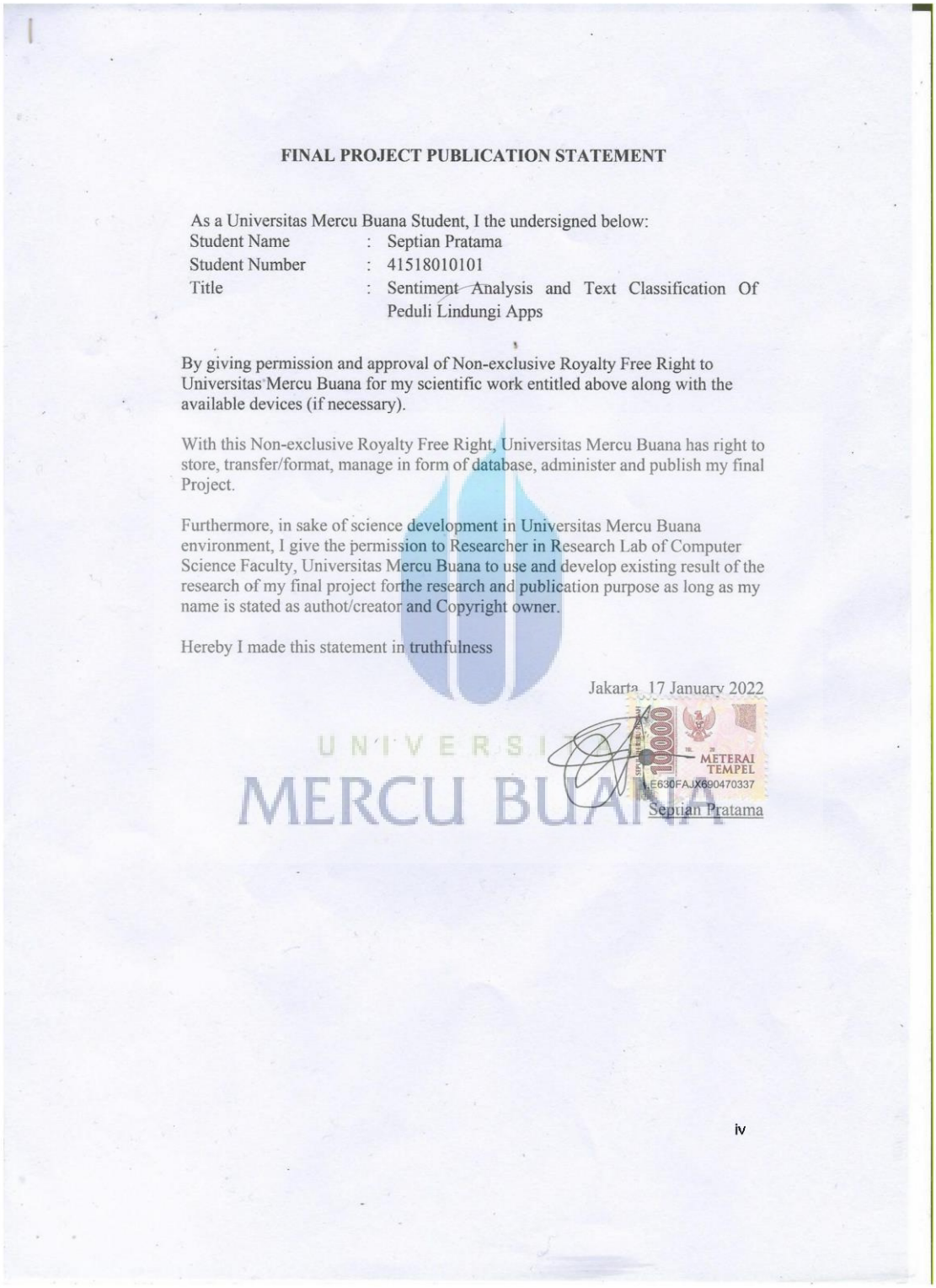
Jakarta, 17 January 2022

METERAI TEMPEL 10000 6203CAJX690470342

Septian Pratama

iii

FINAL PROJECT PUBLICATION STATEMENT



FINAL PROJECT PUBLICATION STATEMENT

As a Universitas Mercu Buana Student, I the undersigned below:

Student Name	:	Septian Pratama
Student Number	:	41518010101
Title	:	Sentiment Analysis and Text Classification Of Peduli Lindungi Apps

By giving permission and approval of Non-exclusive Royalty Free Right to Universitas Mercu Buana for my scientific work entitled above along with the available devices (if necessary).

With this Non-exclusive Royalty Free Right, Universitas Mercu Buana has right to store, transfer/format, manage in form of database, administer and publish my final Project.

Furthermore, in sake of science development in Universitas Mercu Buana environment, I give the permission to Researcher in Research Lab of Computer Science Faculty, Universitas Mercu Buana to use and develop existing result of the research of my final project forthe research and publication purpose as long as my name is stated as authot/creator and Copyright owner.

