

Sentiment Analysis of Google Play Store User Reviews on Digital Population Identity App Using K-Nearest Neighbors

Rudi Kurniawan^{[1]*}, Harma Oktafia Lingga Wijaya^[2], Rani Purnama Aprisusanti^[3]

Department of Computer System Engineering, Faculty of Engineering Science^[1]

Department of Information System, Faculty of Engineering Science^{[2] [3]}

Universitas Bina Insan Lubuklinggau

Lubuk Linggau, Indonesia

rudi.kurniawan@univbinsan.ac.id^[1], harmaoktafialingga@univbinsan.ac.id^[2], ranigeulis08@gmail.com^[3]

Abstract— The Digital Population Identity Application provides convenience for users to access and manage their population data digitally. Based on the increasing usage of the Digital Population Identity Application on the Google Play Store, various user reviews of the application have emerged. Therefore, sentiment analysis is needed to provide a deeper understanding of user perceptions and to classify user reviews of the Digital Population Identity Application. Sentiment analysis is a computational study of opinions, feelings, and emotions expressed in text, using the K-Nearest Neighbors method, which is a classification method based on the closest distance or similarity to objects in the training data. Using 5000 relevant review data from September 2022 to December 2023, after labeling them into positive, negative, and neutral sentiment classes, the results show 3581 negative sentiments, 1031 positive sentiments, and 388 neutral sentiments. Testing was conducted by applying the K-Nearest Neighbors method in the classification stage, testing this method by varying K values from 1 to 10. The best results were obtained with a training data ratio of 90% to testing data ratio of 10%. The best results were achieved at K values of 8, 9, and 10, with an accuracy of 81%, precision of 82%, recall of 95%, and an F1-Score of 88%. With a training data ratio of 70% to testing data ratio of 30%, the best results were obtained at K values of 6, 7, 8, 9, and 10, with an accuracy of 80%, precision of 81%, recall of 95%, and an F1-Score of 88%. Based on the results of this research, the K-Nearest Neighbors method can be used for sentiment classification of user reviews with good results.

Keywords— Sentiment Analysis, Digital Population Identity, K-Nearest Neighbors.

I. INTRODUCTION

In the era of 5.0 society with advancing technology, the government, as a public servant, enhances innovation in providing services to the community by implementing E-Government [1]. One of the implementations is by the Directorate General of Civil Registration of the Ministry of Home Affairs launching the Digital Population Identity Application. This application provides users with ease of access and management of their population data digitally.

With the increasing usage of the Digital Population Identity Application on the Google Play Store, various user reviews of the application have emerged. Therefore, sentiment analysis is necessary to provide a deeper understanding of user perceptions and to classify user reviews of the Digital Population Identity Application. By analyzing user reviews, developers and stakeholders gain valuable insights into user satisfaction levels. Understanding the sentiments expressed by users helps identify areas of improvement and areas where the application excels. Sentiment analysis is a computational study of opinions,

feelings, and emotions expressed in text [2].

Given the abundance of reviews, it is necessary to classify them into positive, neutral, and negative reviews [3]. The author utilizes the K-Nearest Neighbors method, a classification method based on the closest distance or similarity to objects in the training data. In the training phase, this algorithm stores feature vectors and classifies training data. The advantage of the K-Nearest Neighbor method lies in its proven accuracy and consistency with calculations used in applications [4].

Previous research on sentiment analysis, such as "Sentiment Analysis on Maxim Application Reviews on Google Play Store using K-Nearest Neighbor," yielded accuracies, precisions, and recalls of 90.23%, 90.23%, and 72.38%, respectively, using 90% training data and 10% testing data with $k = 5$ [5].

Similarly, research with titled "Sentiment Analysis of Gojek Application Using SVM and K-Nearest Neighbor" obtained accuracies using the KNN method with $k = 22$ resulting in 82.14% accuracy, 82.28% precision, and 95.43% recall, while SVM method with linear kernel and parameter $C = 1$ achieved accuracies of 87.98%, 88.55%, and 95.43% respectively [6].

Furthermore, in sentiment analysis regarding the use of E-Wallet on Google Play using Lexicon-Based and K-Nearest Neighbor, the highest accuracies were observed for Dana at 78% with $k = 6$, Ovo at 75.33% with $k = 9$, and LinkAja at 73.5% with $k = 8$ [7].

Another study on titled "Sentiment Analysis on User Reviews of Bibit and Bareksa Applications using KNN Algorithm" yielded accuracies, precision, and recall for Bibit at 85.14%, 91.91%, and 76.44% respectively, and for Bareksa at 81.70%, 87.15%, and 75.73% respectively [8].

The same study on sentiment analysis regarding the Implementation of the K-Nearest Neighbor Method (K-NN) for Analyzing User Satisfaction Sentiments of the Financial Technology Application FLIP shows that 77.67% of the test data was correctly classified into the positive review class with high precision and recall values of 82.67% and 86.92%, respectively [9].

Based on previous research, it's evident that the K-Nearest Neighbor classification method yields good accuracy with a high level of precision and recall. Therefore, it can be utilized in this study, with the determination of the optimal K value based on available data to reduce noise effects [10].

Given the provided background explanation, the author is interested in conducting research titled "Sentiment Analysis on User Reviews of Digital Population Identity Application on Google Play Store Using K-Nearest Neighbor (KNN) Method."

