ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 1, March 2023

ALTERNATIVE TEXT PRE-PROCESSING USING CHAT GPT OPEN AI

Indri Tri Julianto[1], Dede Kurniadi[2], Yosep Septiana[3], Ade Sutedi[4]

[1-4] Department of Computer Science, Institut Teknologi Garut

email: indritrijulianto@itg.ac.id[1], dede.kurniadi.ac.id[2], yseptiana@itg.ac.id[3], adesutedi@itg.ac.id

Abstract

Text Pre-Processing is the first step in Sentiment Analysis. Categorizing a sentiment in a dataset is part of the Text-Preprocessing stage to get the optimal model accuracy value. Generative Pretrained Transformer, often known as Chat GPT, is a Machine Learning model that can automatically generate realistic and meaningful text. This study aims to examine the capability of GPT Chat as an alternative in the Text-Pre-Processing stage by utilizing GPT Chat 3 from the openai.com website in the Text-Pre-Processing stage of the collected tweet data. The data used in this research is the result of crawling Twitter by inserting the keyword "Chat GPT". This study method was carried out by measuring performance using the K-Nearest Neighbor and Naïve Bayes Algorithms to find the best performance value and compare it with the Text-Preprocessing generated by Rapidminer. It is shown that the performance accuracy produced using the K-Nearest Neighbor Algorithm is 73.57% using the Linear Sampling method. The comparison result with the Text-Preprocessing method using Rapidminer indeed shows a better accuracy of 75.33%, which means it has a narrow difference of 1.76% with the Chat GPT Text Pre-Processing method. However, both are still in the same category, which is Fair Classification. The results of this research show that Chat GPT can be an alternative in Text-Preprocessing datasets for sentiment analysis.

Keywords : Algorithm, Chat GPT, K-Nearest Neighbour, Naïve Bayes, Text Pre-Processing

Received: 16-03-2023 | **Revised:** 30-03-2023 | **Accepted:** 31-03-2023
DOI: https://doi.org/10.23887/janapati.v12i1.59746

INTRODUCTION

Twitter is a social media that is still popular with the public because it has extensive connection capabilities [1]. Through Twitter, users can express their emotions towards something according to their wishes. The content can be opinions, sentiments, or emoticons [2]. The results of these tweets can be data to analyze trends or specific topics. Twitter provides an application programming interface (API) to collect sentiment data. There are two types of API available, namely RESTAPI and Streaming API. RESTAPI is used to access the user's status and timeline, while the Streaming API is used to access keywords, hashtags, user ID and location [3].

Crawling data from Twitter cannot be used directly use crawling data from Twitter in the sentiment analysis process. This happens because Twitter has rules for writing a maximum of 140 characters in one tweet and many non-standard words. After all, people usually devote their tweets to everyday and informal language. [4]. Twitter is a social media platform with more than 140 million active users and more than 400 million iconic Tweets daily [2].

The process of sentiment analysis will show the tendency of a person's opinion on a topic or problem by building a sentiment classification into two or more classes [2]. Sentiment analysis is the process of extracting keywords from user reviews and text classified into positive, negative, and neutral [5], [6]. Sentiment analysis and opinion mining are fields that focus on analyzing people's opinions, sentiments, evaluations, attitudes, and emotions [7]. To perform sentiment analysis, there are various data mining techniques available, and one can still use these techniques to improve the accuracy of the model being built [8].

Chat GPT is an artificial intelligence language model introduced by OpenAI in November 2022. Where this model is the solution with a combination of Reinforcement Learning Algorithm and human input of more than 150 billion parameters [9]. Chat GPT works as if it were a personal assistant for humans to answer questions given in a dialogue format and then this model can be used for various purposes, such as creating automated chats in chat apps, helping with content creation, or even helping translate different languages with

different levels of accuracy for each language [10]. Open AI is a developer of Chat GPT, where Open AI itself is a research laboratory on artificial intelligence located in the United States.