Hereby I made this statement in truthfulness

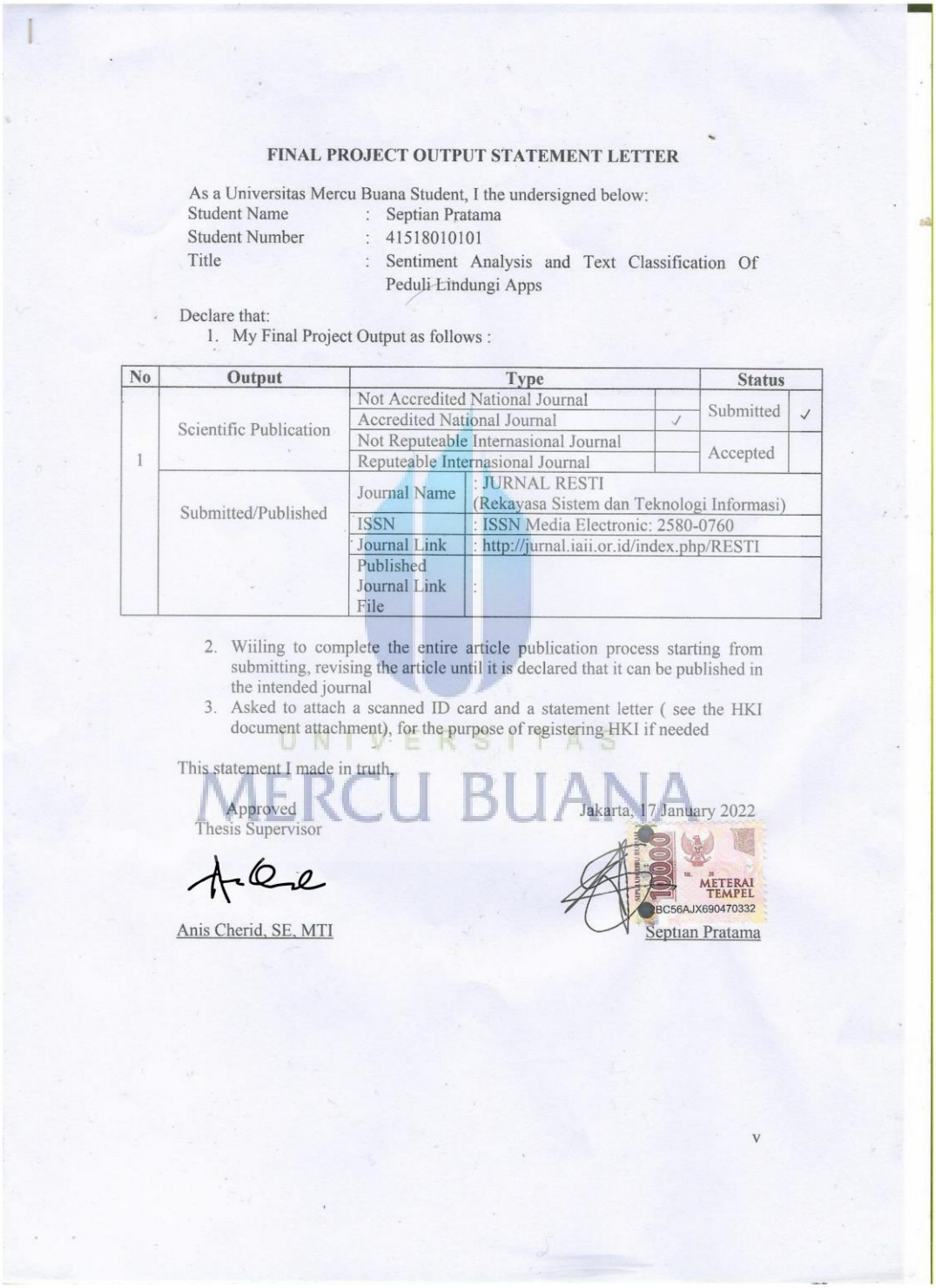
Jakarta, 17 January 2022

METERAI TEMPEL 10000

Septian Pratama

iv

FINAL PROJECT OUTPUT STATEMENT LETTER



FINAL PROJECT OUTPUT STATEMENT LETTER

As a Universitas Mercu Buana Student, I the undersigned below:

Student Name : Septian Pratama
Student Number : 41518010101
Title : Sentiment Analysis and Text Classification Of Peduli Lindungi Apps

Declare that:

1. My Final Project Output as follows :

No	Output	Type		Status	
1	Scientific Publication	Not Accredited National Journal		Submitted	✓
		Accredited National Journal	✓		
		Not Reputeable Internasional Journal		Accepted	
		Reputeable Internasional Journal			
	Submitted/Published	Journal Name	: JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)		
		ISSN	: ISSN Media Electronic: 2580-0760		
		Journal Link	: http://jurnal.iaii.or.id/index.php/RESTI		
		Published Journal Link File	:		

2. Wiiling to complete the entire article publication process starting from submitting, revising the article until it is declared that it can be published in the intended journal
3. Asked to attach a scanned ID card and a statement letter (see the HKI document attachment), for the purpose of registering HKI if needed

This statement I made in truth,

Approved
Thesis Supervisor

Anis Cherid, SE, MTI

Jakarta, 17 January 2022

Septian Pratama

EXAMINER APPROVAL SHEET

Student ID	:	41518010101
Student Name	:	SEPTIAN PRATAMA
Title	:	Sentiment Analysis and Text Classification Of Peduli Lindungi Apps

This thesis has been examined and heard as one of the requirements to obtain a Bachelor's degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta,

Approved,

Dr. Leonard Goeirmanto, M.Sc

EXAMINER APPROVAL SHEET

Student ID	:	41518010101
Student Name	:	SEPTIAN PRATAMA
Title	:	Sentiment Analysis and Text Classification Of Peduli Lindungi Apps

This thesis has been examined and heard as one of the requirements to obtain a Bachelor's degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta,

Approved,

Dr. Rahmat Budiarto

EXAMINER APPROVAL SHEET

Student ID	:	41518010101
Student Name	:	SEPTIAN PRATAMA
Title	:	Sentiment Analysis and Text Classification Of Peduli Lindungi Apps

This thesis has been examined and heard as one of the requirements to obtain a Bachelor's degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta,

Approved

Emil Robert Kaburuan, Ph.D

COMMITTEE APPROVAL SHEET

Student Number	:	41518010101
Student Name	:	SEPTIAN PRATAMA
Title	:	Sentiment Analysis and Text Classification Of Peduli Lindungi Apps

This thesis has been examined and heard as one of the requirements to obtain a Bachelot's degree in the Informatics Engineering Study Program. Faculty of Computer Science, Universitas Mercu Buana.