II. LITERATURE REVIEW

A. Definition of Analysis

Analysis is the process of investigating an event to gain a deep understanding of it (essays, actions, etc.) to determine the actual circumstances [11]. Analysis involves grouping points into several parts, examining those parts themselves, and the relationships among them are part of accurate and meaningful understanding [12].

B. Definition of Sentiment Analysis

Sentiment analysis is the process of identifying and classifying sentiments contained within text, also known as opinion mining. It is a field of study that examines how to understand opinions, evaluations, attitudes, and emotions expressed in text [13].

Sentiment analysis is the process of understanding the emotions, perceptions, and judgments perceived by an individual or researcher towards a model, material, or design [14].

C. User Review

User reviews are one of the features provided by the Google Play Store platform that allows users to provide feedback or opinions in the form of ratings and reviews for downloaded applications. There is a review feature that enables users to comment on the applications they use and provide feedback to the application developers. This research creates a system that can determine the ranking of user reviews on the Google Play Store [15].

D. Application

An application is a program created in software on a computer that is designed or created to facilitate a particular job or activity through information flow processes and procedures that meet the needs of information technology infrastructure. [16].

E. Google Play Store

Google Play Store, formerly known as Android Market, is a digital distribution service operated and developed by Google. It is the official app store for the Android operating system that allows users to browse and download apps developed using the Android Software Development Kit (SDK) and released by Google. Google Play Store also acts as a digital media store, offering music, books, movies, and Television shows. Previously they offered Google devices for purchase to launch separate network hardware Translated with DeepL.com (free version) [17].

F. Text Mining

Text Mining is the search for unknown information using automated data extraction from large amounts of unstructured text [18]. The purpose of text mining is to analyze the opinions,

feelings, judgments, attitudes, assessments, feelings of a person to find out whether a topic, service, organization or person is related to a particular activity [19].

G. Text Preprocessing

Preprocessing is the process of preparing the data to fit the classification requirements. This can be done by removing noise, homogenizing word forms, and reducing data size. Thus, a data set is obtained that is ready to be used for the classification process [20].

H. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm is a machine learning algorithm that does not rely on parameters and learns directly. It does not make any assumptions about the distribution of the underlying data. In other words, there is no fixed number of parameters or parameter estimates in the model, regardless of the size of the data [21].

The K-Nearest Neighbor (K-NN) algorithm is a core example-based classification method that does not make explicit declarations of categories but relies on those categories class entries in training documents are similar to test documents. In the process of testing documents, the system finds k nearest neighbors among the training documents. The average similarity between the document and the nearest neighbors of the test document is used as the class scale for the neighboring document [22].

The K-Nearest Neighbors (KNN) algorithm is a machine learning algorithm that works by calculating the distance between test data and training data. The training data with the closest distance will be used to determine the prediction of the test data. This algorithm was chosen because it has a high accuracy rate. The k value can be determined based on the highest accuracy rate [23].

I. Python

Python is an object-oriented scripting language. It can be used for software development and can run on various operating systems. Currently, Python is also a popular language in the field of computer science and analytics. This is due to Python's support for libraries that provide data analysis and machine learning functions, data processing tools, and data visualization [24].

Python is said to be a programming language that combines functionality and power with a very clear code syntax and a large and comprehensive standard library of functions. Python supports various programming paradigms, but is not limited to object-oriented programming, imperative programming, and active programming. One of the distinctive features of Python is that the language is dynamically equipped with automatic memory management. Python can be used for many purposes of software development needs and can work on different operating system platforms [25].

In analyzing data, Python has libraries that are often used, namely:

1. NumPy is a library that is used to perform mathematical and scientific calculations.
2. Matplotlib is a library used for data visualization.

3. Pandas is often used to move data when analyzing or removing punctuation data and performing statistical summaries.
4. Scraper or better known Scraping is a tool or library in Python that is used to collect data from the Goggle Play Store.





J. Flowchart

A flowchart is a diagram used to explain the steps of a process. This diagram uses standard symbols that represent the steps. The steps are connected with arrows to show their order [26].

flow chart is a diagram that depicts the sequential steps of instructions in a system. This diagram is used by system analysts to explain the logical description of the system to be built to programmers [27].

Here are the flowchart symbols used has been shown in TABLE I.

TABLE I. Flowchart Symbol

No	Symbol	Information
1.		Flow Direction Symbol/ Connecting Line; Is a symbol that serves to connect one symbol with another and also states the flow of a process.
2.		Terminal; functions to start or end the program.
3.		Input/ Output; serves to declare inputs and outputs regardless of their type.
4.		Processing; serves as a pointer to the processing that will be done in the computer.

III. RESEARCH METODOLOGY

A. Research Method

The research method used by the author in this study uses quantitative methods. Quantitative methods are a type of method that helps writers solve problems that produce statistical data in a logical step-by-step manner, showing mathematical processes help find problem solutions and present information in the form of numbers [28].

In this study the authors used review data from the Google Play Store by taking 5000 review data from September 2022 - December 2023. The author also uses several methods to obtain information in solving problems.

B. Data Collection Method

In order to get meaningful information, the author needs data that can be used in this study. The data collection methods carried out by the author are as follows:

- 1) Primary data, is data that is obtained directly from the data source to be used. So the primary data used in this study are user reviews of the Digital Population Identity application on the Google Play Store as much as 5,000 Indonesian-language review data using the Scrapping technique through Python using Google Colab. User reviews contain responses or comments from users who have downloaded the Digital Population Identity application which consists of various comments including positive, neutral and negative comments.
- 2) Secondary data is obtained from existing sources to support the author in conducting research. Secondary data can be in the form of literature studies such as from books, journals and scientific articles or social media that can be accessed using the internet.