There are several previous studies regarding Text-Preprocessing, and in general these studies use Text-Preprocessing to improve the performance of the resulting model [4], [11]–[14]. Stopword removal, stemming, and feature selection methods can increase performance by 20.4% [15]. Research on preprocessing cleaning, case folding, and stemming techniques without using stopword removal can increase accuracy by 94.24% [16]. Then research on sentiment analysis using the K-Nearest Neighbor Algorithm is applied to the Online Lecture dataset, which is a distance learning medium resulting in an accuracy rate of 56% [2]. The latest research regarding sentiment analysis of online learning from Twitter media using the Naïve Bayes Algorithm produces an accuracy of 76.39% [3].

Previous research on the Text Pre-Processing stage has mostly utilized Rapidminer, as seen in studies [17]–[21] while Python was used in studies [22]–[26]. In order to identify the gaps between previous research and the current study, a table has been presented as shown in Table 1.

Table 1. Research Gap

No	Research	Text Pre-Processing Technique
1	[17]	Rapidminer
2	[18]	Rapidminer
3	[19]	Rapidminer
4	[20]	Rapidminer
5	[21]	Rapidminer
6	[22]	Python
7	[23]	Python
8	[24]	Python
9	[25]	Python
10	[26]	Python
11	Present	Chat GPT & Rapidminer

This research is interested in using Chat GPT as an alternative to Text Pre-Processing to the Twitter dataset for sentiment analysis. This research is essential to do because there is still little research on Chat GPT, and to be able to find out the effect of Text-Pre-Processing using Chat GPT on the performance produced with the K-Nearest Neighbor and Naïve Bayes Algorithms.

This research will fill the gap with previous research by using the Chat GPT Text-Preprocessing alternative in the text cleaning and sentiment labeling sections. Then there will be Tokenizing, Transform Cases, Stopwords Filters, Token Filters (by length), and Stemming. After that, TF-IDF was carried out, and then model testing was carried out using the K-Nearest Neighbor and Naïve Bayes Algorithms by comparing three sampling methods: Linear, Stratified, and Shuffle using the Rapidminer Studio application.

The K-Nearest Neighbour algorithm was selected due to its advantages in handling training data with a high amount of noise, fast training execution, ease of understanding, and ability to handle large amounts of data [27]. This algorithm has been proven effective, as demonstrated in the study "Application of K-Nearest Neighbor Method in Twitter User Sentiment Analysis on the G20 Summit in Indonesia" [28], where the classification results using K-Nearest Neighbour were categorized as Excellent Classification.

The Naïve Bayes algorithm was also chosen for its advantage in the classification process, where it only requires a small amount of training data to select the necessary parameters, and its process is both fast and efficient [29]. The effectiveness of the Naïve Bayes algorithm has been proven in the study "Sentiment Analysis of Brimo Application Reviews on Google Play Using Naive Bayes Algorithm" [30], where the accuracy results were categorized as Good Classification.

METHOD

The proposed research framework is in chart form, as shown in Figure 1.