Jakarta,

Approved,

( )
Dr. Rahmat Budiarto

( )
Emil Robert Kaburuan, Ph.D

( )
Dr. Leonard Goeirmanto, M.Sc

VALIDATION SHEET

Student Name	:	SEPTIAN PRATAMA
Title	:	Sentiment Analysis and Text Classification Of Peduli Lindungi Apps

This Final Project has been examined and defenced as one of the requirements for obtaining a Bachelor's degree in the Informatics Engineering Study Program, Faculty of Compuer Science, Universitas Mercu Buana.

Jakarta,

Approved,

(Anis Cherid, SE, MTI)
Thesis Supervisor

Acknowledged,

(Wawan Gunawan, S.Kom, MT))	(Emil R. Kaburuan, Ph.D)
Informatics Thesis Coordinator	Head of Informatics Department

ABSTRAK

Nama	:	SEPTIAN PRATAMA
NIM	:	41518010101
Pembimbing TA	:	Anis Cherid, SE, MTI
Judul	:	Sentiment Analysis and Text Classification Of Peduli Lindungi Apps

Berbagai upaya telah dilakukan oleh pemerintah Republik Indonesia untuk menangani penyebaran virus covid19. Salah satu terobosan dari pemerintah adalah dengan membuat aplikasi PeduliLindungi, aplikasi ini diharapkan mampu memberikan peringatan kepada masyarakat saat memasuki wilayah terdampak COVID-19, lokasi fasilitas kesehatan dan tracking jika ada masyarakat yang berpotensi terinfeksi virus covid19.Dalam penelitian ini peneliti menggunakan ulasan komentar dan opini publik pada aplikasi PeduliLindungi, baik positif maupun negatif dalam bahasa Indonesia. Di era pandemi ini, aplikasi PeduliLindungi merupakan salah satu aplikasi yang banyak digunakan oleh masyarakat, oleh karena itu penelitian ini sangat menarik untuk diolah menjadi informasi. Penggunaan informasi ini memerlukan teknik analisis sehingga dapat menghasilkan dan membantu pengembang atau pemerintah mendengar opini dari masyarakat dalam rangka penyempurnaan aplikasi PeduliLindungi di era pandemi Covid-19. Penelitian ini menggunakan metode Multinomial Naive Bayes. Metode MultinomialNB untuk mencari sentimen yang menghasilkan Akurasi sebesar 76% dan untuk mengklasifikasikan Review pada kelas yang tepat untuk menghasilkan Akurasi 70% dengan menggunakan 2000 dataset untuk mencari sentimen dan 1081 ulasan komentar yang telah di labeling secara manual oleh peneliti untuk mencari kategori class yang tepat pada sebuah ulasan komentar. Tujuan dari penelitian ini adalah untuk membantu pemerintah meningkatkan kualitas aplikasi pedulilindungi dengan menggunakan ulasan komentar yang didapatkan dari google playstore dan juga untuk mengklasifikasikan ulasan komentar ke dalam kelas yang tepat menggunakan pembelajaran mesin.

Kata kunci: Sentiment, Pedullindungi, klasifikasi text, komentar, *Multinomial Naive Bayes*

ABSTRACT

Nama	:	SEPTIAN PRATAMA
NIM	:	41518010101
Pembimbing TA	:	Anis Cherid, SE, MTI
Judul	:	Sentiment Analysis and Text Classification Of Peduli Lindungi Apps

Various efforts have been made by the government of the Republic of Indonesia to deal with the spread of the COVID-19 virus. One of the breakthroughs from the government is to create an application PeduliLindung, this application is expected to be able to provide warnings to the public when entering areas affected by COVID-19, the location of health facilities and tracking if there are people who have the potential to be infected with the Covid-19 virus. In this study, researchers used comments and opinions. Public on the PeduliLindung application, both positive and negative in Indonesian. In this pandemic era, the PeduliLindung application is one application that is widely used by the public, therefore this research is very interesting to be processed into information. The use of this information requires analytical techniques so that it can generate and help developers or the government hear opinions from the public in order to improve the PeduliLindung application in the Covid-19 pandemic era. This study uses the Multinomial Naive Bayes method. MultinomialNB to find the sentiment that produces an Accuracy of 76% and to classify the Review at an appropriate class to produce Accuracy 70% by using the 2000 dataset to search for the sentiment and 1081 review comments that have been in the labeling manually by researchers to find a category class which is appropriate in a review comment.. The purpose of this study is to help the government improve the quality of the application pedulilindungi by using the review comments obtained from google playstore and also to classify comment reviews into proper classes using machine learning.

Keywords: Sentiment, Pedullindung, text classification, review, Multinomial Naive Bayes

PREFACE

Praise our gratitude for the presence of Allah SWT, because with His grace & guidance author could complete this thesis report, as a condition for completing the Bachelor degree (S1) in Informatika Engineering at Universitas Mercu Buana. The author is fully aware that in completing this thesis report will not escape the support and guidance of the closest people, therefore the author would like to express my gratitude as psossible to:

1. Dr. Ngadino Surip as the Chancellor of Mercu Buana University who has provided many positive changes and progress for our university.
2. Yaya Sudarya Triana, M.Kom., Ph.D. as Dean of the Faculty of Computer Science, Mercu Buana University
3. Emil R. Kaburuan, Ph.D., as Head of the Department of Informatics Engineering, Mercu Buana University
4. Anis Cherid, SE, MTI, as Head of the International Department of Informatics, Mercu Buana University, as well as academic supervisor, Thank you for the knowledge that you have discussed with me as a guide in completing the thesis report.
5. Lecturer of the Department of Informatics for the knowledge, dedication, and motivation given during the lecture period.
6. Staff Mercu Buana University who has provided invaluable assistance for the author to complete this thesisthis final project
7. Parents and family who always pray for and support the author in completing.
8. Classmates from the English Informatics Class who have been together for 3 years and continue to motivate the author to complete this thesis report.