C. Analysis Method

In this research, the author uses the sentiment analysis method with K-Nearest Neighbors (KNN) as the machine learning method. The K-Nearest Neighbors algorithm is a supervised learning algorithm in which the results of new cases are classified based on the category of K nearest neighbors. The purpose of this algorithm is to classify objects based on attributes and training data samples [5].

D. Testing Method and Data Processing

1) Testing Method

The author uses Confusion Matrix to evaluate the performance of the K-Nearest Neighbors algorithm. In general, there are several Confusion Matrix indicators, including::

- a. Accuracy is the closeness of the actual value obtained. Here is the formula for determining accuracy:

$$Acc = \frac{TN + TP}{TN + FN + TP + FP}$$

definition:

- TN = True Negative
- TP = True Positive
- FN = False Negative
- TN = True Negative

- b. Precision is a level of compatibility of the data taken with existing information. Here is the formula for determining precision:

$$Precision = \frac{TP}{TP + FP}$$

- c. Recall is the degree of closeness of the actual value to the value obtained. The following is the formula for calculating Recall:

$$Recall (N) = \frac{TP}{TN + FN}$$

$$Recall (P) = \frac{TP}{TN + FP}$$

- d. F1 Score is a comparison of the weighted average

precision and recall. The following is the formula for calculating F1 Score:

$$F-1 \text{ Score} = 2X \frac{\text{Recall} \times \text{Presisi}}{\text{Recall} + \text{Presisi}}$$

Table II described the confusion matrix illustration

TABLE II. Confusion Matrix

Actual	Predicted Condition	
	Positive	Negative
Positive	(TP)	(FN)
Negative	(FP)	(TN)

2) *Data Processing*

In conducting this research, processing is carried out on user review data with the stages presented in the following figure 1:

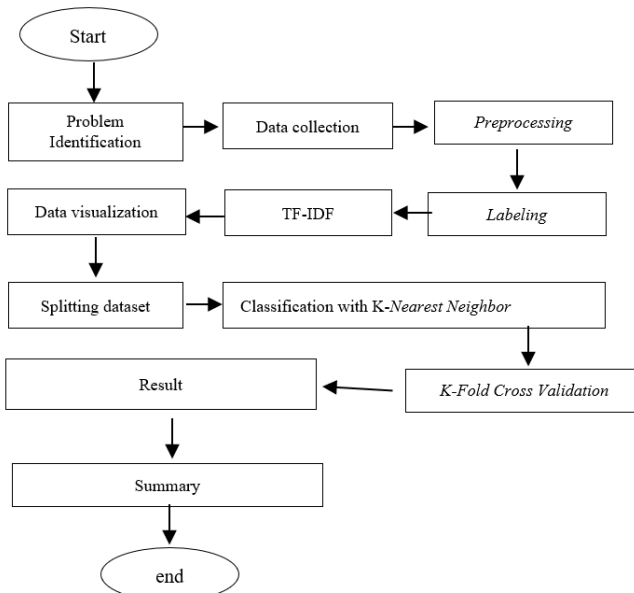


Fig 1. Research Framework

a. *Problem Identification*

Based on the increasing usage of the Digital Population Identity Application on Google Play Store, various user reviews of the application have emerged. Therefore, sentiment analysis is needed to provide a deeper understanding of user perceptions and to classify user reviews of the Digital Population Identity Application..

b. *Preprocessing Stage*

During the data processing phase, preprocessing was conducted using the Python programming language, which consisted of several stages such as retrieving user reviews data of the Digital Population Identity Application from the Google Play Store through Web Scraping using Python. After obtaining the data, the user reviews were then processed using Python libraries. The following are the preprocessing stages:

1. Case Folding is a step where review text or strings are converted to uppercase or lowercase to make the entire text lowercase.
2. Remove Punctuation is the process of removing components contained in user reviews in the form of

characters or punctuation marks.

3. Tokenizing is the process of breaking down a set of characters in a text into units of words.
4. Normalization involves changing and correcting abbreviations into correct words that have the same meaning according to KBBI so that information can be processed easily, for example "utk" becomes "for", "yg" becomes "yang", and so on. The normalization stage uses a collection of words in the form of a CSV (Comma Delimited) file.
5. Stopword Removal is the process of taking important words from the results obtained at the tokenizing stage by removing stopwords or stoplists [8].
6. Stemming is the process of changing each word to its basic form in accordance with the rules of the Big Indonesian Dictionary (KBBI) [5]. At the Stemming stage, it is carried out by installing the Sastrawi dictionary which will change into a standard language form without affixes into a basic word form.

c. *Labelling*

Labelling of data is done to analyze a word or sentence whether it falls into the Positive, Neutral or Negative category.

d. *TF-IDF Weighting Phase*

The TF-IDF method is a method for determining the relative frequency of each word or character where each word is given a weight according to its importance in the form of a value whether the word is important or not, based on the number of occurrences of the word in a document and comparing these words with all existing documents [8].

e. *Cosine Similarity*

Cosine similarity is one method that can be used to measure the similarity between two documents. This method works by converting the document into vector form and then calculating the angle between the two vectors..