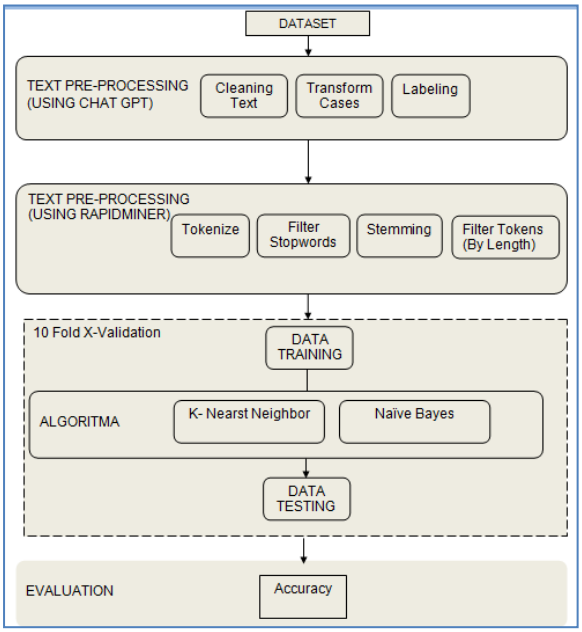


Figure 1. Research Method

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 1, March 2023

The first stage is the dataset. Crawling is a data extraction technique from Twitter that provides an application program interface (API) that is used to access the information contained in it [3]. The dataset was collected by crawling Twitter using the Rapidminer application by entering the keyword “Chat GPT”, here the data was collected was 100 data. The results of the Crawling dataset from Twitter are presented as shown in Figure 2.

Id	Created-At	Language	Source	Retweet-Count	Text
1639155912...	Mar 24, 2023 ...	en	<a href="http...	544	AI tools that didn't exist 1 year ago will save you thousan...
1639299961...	Mar 24, 2023 ...	en	<a href="http:/...	176	□BREAKING □
1639065836...	Mar 24, 2023 ...	en	<a href="http...	195	"How to chat with a 56-page PDF"
1639545629...	Mar 25, 2023 ...	en	<a href="http...	0	is there any service that uses gpt4/chat gpt to summari...
1639545546...	Mar 25, 2023 ...	en	<a href="http:/...	13	RT @kwesi_winfred: With the help of Chat GPT we woul...
1639545300...	Mar 25, 2023 ...	en	<a href="http:/...	1	RT @LCHistoryTutor: My interview with @emma_okelly ...
1639545163...	Mar 25, 2023 ...	en	<a href="http:/...	13	RT @kwesi_winfred: With the help of Chat GPT we woul...
1639545149...	Mar 25, 2023 ...	en	<a href="http:/...	0	@HamishH1931 □Optimus AI - is a project that provide...
1639545028...	Mar 25, 2023 ...	en	<a href="http:/...	0	Domain name for sale:https://t.co/javC891pCx
1639544971...	Mar 25, 2023 ...	en	<a href="http:/...	2	RT @ajemiwa_AO: Day 9 of 60: Interaction Design.
1639544946...	Mar 25, 2023 ...	en	<a href="http:/...	13	RT @kwesi_winfred: With the help of Chat GPT we woul...
1639544929...	Mar 25, 2023 ...	en	<a href="http:/...	0	CHAT GPT isn't better than human beings, it's just bette...
1639544810...	Mar 25, 2023 ...	en	<a href="http:/...	544	RT @simonholdorf: AI tools that didn't exist 1 year ago ...
1639544792...	Mar 25, 2023 ...	en	<a href="http:/...	13	RT @kwesi_winfred: With the help of Chat GPT we woul...

Figure 2. Crawling Dataset from Twitter

Figure 2 explains that there are 6 attributes, namely:

1. Id
2. Created at;
3. Language;
4. Source;
5. Retweet count;
6. Text.

Based on these results, the attribute that will be used is only the Text attribute. Then the data that has been collected does not have sentiment categories, such as Positive, Neutral, and Negative.

The second stage is Text-Preprocessing which is a stage to eliminate problems that can interfere with data processing results [2], [31]. This stage is divided into two activities. The first activity was carried out using the ChatGPT 3. The use of Text-Preprocessing is carried out by visiting the Chat GPT web page, then the gathered data is entered into the chat column provided to perform the following steps:

1. The text cleaning stage is almost always included when preprocessing text because the data is only sometimes structured and consistent. The role of cleaning is to remove punctuation, generalize the use of capital letters, remove duplicate tweet data and correct spelling [2]. Text cleaning to remove characters from Twitter such as, (@, RT,#,link);
2. Then Transform Cases is the process of converting text into all small print or vice versa [32], [33];
3. Labeling sentiment into three categories, namely, Positive, Neutral, and Negative.