In writing this thesis, the author realizes that this is still not perfect, therefore constructive criticism and suggestions from all parties are highly expected. Hopefully this thesis report can add knowledge to the parties concerned. The author would like to thank you very much for the guidance and all the support given, may Allah SWT bestow His grace and gifts.

Jakarta,17 January 2022

Septian Pratama

TABLE OF CONTENT

JOURNAL

Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 5 No. x (2021) x - x ISSN Media Elektronik: 2580-0760

SENTIMENT ANALYSIS AND TEXT CLASSIFICATION OF PEDULI LINDUNGI APPS

Septian Pratama
Informatics, Fasilkom, Universitas Mercu Buana Jakarta
septiannn009@gmail.com

Abstract

Various efforts have been made by the government of the Republic of Indonesia to deal with the spread of the COVID-19 virus. One of the breakthroughs from the government is to create an application PeduliLindung, this application is expected to be able to provide warnings to the public when entering areas affected by COVID-19, the location of health facilities and tracking if there are people who have the potential to be infected with the Covid-19 virus. In this study, researchers used comments and opinions. Public on the PeduliLindung application, both positive and negative in Indonesian. In this pandemic era, the PeduliLindung application is one application that is widely used by the public, therefore this research is very interesting to be processed into information. The use of this information requires analytical techniques so that it can generate and help developers or the government hear opinions from the public in order to improve the PeduliLindung application in the Covid-19 pandemic era. This study uses the Multinomial Naive Bayes method. MultinomialNB to find the sentiment that produces an Accuracy of 76% and to classify the Review at an appropriate class to produce Accuracy 70% by using the 2000 dataset to search for the sentiment and 1081 review comments that have been in the labeling manually by researchers to find a category class which is appropriate in a review comment.. The purpose of this study is to help the government improve the quality of the application pedulilindungi by using the review comments obtained from google playstore and also to classify comment reviews into proper classes using machine learning.

Keywords: *Sentiment, Pedullindung, text classification, review, Multinomial Naive Bayes*

Abstrak

Berbagai upaya telah dilakukan oleh pemerintah Republik Indonesia untuk menangani penyebaran virus covid19. Salah satu terobosan dari pemerintah adalah dengan membuat aplikasi PeduliLindungi, aplikasi ini diharapkan mampu memberikan peringatan kepada masyarakat saat memasuki wilayah terdampak COVID-19, lokasi fasilitas kesehatan dan tracking jika ada masyarakat yang berpotensi terinfeksi virus covid19.Dalam penelitian ini peneliti menggunakan ulasan komentar dan opini publik pada aplikasi PeduliLindungi, baik positif maupun negatif dalam bahasa Indonesia. Di era pandemi ini, aplikasi PeduliLindungi merupakan salah satu aplikasi yang banyak digunakan oleh masyarakat, oleh karena itu penelitian ini sangat menarik untuk diolah menjadi informasi. Penggunaan informasi ini memerlukan teknik analisis sehingga dapat menghasilkan dan membantu pengembang atau pemerintah mendengar opini dari masyarakat dalam rangka penyempurnaan aplikasi PeduliLindungi di era pandemi Covid-19. Penelitian ini menggunakan metode Multinomial Naive Bayes. Metode MultinomialNB untuk mencari sentimen yang menghasilkan Akurasi sebesar 76% dan untuk mengklasifikasikan Review pada kelas yang tepat untuk menghasilkan Akurasi 70% dengan menggunakan 2000 dataset untuk mencari sentimen dan 1081 ulasan komentar yang telah di labeling secara manual oleh peneliti untuk mencari kategori class yang tepat pada sebuah ulasan komentar. Tujuan dari penelitian ini adalah untuk membantu pemerintah meningkatkan kualitas aplikasi pedulilindungi dengan menggunakan ulasan komentar yang didapatkan dari google playstore dan juga untuk mengklasifikasikan ulasan komentar ke dalam kelas yang tepat menggunakan pembelajaran mesin.

Kata kunci: Sentiment, Pedullindungi, klasifikasi text, komentar, *Multinomial Naive Bayes*

1. Introduction

Coronavirus 19 is a highly contagious viral infection caused by the acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) [1]. It began to spread widely in Wuhan, Hubei Province, China. Most of the patients were epidemiologically associated with the Huanan Seafood Wholesale Market [2]. The spread of Coronavirus has become a major global public health event, linking people's physical and mental health and even mental safety [3]. Indonesia is the fourth most populous country in the world and is expected to be significantly affected by Coronavirus over a longer period [4].

Monitoring the spread of Coronavirus in Indonesia is handled by the Government of Indonesia in various ways, one of which is through the Android application installed by Google Play. This application made by the government is called Peduli Lindung which was developed to assist relevant government agencies in tracking to stop the spread of Coronavirus Disease. This application relies on community participation to share location data with each other while traveling so that contact history tracing with Coronavirus sufferers can be carried out and users of this application will also receive notifications if they are in a crowd or are in a red zone, namely an area or sub-district that has been recorded that there are People who are infected with Coronavirus are positive or there are Patients Under Supervision of the Applications, Care for this is welcomed by the community in dealing with Coronavirus. The Pedulilindungi for allows users to provide reviews about user satisfaction with the applications used, this aims to evaluate application performance so that improvements can be made. To find this out, it is necessary to do a sentiment analysis.

Sentiment analysis itself or also commonly referred to as opinion mining is one part of text mining. This field conducts the study of people's opinions, sentiments, evaluations, behavior and emotions textually towards a service entity, organization, individual, problem, topic, event and its attributes. The use of sentiment analysis is generally used to analyze a product in improving product quality in the future. In this case sentiment analysis can be applied to application reviews [5].

2. Research Method

Research starts from the existence of a problem that is important, interesting and needs a solution. To create an effective and efficient research, a structured framework of thinking is needed and conveyed through pictures with structured stages related to the actions to be taken, which can be seen in Figure 1.