f. *Data Visualization*

After labeling, the data will be converted into chart form for more attractive data visualization using the Matplotlib library. Matplotlib is a cross-platform Python library for creating high-quality 2D plots. Matplotlib makes it easy to create graphs, histograms, spectra, bar charts with ease..[25]

g. *Data Splitting*

Data that has been preprocessed and labeled is then divided between training data and test data. Training data is data for creating classification models, while test data is used for testing existing models.

h. *Classification*

The K-Nearest Neighbors algorithm functions to classify data based on training data and test data taken from the K nearest neighbors. After weighing the data, the next step is the classification process. This classification process uses the K-Nearest Neighbors library with python.

i. *K-Fold Cross Validation*

The K-Folds Cross-Validation method divides the term-weighted data into K parts, where K is the size of the terms. K trials are performed in a set of experiments, using one part as test data [29].

j. *Result*

Perform data processing based on the research that has been done and then analyze the results.

k. *Summary*

The conclusion is the final result taken based on the research that has been done.

IV. RESULT AND DISCUSSION

In this research, the implementation was done using Python and Google Colab. To start a Python program in Google Colab, a library is required. A library is a collection of code that can be used repeatedly in different programs. When performing sentiment analysis with Python, some libraries that must be installed are csv, pandas, numpy, sastrawi, scikit-learn or sklearn. Library installation can be done using the "!pip install" command. Once installed, the library can be imported by using the "import" command.

To get the results in this study, several stages are needed, such as the data collection stage by scrapping review data, the preprocessing stage, the labelling, classification and evaluation stages.

1. *Data Collection*

In the data collection stage, the author uses the scraper library in the Python programming language. To use the library, the first step is to install it on Google Colab. Then install the library with the source code !pip install google-play-scraper. After that, the scrapping stage can be carried out or the process of retrieving user review data on the Digital Population Identity application from the Google Play Store which consists of several attributes that will be used such as user name, score, time and user reviews.

In this study the authors used 5000 of the most relevant user review data. Review data collected from September 2022 - December 2023. Furthermore, the results of data collection will be downloaded and stored in csv (comma delimited) file format to be used in the next stage. The period from September 2022 to December 2023 was likely selected to capture a broad range of user experiences over time. This timeframe allows researchers to observe any temporal trends or changes in user perceptions of the Digital Population Identity (DPI) application.

2. *Preprocessing Phase*

The next stage is Preprocessing, which is a stage to prepare data so that it can be used in the next step. The following is the Preprocessing stage:

a. *Case Folding*

In the Case Folding stage, the review data is processed to convert letters into lowercase letters. Figure 2 below presents the case folding process.

	score	content	case folding
0	1	Katanya ktp digital, ak mau buat sendiri aja h...	katanya ktp digital, ak mau buat sendiri aja h...
1	1	Yg di sayangkan knp harus datang ke dukcapil, ...	yg di sayangkan knp harus datang ke dukcapil, ...
2	1	Kurang sip punya ku lupa pin pun gak bisa di b...	kurang sip punya ku lupa pin pun gak bisa di b...
3	1	aplikasinya ga bisa di buka,keterangan pada ap...	aplikasinya ga bisa di buka,keterangan pada ap...
4	1	Bukanya data e-KTP sudah terekam kok yg ini ma...	bukanya data e-ktp sudah terekam kok yg ini ma...
...
4995	5	Mantap aplikasi ini, klo semuanya udah pakai a...	mantap aplikasi ini, klo semuanya udah pakai a...
4996	5	Aplikasi sangat berguna membawa dokumen tidak ...	aplikasi sangat berguna membawa dokumen tidak ...
4997	5	Mantap. Aplikasi yang sangat membantu. Sukses ...	mantap. aplikasi yang sangat membantu. sukses ...
4998	5	ga perlu pusing mikirin ktp kapan jadi. sekara...	ga perlu pusing mikirin ktp kapan jadi. sekara...
4999	5	Mantap, tp agak lama loginnya, mudah2an ke dep...	mantap, tp agak lama loginnya, mudah2an ke dep...

Fig 2. Case Folding Process

b. *Remove Punctuation*

In the Remove Punctuation stage, numbers and characters such as punctuation marks are removed from user reviews. Figure 3 below presents the results of the remove punctuation process.

	case folding	remove punctuation
	katanya ktp digital, ak mau buat sendiri aja h...	katanya ktp digital ak mau buat sendiri aja ha...
	yg di sayangkan knp harus datang ke dukcapil, ...	yg di sayangkan knp harus datang ke dukcapil u...
	kurang sip punya ku lupa pin pun gak bisa di b...	kurang sip punya ku lupa pin pun gak bisa di b...
	aplikasinya ga bisa di buka,keterangan pada ap...	aplikasinya ga bisa di bukaketerangan pada apk...
	bukanya data e-ktp sudah terekam kok yg ini ma...	bukanya data e-ktp sudah terekam kok:yg ini ma...
...
	mantap aplikasi ini, klo semuanya udah pakai a...	mantap aplikasi ini klo semuanya udah pakai ap...
	aplikasi sangat berguna membawa dokumen tidak ...	aplikasi sangat berguna membawa dokumen tidak ...
	mantap. aplikasi yang sangat membantu. sukses ...	mantap aplikasi yang sangat membantu sukses se...