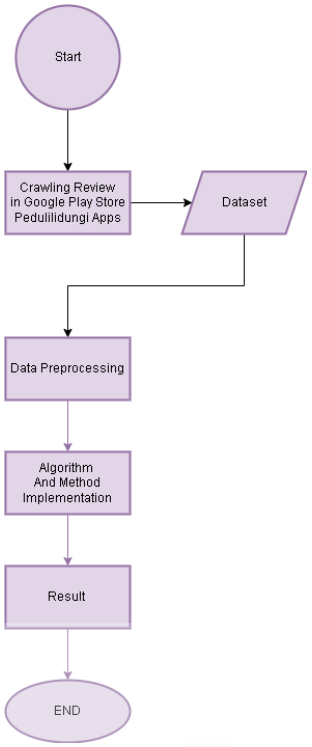


Figure 1. Research stage

2.1. Data Collection

This data collection was carried out to strengthen the reason why this research should be carried out, data collection was taken from the source directly, namely from users of the PeduliLindungi application on the Google Play Store, attributes are shown in Table 1.

Table 1. Dataset attribute Pedulilindungi

Attribute Name	Description
Content	Review User
Value	Category Class
Sentiment	Rating Score from user

Table 2. Example Dataset Pedulilindungi

Content	label	sentiment
Operator yang input umum teramat sangat super lelet sekali, harus berapa lama saya nunggu sertifikat vaksin muncul. #tolong jangan bot yang jawab.	sertifikat	NEGATIF
Aplikasi gk bisa Login,, pilih warga negara indonesia, dicentang, cuma diem aja,, gk beralih ke halaman login atau gimana,, tolong d perbaiki ya...	login	NEGATIF

Aplikasi luar biasa, bentuk kepedulian pemerintah pada kesehatan dan keselamatan masyarakat	Pujian	POSITIF
Jangan takut paksin, karna itu semua demi kesehatan dan keamanan kita sendiri	umum	NEGATIF
Tolong pengembang aplikasi atau yang bertanggung jawab atas aplikasi ini. Tolong semua pertanyaan dijawab. Ini aplikasi penting loh. Sepele tapi penting. Mau scan kartu kalo gk ada barkode ya gak bisa. Terus buat apa aplikasi ini. Tolong cerdas dalam informasi dan teknologi??.	Komplain	NEGATIF

The data used for the classification text uses a different amount of data on sentiment analysis because the data in the sentiment analysis found that the Umum category and Sertifikat are more numerous than other categories and this can make the model not work well.

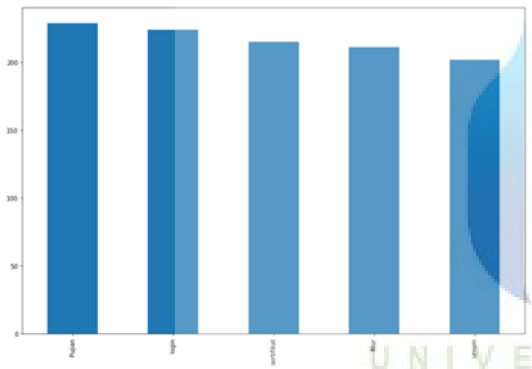


Figure 2. Category label

Table 3. Description of category label

Cateogry Class	Total Review in class
Pujian	229
login	224
Sertifikat	215
Fitur	211
Umum	202

2.2. Data Labeling

This study uses 2 stages of labeling. The first uses a rating score from the Pedulilindungi application review and the second manually by the researcher to assign a category class to the Pedulilindungi application review, 2000 data taken from Google Play Store will be analyzed as research material.

2.3. Data Preprocessing

Data Preprocessing is a process for making low quality data into high quality data making it easy to process [6].

In the process of preprocessing, attributes that have less influence on the Process of Classification will be reduced.. The data entered at this stage is still raw data, so the result of this process is a quality document that will facilitate the process classification. There are several data preprocessing techniques used in this research; case folding, tokenization, filtering, stemming

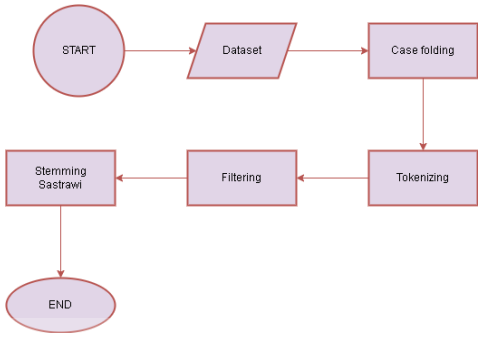


Figure 3. Preprocessing process

2.3.1 Case Folding

The first stage is Case folding, which is the stage of uniformization of the form of letters to lowercase or uppercase, and the elimination of numbers and punctuation.

Table 4. Result of Case folding

Review	Case folding
Sudah lancar dan bisa masuk ke akun saya, terima kasih... Kembangkan terus aplikasinya..	sudah lancar dan bisa masuk ke akun saya terima kasih kembangkan terus aplikasinya

2.3.2. Tokenizing

Tokenizing is the process of separating text into pieces called tokens for later analysis. Words, numbers, symbols, punctuation marks and other important entities can be considered tokens. In NLP, tokens are defined as "words" although tokenize can also be done in paragraphs or sentences.The split() function in python can be used to split text.

Table 5. Result of Tokenizing

Review	Tokenizing
sudah lancar dan bisa masuk ke akun saya terima kasih kembangkan terus aplikasinya	[sudah, lancar, dan, bisa, masuk, ke, akun, saya, terima, kasih, kembangkan, terus, aplikasinya]

2.3.3. Filtering (Stopword Removal)

Filtering is the stage of taking important words from the token results by using a stoplist algorithm (discarding less important words) or wordlists (saving important words). Stopwords are common words that usually appear in large numbers and are considered meaningless. Examples of stopwords in Indonesian are "yang", "and", "di", "from", etc. The meaning behind using stopwords is that by removing low-information words from a text, we can focus on the important words instead.

Table 6. Result of Filtering

Review	Filtering
sudah lancar dan bisa masuk ke akun saya terima kasih kembangkan terus aplikasinya	['sudah', 'lancar', 'masuk', 'akun', 'terima', 'kasih', 'kembangkan', 'aplikasinya']

2.3.4 Stemming Sastrawi

Stemming is the process of removing the inflection of words to their basic form. For example, in English text, the only process required is removing the suffix. Meanwhile, in Indonesian texts, all affixes, both suffixes and prefixes, are also omitted.

Table 7. Result of Stemming Sastrawi

Review	Stemming Sastrawi
sudah lancar dan bisa masuk ke akun saya terima kasih kembangkan terus aplikasinya	lancar masuk akun terima kasih kembang aplikasi

2.4. Term Weighting

The TF-DF algorithm is an algorithm based on statistical values showing the occurrence of a word in the document [7]. TF (Term Frequency) states the number of words that appear in a document. DF (Document Frequency) states the number of documents containing a word in one publication segment.

$$IDF = log(\frac{N}{DF(w)}) \quad (1)$$

$$TF - IDF(w,d) = TF(w,d)x\ IDF(w) \quad (2)$$

TF-lDF (w,d) is the weight of a word in the whole document, w is a word (word), d is a document (document), TF(w,d) is the frequency of occurrence of a word w in the document d, IDF(w) is inverse DF of the word , N is the total number of documents, DF(w) is the number of documents containing the word w

2.2.5 Multinomial Naïve Bayes

Multinomial NBC is a development model of the Bayes algorithm which is suitable for classifying text or documents. In the Multinomial Naive Bayes Classifier formula, the document class is not only determined by the words that appear but also the number of occurrences [8]. The naive Bayes classifier method consists of two stages in the text classification process, the training stage and the classification stage. This method is used because MNB is a special method of Naïve Bayes. Naïve Bayes is a method that is fast, easy to implement with a simple and effective structure [9]. And in the previous study it was stated that the Naive Bayes Multinomial was better than the Naive Bayes method for Indonesian text classification [10]. The equation for class selection or classification calculations using the Multinomial Naïve Bayes method can be defined in equation (6) by looking for the Prior results in equation (4), Conditional Probability in equation (5) and the illustration of using the equation in table 7 using the responses in table 6.

$$P(c) = \frac{N_c}{N} \quad (4)$$

$$P\ (w|c) = \frac{count(w,c)+1}{count(c)+|V|} \quad (5)$$

$$P(c|d) \propto P(c)P(w|c) \quad (6)$$

w is Word, c is Class (positive or negative), N is Sum of response, d is Response, Count (w, c) is Sum of words w in class c, Count(c) is Sum of words in class c, |V| is Sum of vocabulary, P(c|d) is Posterior Probability, the opportunity of class c for response d, P(w|c) is Conditional Probability (likelihood), the opportunity arises the word w in class c, P(c) is Class Prior Probability, the opportunity arises the class c, $\propto$ is Proportional

Table 8. Data for calculation of MultinomialNB

Description	Review	Sentiment
Data Training	Aplikasi sudah bagus (Good Application)	1
	Bagus(Good)	1
	Jelek (Bad)	0
Data Testing	Sudah bagus (Already Good)	?

Table 9. Calculation of MultinomialNB

Function	POSITIVE	NEGATIVE
P(c)	2/3 = 0,66	1/3 = 0,33
P('aplikasi'|c)	(1+1) / (3+4) = 2/7 = 0,29	(0+1) / (1+4) = 1/5 = 0,2
P('sudah'|c)	(1+1) / (3+4) = 2/7 = 0,29	(0+1) / (1+4) = 1/5 = 0,2
P('bagus'|c)	(2+1) / (3+4) = 3/7 = 0,43	(0+1) / (1+4) = 1/5 = 0,2
P('jelek'|c)	(0+1) / (3+4) = 1/7 = 0,14	(1+1) / (1+4) = 2/5 = 0,4
P(c|'sudah','bagus')	0,66 * 0,29 * 0,43 = 0,08	0,33 * 0,2 * 0,2 = 0,01

From the results of the calculations in table 9, the highest posterior results for the test data response "already good"

or already good in table 6 is 0.08 which is obtained from the results of the previous multiplication in the positive class is 0.66 and the possibility of good words and in the positive class is 0.66 0.29 and 0.43. And it can be concluded that "already good" is a positive response.

2.5. Data Mining

Data Mining is a process that employs one or more computer learning techniques to analyze and extract knowledge automatically. Data Mining is an iterative and interactive process to suggest new patterns or models that are perfect, useful and understandable in a very large database.

Data Mining contains the search for the desired trend or pattern in a large database to help make decisions in the future. These patterns are recognized by certain tools which can provide a useful and insightful analysis of data which can then be studied more thoroughly, possibly using other decision support tools. [11]

2.6. Peduli Lindungi Application

PeduliLindungi is an application developed to assist relevant government agencies in tracking to stop the spread of Coronavirus Disease (COVID-19).

This application relies on community participation to share location data with each other while traveling so that contact history tracing with COVID-19 sufferers can be carried out.

Users of this application will also get a notification if they are in a crowd or are in a red zone, namely an area or sub-district where it has been recorded that there are people who are infected with positive COVID-19 or there are patients under surveillance.