Fig 3. Remove Punctuation Process

c. *Tokenizing*

In the Tokenizing stage, tokenization is carried out or separating sentences into word units. Figure 4 below presents the Tokenizing process.

	remove punctuation	tokenisasi
	katanya ktp digital ak mau buat sendiri aja ha...	[katanya, ktp, digital, ak, mau, buat, sendiri...
	yg di sayangkan knp harus datang ke dukcapil u...	[yg, di, sayangkan, knp, harus, datang, ke, du...
	kurang sip punya ku lupa pin pun gak bisa di b...	[kurang, sip, punya, ku, lupa, pin, pun, gak, ...
	aplikasinya ga bisa di bukaketerangan pada apk...	[aplikasinya, ga, bisa, di, bukaketerangan, pa...
	bukanya data e-ktp sudah terekam kok:yg ini ma...	[bukanya, data, e-ktp, sudah, terekam, kok, yg...
...

Fig 4. Tokenizing Process

d. *Normalization*

At the Normalization stage is the process of changing and correcting abbreviations into complete words and have the same meaning according to the Big Indonesian Dictionary (KBBI). at this stage using a list with a CSV file format containing the form of words that have been corrected. Figure 5 below presents the results of the normalization process.

tokenisasi	normalisasi
[katanya, ktp, digital, ak, mau, buat, sendri...]	[katanya, ktp, digital, saya, mau, buat, sendri...]
[yg, di, sayangkan, knp, harus, datang, ke, du...]	[yang, di, sayangkan, kenapa, harus, datang, k...]
[kurang, sip, punya, ku, lupa, pin, pun, gak, ...]	[kurang, sip, punya, ku, lupa, pin, pun, tidak...]
[aplikasinya, ga, bisa, di, bukaketerangan, pa...]	[aplikasinya, tidak, bisa, di, bukaketerangan, ...]
[bukanya, data, e-ktp, sudah, terekam, kok, yg...]	[bukanya, data, e-ktp, sudah, terekam, kok, ya...]

Fig 5. Normalization Process

e. Stopword Removal

In the stopword removal step, words that are not important in the document are removed. The Stopword Removal stage uses a list in the form of a CSV file containing a list of stopword of words. Figure 6 below presents the Stopword Removal process.

normalisasi	stopword removal
[katanya, ktp, digital, saya, mau, buat, sendri...]	[ktp, digital, scan, barcod, kedukcapil, bamba...]
[yang, di, sayangkan, kenapa, harus, datang, k...]	[sayangkan, dukcapil, barcode, inovasi, djp, o...]
[kurang, sip, punya, ku, lupa, pin, pun, tidak...]	[sip, ku, lupa, pin, buka, aturannya, mudah, n...]
[aplikasinya, tidak, bisa, di, bukaketerangan, ...]	[aplikasinya, bukaketerangan, apkterjadi, kesa...]
[bukanya, data, e-ktp, sudah, terekam, kok, ya...]	[bukanya, data, e-ktp, terekam, disuruh, wara...]

Fig 6. Stopword Removal Process

f. Stemming

In the Stemming stage, each word is converted into the basic form of the word using the Literary Library. Figure 7 presents the results of the stemming process.

normalisasi	stopword removal	stemming
[katanya, ktp, digital, saya, mau, buat, sendri...]	[ktp, digital, scan, barcod, kedukcapil, bamba...]	[ktp, digital, scan, barcod, kedukcapil, bamba...]
[yang, di, sayangkan, kenapa, harus, datang, k...]	[sayangkan, dukcapil, barcode, inovasi, djp, o...]	[sayang, dukcapil, barcode, inovasi, djp, onli...]
[kurang, sip, punya, ku, lupa, pin, pun, tidak...]	[sip, ku, lupa, pin, buka, aturannya, mudah, n...]	[sip, ku, lupa, pin, buka, atur, mudah, nengok...]
[aplikasinya, tidak, bisa, di, bukaketerangan, ...]	[aplikasinya, bukaketerangan, apkterjadi, kesa...]	[aplikasi, bukaketerangan, apkterjadi, salah, ...]
[bukanya, data, e-ktp, sudah, terekam, kok, ya...]	[bukanya, data, e-ktp, terekam, disuruh, wara...]	[buka, data, e-ktp, rekam, suruh, wara-wiri, k...]

Fig 7. Stemming Process

After the Preprocessing stage, the data can be saved into a CSV (Comma Delimited) file format to be used in the next stage.

g. Labelling

Labeling is used to assign a class to each data. Figure 8 below presents the labeling process. Labeling of data is done to analyze a word or sentence whether it falls into the Positive, Neutral, or Negative category. In the dictionary-Based Approach, This technique involves using predefined dictionaries or lexicons containing words associated with positive, neutral, or negative

sentiment. Each word in the text is matched against entries in the dictionaries, and its sentiment category is determined based on the presence and context of these words.

	score	stemming	label
0	1	['ktp', 'digital', 'scan', 'barcod', 'kedukcap...]	negatif
1	1	['sayang', 'dukcapil', 'barcode', 'inovasi', '...]	negatif
2	1	['sip', 'ku', 'lupa', 'pin', 'buka', 'atur', '...]	negatif
3	1	['aplikasi', 'bukaketerangan', 'apkterjadi', '...]	negatif
4	1	['buka', 'data', 'e-ktp', 'rekam', 'suruh', 'w...]	negatif
...
4995	5	['mantap', 'aplikasi', 'pakai', 'aplikasi', 'p...]	positif
4996	5	['aplikasi', 'guna', 'bawa', 'dokumen', 'ribet...]	positif
4997	5	['mantap', 'aplikasi', 'bantu', 'sukses']	positif
4998	5	['pushing', 'mikirin', 'ktp', 'identitas', 'dud...]	positif
4999	5	['mantap', 'loginnya', 'mudah', 'ngacir', 'apl...]	positif