2.7. Sentiment Analysis

Sentiment Analysis is the process of understanding and processing textual data automatically to obtain sentiment information contained in a sentence or text in the form of an opinion. The purpose of sentiment analysis is to see the views or opinions of the text related to a problem or object, whether it tends to have a positive or negative view. Sentiment analysis consists of natural language processing, text analysis and computational linguistics to identify the sentiments of a document.[12]

Figure 4. Sentiment analysis process

2.8. Text Classification

Text classification is a field of research in information acquisition that develops methods to determine or categorize a document into one or more previously known groups automatically based on the contents of the document [13]. Document classification aims to classify unstructured documents into groups that describe the contents of the document

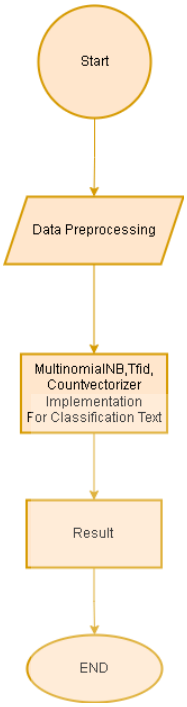


Figure 5. Text classification process

2.9 Word Cloud

Word cloud is a picture that shows a list of words used in a text, generally the more words that are used the greater the size of the word in picture. At this study will use data scraping results from the Google Play Store page PeduliLindungi Applications. The scraped data itself has been cleaned simply to remove punctuation marks, reduce letters, etc.

3. Result and Discussion

In the study to determine sentiment analysis and classification text using the Multinomial Naïve Bayes model, where the results of sentiment and data classification methods will be followed by a calculation process to determine the accuracy and accuracy of predictions using the python programming language on google colab cloud tools.

3.1. Sentiment Analysis

In sentiment analysis on a Dataset Pedulindung with the same stage using the text processing and after that the implementation of the validation Algorithm is the K-fold, the researchers used the K-10 and the result is not working properly, the accuracy is given for the K-10 is only given 53%.Meanwhile, MultinomialNB in the PeduliLindung dataset uses 1500 training data and 500 data testing and the results get 76% accuracy. The researcher assumes that K-fold Validation does not work well because there are too many datasets provided so that K-fold does not work well and also because each dataset will be randomly trained 10x and make K-fold performance decrease compared to Naive Bayes.

In sentiment analysis of words that often appear in negative sentiment is a problem that is always obtained by user applications such as difficult to enter / login, certificates are always late update, harder registration and slow the info vaccination, can be seen in [Figure 6]

While the positive sentiment in words that often Appears are words of praise for protective care applications such as, thank you, applications, easy login / login, easy registration, quick vaccination info.[Figure 7]

Table 10. Result of Confusion Matrix Naïve Bayes and K-fold

Confusion matrix Naïve Bayes:	
[[203 54]	
[[66 177]]	
MultinomialNB Accuracy	: 0.76
MultinomialNB Precision	: 0.75
MultinomialNB Recall	: 0.78
MultinomialNB F1 Score	: 0.77
Confusion matrix K-fold:	
[[36 2]	
[[100 62]]	
MultinomialNB Accuracy	: 0.53
MultinomialNB Precision	: 0.53
MultinomialNB Recall	: 0.46
MultinomialNB F1 Score	: 0.46

Table 11. Result of Sentiment Naïve Bayes and K-fold

Naïve Bayes	Precision	Recall	F1-Score	Support
NEGATIF	0.75	0.79	0.77	257
POSITIF	0.77	0.73	0.75	243
K-Fold	Precision	Recall	F1-Score	Support
NEGATIF	0.26	0.95	0.41	38
POSITIF	0.97	0.38	0.55	162

Figure 6. Word cloud negative sentiment

Figure 7. Word cloud positive sentiment

3.2. Text Classification

In this text classification stage, the researcher initially had 10 class categories for modeling, such as; umum,fitur,komplain,sertifikat,OTPcode,pujian,registrasi,update,scan, dan login [Figure 8]during modeling, and it turns out that the results given in 10 categories get 30% accuracy for text classification modeling, researchers have done 3x modeling by making balance the dataset in each category and reduce the dataset category, the researcher conducted a 4x experiment with 5 categories and training data as much as 810 data [Figure 9] and balanced a dataset resulting in better accuracy than before. The 5 categories that were deleted were, scan, OTP code, certificate, registration and update, the researcher assumed that the five categories could be included in the general category and complaints. The way the model works is to find the most occurrences of words in each document in each category and provide the results of occurrences in each category after that stage, the highest results in each category will be considered as the correct category in the existing document.[Figure 10]

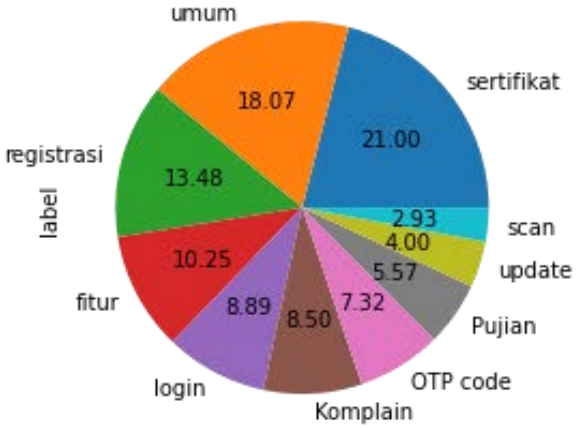


Figure 8.Ten (10) category

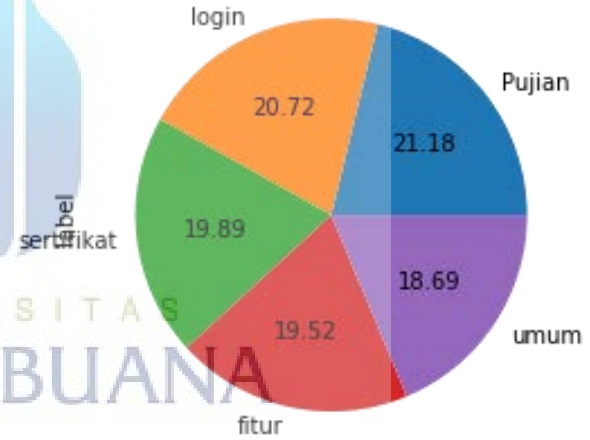


Figure 9.Five (5) category

```
X_test1= ['Sangat membantu, karena mudah untuk scan kalau bepergian']
predicted = text_clf.predict(X_test1)
prob=text_clf.predict_proba(X_test1)

[ ] str(predicted[0])

    'Pujian'

[ ] prob

    array([[0.28424846, 0.27971103, 0.14780626, 0.12728488, 0.16094937]])

[ ] text_clf.classes_

    array(['Pujian', 'fitur', 'login', 'sertifikat', 'umum'], dtype='<U10')

[ ] max(prob[0])

    0.2842484598283412
```