5000 rows x 3 columns

Fig 8. Labelling Result

h. TF-IDF Weighting (Term Frequency – Inverse Document Frequency)

At this stage, the weight of each frequently occurring word will be calculated.

i. Calculating Cosine Similarity

Cosine similarity measures the similarity between documents by converting the documents into vector form and then calculating the angle between the two vectors. Figure 9 below presents the calculation results of cosine similarity. In the context of sentiment analysis, cosine similarity can be used to compare the sentiment of different user reviews or documents. By computing the cosine similarity between vectors representing the sentiment of each document, researchers can assess how closely related or similar the sentiments expressed in the reviews are.

```
PROSES COSINE SIMILARITY

[ ] from sklearn.metrics.pairwise import cosine_similarity

[ ] #Compute similarity using cosine similarity
cos_sim=cosine_similarity(text_tf, text_tf)

print(cos_sim)

[[[1. 0. 0.02889552 ... 0. 0.11952169 0.03933563]
 [0. 1. 0. ... 0. 0. 0. ]
 [0.02889552 0. 1. ... 0. 0.0880262 0.03561176]
 ...
 [0. 0. 0. ... 1. 0. 0.28735186]
 [0.11952169 0. 0.0880262 ... 0. 1. 0. ]
 [0.03933563 0. 0.03561176 ... 0.28735186 0. 1. ]]]
```

Fig 9. Cosine Similarity Process

j. Data Visualization

After the review data has gone through the labeling stage, data visualization will then be carried out to make it easier to understand and provide an attractive appearance. At this visualization stage, the Matplotlib library is used which will visualize into a diagram to see the amount of data with positive, neutral and negative labels. Figure 10 presents the distribution of data in each class.

TABLE V. Result of ACC, Precision, Recall, and F-1 Score (Data Split 70:30)

K	Data Split (70:30)			
	Accuracy	Precision	Recall	F-1 Score
1	0.75	0.97	0.85	0.85
2	0.74	0.74	0.99	0.85
3	0.77	0.82	0.90	0.86
4	0.79	0.80	0.95	0.87
5	0.79	0.82	0.94	0.87
6	0.80	0.81	0.96	0.88
7	0.80	0.82	0.95	0.88
8	0.80	0.81	0.95	0.88
9	0.80	0.81	0.95	0.88
10	0.80	0.81	0.95	0.88

At a ratio of 70% Training Data: 30% Test Data, there are the highest results at K = 6, 7, 8, 9 and 10 which produce an accuracy of 80%, a precision value of 81%, a recall of 95% and an F1-Score value of 88%. Evaluation of the KNN model testing at the value of K = 10 using the confusion matrix is presented in Figure 13.

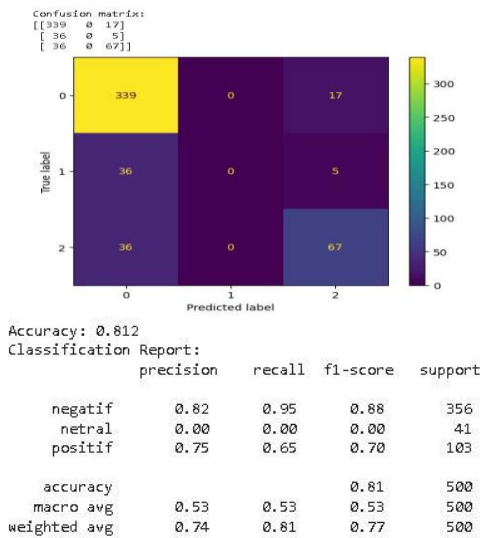


Fig 13. Model Evaluation (Split Data 70:30)

From the classification results using the K-Nearest Neighbors method using a value of K = 10 with two ratios of comparison between training data and test data on data division 90: 10 and 70: 30 there is data imbalance in the Neutral review class. Data imbalance is a situation where there is a significant difference between the number of samples from different classes or groups in a data set. This can be a problem in machine learning as models trained on imbalanced data sets can tend to prioritize the majority class and ignore the minority class.

5. Evaluation with K-Fold Cross Validation

The K-Folds Cross-Validation method divides the data that has been weighted by terms into equal parts. The number of parts is called K. The machine learning model is then trained using K parts of the data, and evaluated using the remaining parts of the data.

This process is done K times, using different parts of the data each time. In this research, we will use a value of K=10,

which means that the dataset will be divided into 10 parts or folds to perform the K-Fold Cross Validation test for 10 repetitions. Table VI presents the results of the evaluation using K-Fold CV.

TABLE VI. Evaluation With K-Fold Cross Validation

K-Fold	Accuracy	Precision	Recall	F-1 Score
1	0.81	0.81	0.97	0.88
2	0.82	0.83	0.97	0.89
3	0.81	0.82	0.95	0.88
4	0.79	0.81	0.94	0.87
5	0.80	0.81	0.96	0.88
6	0.81	0.82	0.96	0.88
7	0.79	0.80	0.94	0.87
8	0.82	0.84	0.96	0.89
9	0.78	0.81	0.92	0.86
10	0.84	0.84	0.98	0.90
	0.80			

After testing K-Fold Cross Validation using the value of K = 10, 10 tests were carried out, the average accuracy value was 0.80. Figure 14 below presents the results of the evaluation using K-Fold CV.

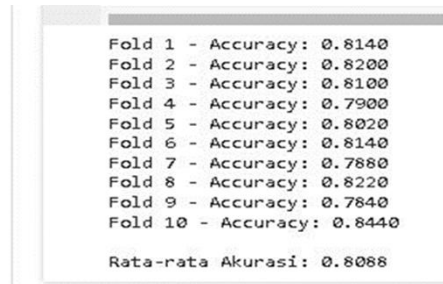


Fig 14. Evaluation with K-Fold CV

The evaluation results show that the tested model achieved an average accuracy of 0.80 or 80%. This means that the model can correctly predict about 80% of the data that has never been trained before.

V. CONCLUSION

From the classification results using the K-Nearest Neighbors method, the best results were obtained with a ratio of 90% training data:10% test data. The best results were obtained at the values of K = 8, K = 9, and K = 10, with Accuracy of 81%, Precision of 82%, Recall of 95%, and F1-Score value of 88%. At a ratio of 70% training data: 30% test data, the best results were obtained at values K = 6, K = 7, K = 8, K = 9, and K = 10, with Accuracy of 80%, Precision of 81%, Recall of 95%, and F1-Score value of 88%.

After the labeling stage produces 3 classes, namely Positive, Negative and Neutral Sentiment classes. The results are 3581 Negative Sentiments, 1031 Positive Sentiments and 388 Neutral Sentiments. From the review data, it explains that user perceptions of the Digital Population Identity Application are dominated by negative reviews.