Figure 10. Predict text classification

4. Conclusion

Based on the results carried out with Multinomial Naive Bayes using the python programming language on Google Colaboratory, the average model accuracy is 76% for sentiment analysis and the final results for negative sentiments produce 75% precision, 79% recall and 77% f-1 score, while for positive sentiment it produces 77% precision, 73% recall and 75% f-1 score, from the results that can be concluded to see the sentiment assessment of the PeduliLindung application, it can be seen from the f-1 score of each result which means negative comments more than positive sentences.

Text classification model obtained with an accuracy of 70%. The results for text classification initially obtained an accuracy of 30% because the category classes were in a dataset of 10 categories, to find these 10 categories the researchers did manual labeling, in the modeling the researchers finally decided to only use 5 category classes in order to produce better results. model accuracy, and the results of the MultinomialNB model using 5 categories resulted in a model accuracy value of 70% with 810 training data, therefore the researchers chose to reduce these categories because the accuracy given was higher. after the data up sample in this study. Research shows that the use of the Naive bayes classifier can cope with sentiment analysis and text classification on a dataset.

Table 12. Result of Sentiment Naïve Bayes

Naïve Bayes	Precision	Recall	F1-Score	Support
NEGATIF	0.75	0.79	0.77	257
POSITIF	0.77	0.73	0.75	243

Table 13. Result of Model Sentiment Naïve Bayes and Text Classification

Description	Result
Sentiment Analysis	76%
Text Classification	70%

References

[1] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses," J. Adv. Res., vol. 24, pp. 91–98, 2020.

[2] L. Yang et al., "Epidemiological and clinical features of 200 hospitalized patients with coronavirus disease 2019 outside Wuhan, China: A descriptive study.," J. Clin. Virol., vol. 129, no. March, p. 104475, 2020

[3] W. Cai, B. Lian, X. Song, T. Hou, G. Deng, and H. Li, "A crosssectional study on mental health among health care workers during the outbreak of Corona Virus Disease 2019," Asian J. Psychiatr., vol. 51, no. March, p. 102111, 2020.

[4] R. Djalante et al., "Review and analysis of current responses to COVID-19 in Indonesia: Period of January to March 2020," Prog. Disaster Sci., vol. 6, p. 100091, 2020.

[5] Christoper Fernandez Zefanya , Dika Rizky Akbar , Yohana Titirloloby , Wendry Rizky Ramadhan , Fransiska Febria Situmorang, "Analisis Sentiment Berdasarkan Ulasan Komentar Terhadap Aplikasi Pedulilidungi Menggunakan Metode NAÏVE BAYES", 2020 November

[6] J Han M Kamber and J Pei Data Mining: Concepts and Techniques 3rd Edition Elsevier 2011.

[7] A. P. Wijaya. 2016. "Klasifikasi Dokumen dengan Naive Bayes Classifier (NBC) untuk Mengetahui Konten EGoverment," Journal of Applied Intelligent System, Vol.1, No. 1, pp. 48-55,

[8] I. H. Witten, F. Eibe and M. A. Hall. 2011. Data mining : Practical Machine Learning I. H. Witten, F. Eibe and M. A. Hall. 2011. Data mining : Practical Machine Learning

[9] X. Wu, V Kumar, J R Quinlan, J Ghosh, Q Yang, H Motoda, G J McLachlan, A Ng, B Liu, P S Yu, Z Zhou, M Steinbach, D J Hand and D Steinberg 2008 Top 10 algorithms in data mining vol. 14 no. 1

[10] R A Aziz, M S Mubarok and Adiwijaya 2016 Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes no. August 139–148

[11] Hermawati, F. A. (2013). Data Mining. Yogyakarta: Penerbit Andi.

[12] G, V., & Chandrasekaran, D. (2015). A comparative performance evaluation of neural network based approach for sentiment classification of online reviews. Journal of King Saud University.

[13] Tenenboim, L., Shapira, B., & Shoval,P., 2008, Ontology-based classification of news in an electronic news paper Paper presented at Intelligent Information and Engineering Systems Conference, Varna, Bulgaria

WORKING SHEET

This working paper is material to complete a journal article entitled "Sentiment Analysis and Text Classification Of Care Protect Apps". This working paper will contain all the research material for the Final Project that has not been published in a journal article. In this paper, the following sections are presented:

1. Literature Review is a section that contains the results of literature studies carried out related to the experiments carried out. Broadly speaking, the literature review conducted on the concept of Data Mining and Naïve Bayes Classifier in a commentary on the PeduliLindung application.

2. Analysis and design is a part that consists of an outline and the stages carried out in this research. This stage uses data preprocessing, data training and testing using the Naive Bayes Algorithm.

3. The source code in this study uses the Python programming language on Google Colab. The use of Python in this study is used to clean up unused words and train the available datasets and process them using the Naïve Bayes algorithm.

4. The dataset describes the entire data taken using the Python programming language in a comment from the PeduliLindung application from 2021.

5. Experimental Stage is a section that contains all experimental stages that are not included in the journal. This section describes the overall technical flow of the research. The stages described in this section include the stages of data preprocessing, data collection, implementation of TF-IDF and the Naïve Bayes Algorithm.

6. The results of all experiments are part of the results of the modeling experiments carried out. using the Naïve Bayes Algorithm.