REFERENCE

- [1] E. Islami And J. P. Islam, "Pemanfaatan Website Sebagai Bentuk Digitalisasi Pelayanan Islam," Pp. 1167–1182, 2022, Doi: 10.30868/Ei.V11i01.2979.
- [2] G. Vinodhini, "Sentiment Analysis And Opinion Mining : A Survey International Journal Of Advanced Research In Sentiment Analysis And Opinion Mining : A Survey," No. June 2012, 2014.
- [3] N. Faridhotun, E. Haerani, And ..., "Analisis Sentimen Ulasan Aplikasi Wetv Untuk Peningkatan Layanan Menggunakan Metode K-Nearest Neighbor," *J. Inf.*, Vol. 4, No. 3, Pp. 855–864, 2023, Doi: 10.47065/Josh.V4i3.3349.
- [4] M. Farid And E. Firdaus, "Analisis Sentimen Tokopedia Pada Ulasan Di Google Playstore Menggunakan Algoritma Naive Bayes Classifier Dan K-Nearest Neighbor," Vol. 9, No. 5, Pp. 1329–1336, 2022, Doi: 10.30865/Jurikom.V9i5.4774.
- [5] R. Ramadhan, M. Afdal, I. Permana, And M. Jazman, "Analisis Sentimen Pada Ulasan Aplikasi Maxim Di Google Play Store Dengan K-Nearest Neighbor," Vol. 10, No. 3, 2023, Doi: 10.30865/Jurikom.V10i3.6396.
- [6] M. N. Shariff, "Tata Kelola Penerapan Teknologi Informasi Pengelolaan Pajak Di Dppkab Kab. Okl Menggunakan Framework Cobit 5," *Semin. Nas. Teknol. Inf. Dan Komun. X*, Pp. 353–358, 2018.
- [7] N. Habibah, E. Budianita, M. Fikry, And I. Iskandar, "Analisis Sentimen Mengenai Penggunaan E-Wallet Pada Google Play Menggunakan Lexicon Based Dan K-Nearest Neighbor," *J. Ris. Komputer*, Vol. 10, No. 1, Pp. 2407–389, 2023, Doi: 10.30865/Jurikom.V10i1.5429.
- [8] S. Rahayu, Y. Mz, J. E. Bororing, And R. Hadiyat, "Implementasi Metode K-Nearest Neighbor (K-Nn) Untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi Finansial Flip," *Edumatic J. Pendidik. Inform.*, Vol. 6, No. 1, Pp. 98–106, 2022, Doi: 10.29408/Edumatic.V6i1.5433.
- [9] A. D. Adhi Putra, "Analisis Sentimen Pada Ulasan Pengguna Aplikasi Bibit Dan Bareksa Dengan Algoritma Knn," *Jatiji (Jurnal Tek. Inform. Dan Sist. Informatika)*, Vol. 8, No. 2, Pp. 636–646, 2021, Doi: 10.35957/Jatiji.V8i2.962.
- [10] J. Homepage, S. R. Cholil, T. Handayani, R. Prathivi, And T. Ardianita, "Ijcit (Indonesian Journal On Computer And Information Technology) Implementasi Algoritma Klasifikasi K-Nearest Neighbor (Knn) Untuk Klasifikasi Seleksi Penerima Beasiswa," *Ijcit (Indonesian J. Comput. Inf. Technol.)*, Vol. 6, No. 2, Pp. 118–127, 2021.
- [11] Y. Yadi, "Analisa Usability Pada Website Traveloka," *J. Ilm. Betrik*, Vol. 9, No. 03, Pp. 172–180, 2018, Doi: 10.36050/Betrik.V9i03.43.
- [12] B. I. Mayang Milasari, Cindi Wulandari, "Analisis Tingkat Kualitas Layanan Sub Menu Pengecekan Data Penerima Bantuan Sosial Pada Website Dtkas Dinas Sosial Kota Lubuklinggau Menggunakan Metode E-Govqual," Pp. 1389–1397, 2022.
- [13] M. Riski, M. Fikry, And Yusra, "Klasifikasi Sentimen Ulasan Aplikasi Whatsapp Di Play Store Menggunakan Metode K-Nearest Neighbor," *Media Online*, Vol. 4, No. 1, Pp. 438–444, 2023, Doi: 10.30865/Klik.V4i1.1050.
- [14] S. Hasanah, I. Purwasih, And ..., "Analisis Sentimen Terhadap Masyarakat Adanya Uang Kertas Baru Menggunakan Algoritma K-Nearest Neighbor (Knn)," *Ikra-lth Inform.*, Vol. 7, No. 2, Pp. 105–114, 2023, [Online]. Available: [Http://journals.upi-yai.ac.id/index.php/ikraith-informatika/article/view/2813%0ahttps://journals.upi-yai.ac.id/index.php/ikraith-informatika/article/download/2813/2065](http://journals.upi-yai.ac.id/index.php/ikraith-informatika/article/view/2813%0ahttps://journals.upi-yai.ac.id/index.php/ikraith-informatika/article/download/2813/2065)
- [15] Amalia Elma Sari, "Klasifikasi Ulasan Pengguna Aplikasi Mandiri Online Di Google Play Store Dengan Menggunakan Metode Information Gain Dan Naive Bayes Classifier," 2019. <https://openlibrary.telkomuniversity.ac.id/pustaka/152491/klasifikasi-ulasan-pengguna-aplikasi-mandiri-online-di-google-play-store-dengan-menggunakan-metode-information-gain-dan-naive-bayes-classifier.html>
- [16] Elmayati, "Elmayati Aplikasi Sistem Informasi Pengajuan Beasiswa Berbasis Web Pada Sekolah Tinggi Manajemen Dan Ilmu Komputer Musi Rawas (Stmik-Mura) Kota Lubuklinggau Jusim , Vol 1 No. 1 , Desember 2016," *Jusim*, Vol. 1, No. 1, Pp. 8–18, 2016, [Online]. Available: [Http://jurnal.univbinainsan.ac.id/index.php/jusim/article/download/26/190](http://jurnal.univbinainsan.ac.id/index.php/jusim/article/download/26/190)
- [17] Wikipedia, "Google Play Store." [Online]. Available: https://id.wikipedia.org/wiki/Google_Play
- [18] A. Hermawan, I. Jowensen, J. Junaedi, And Edy, "Implementasi Text-Mining Untuk Analisis Sentimen Pada Twitter Dengan Algoritma Support Vector Machine," *Jst (Jurnal Sains Dan Teknol.)*, Vol. 12, No. 1, Pp. 129–137, 2023, Doi: 10.23887/jstundiksha.V12i1.52358.
- [19] F. Prasetya And F. Ferdiansyah, "Analisis Data Mining Klasifikasi Berita Hoax Covid 19 Menggunakan Algoritma Naive Bayes," *J. Sist. Komput. Dan Inform.*, Vol. 4, No. 1, P. 132, 2022, Doi: 10.30865/Json.V4i1.4852.
- [20] J. K. Antartika, "Analisis Sentimen Pada Ulasan Aplikasi Home Credit Dengan Metode Svm Dan Knn," Vol. 1, Pp. 174–181, 2023.
- [21] Adminlp2m, "Algoritma K-Nearest Neighbors (Knn) – Pengertian Dan Penerapan," 2023, [Online]. Available: <https://lp2m.uma.ac.id/2023/02/16/algoritma-k-nearest-neighbors-knn-pengertian-dan-penerapan/>
- [22] J. Informatika And S. Informasi, "Informasi (Jurnal Informatika Dan Sistem Informasi) Volume 15 No.1 / Mei / 2023," Vol. 15, No. 1, Pp. 1–17, 2023.
- [23] D. Sartika, "Implementasi Algoritma K-Nearest Neighbour Dalam Menganalisis Sentimen Terhadap Program Merdeka Belajar Kampus Merdeka (Mbkmm)," Pp. 69–76, 2020.
- [24] D. A. Manalu And G. Gunadi, "Implementasi Metode Data Mining K-Means Clustering Terhadap Data Pembayaran Transaksi Menggunakan Bahasa Pemrograman Python Pada Cv Digital Dimensi," *Infotech J. Technol. Inf.*, Vol. 8, No. 1, Pp. 43–54, 2022, Doi: 10.37365/jti.V8i1.131.
- [25] F. David, P. Studi, T. Informatika, And F. Teknologi, "Visualisasi Data Dalam Bentuk 3 Dimensi Dengan Abstrak Seminar Nasional Pimimd-5 , Itp , Padang," Pp. 1–6, 2019, Doi: 10.21063/pimimd5.2019.1.
- [26] B. I. Pendahuluan, "Panduan Pembuatan Flowchart," 2017.
- [27] R. Rosaly, "Pengertian Flowchart Beserta Fungsi Dan Simbol-Simbol Flowchart Yang Paling Umum Digunakan".
- [28] S. Silvillestari, "Data Mining Menggunakan Algoritma K-Nearest Neighbor Dalam Menentukan Kredit Macet Barang Elektronik," *J. Media Inform. Budidarma*, Vol. 5, No. 3, P. 1063, 2021, Doi: 10.30865/Mib.V5i3.3100.
- [29] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, And W. Gata, "Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi," *J. Teknoinfo*, Vol. 14, No. 2, P. 115, 2020, Doi: 10.33365/jti.V14i2.679.