The three stages performed using Chat GPT do not require much time as long as the given instructions are sufficiently detailed and clear. The number of datasets used in this research is 100. The time required by Chat GPT is less than one minute. The specifications of the devices used in this research are as follows:

1. Processor Intel Pentium P6000 (1.86 GHz, 3 MB L3 Cache);
2. 3 GB DDR 3 Memory;
3. 320 GB HDD.

These specifications are the minimum requirements, so if the specifications are better, the performance of Chat GPT will be faster. Therefore, if using data with a large number, it can still be processed by Chat GPT.

Then the second activity was carried out using Rapidminer:

1. Tokenize is a process to separate text from documents into squential tokens [5], [8], [34]–[36];
2. Filter Stopwords is a process of removing words that often appear like

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 1, March 2023

("a," "the," "of," "and," and "an") [8], [13], [37];
3. Stemming is a transformation into essential words [3], [38];
4. Filter Token (by length), is the process of deleting words with a certain number of letters with parameters min - chars 4 and max chars 25 to limit the number of letters in words to a minimum of 4 and a maximum of 25 in text using the Rapidminer application [37].

The third stage is Validation. This stage will produce an accuracy value of the model being built. The algorithm used to measure accuracy performance is the K-Nearest Neighbor and Naïve Bayes using the K-Fold Cross Validation operator. Cross-validation is a model validation technique applied to evaluate how the statistical analysis results are generalized into an independent dataset [39]. KVC will partition k parts of data and do as many k iterations. Whenever a part of the dataset is selected, the first k – 1 are used as learning data while the rest are used as testing data. This process will be repeated k-times and then the average deviation (error) value of the k different test results will be calculated . The illustration for KVC with a value of K-10 is presented in the form of pictures, as shown in 3.

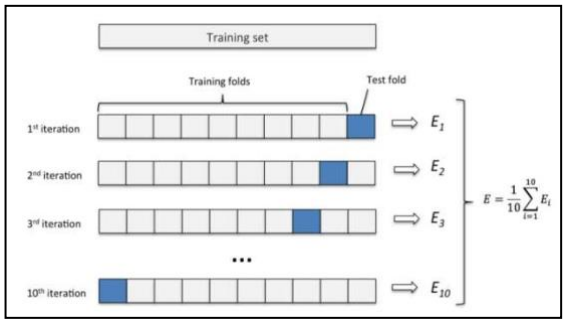


Figure 3. K-Fold Cross Validation

K-Nearest Neighbour Algorithm is often used for classification. The way this algorithm works is grouping data into a class that has been determined based on the closest distance or similarity to the existing data set or training data [40]. The stages of this algorithm are as follows:

1. Determine the value of k;
2. Calculate the distance between the data that will be classified against the label data;
3. Determine the smallest value of k;
4. Classify data based on a distance metric.

Calculation of proximity using a distance matrix can use the following formula:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{21})} \quad (1)$$

Naïve Bayes Algorithm is a classification method derived from the Bayes theorem, which can predict future opportunities based on opportunities that existed in the past [41]. The equation is as follows:

$$P(X|Y) = \frac{P(X|Y)P(X)}{P(Y)} \quad (2)$$

From equation (2) before the occurrence of Y, it can be simplified again. This happens because for each class is always the same value. The equation is as follows:

$$P(X|Y) = P(X|Y)P(X) \quad (3)$$

Mean and Variance are the two parameters used in the gaussian distribution, so for calculations likelihood can use the following equation:

$$\mu = \frac{1}{n}\sum_{n=0}^{N} xn \quad (4)$$

$$\sigma^2 = \frac{1}{n}\sum_{n=0}^{N}(xn - \mu)^2 \quad (5)$$

$$P(Y_1, \ldots, Y_N|\mu, \sigma^2) = \sum_{Y=1}^{Y_N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(xy-\mu)^2}{2\sigma^2}} \quad (6)$$

Note:
n = n data
N = lots of data in 1 feature
Xn = n data value
μ = average value
σ^2 = variance
$Y_{1,\ldots,}Y_N$ = many features 1 to feature N
e = exponential

The fourth stage is Evaluation. Evaluation is carried out on the results of applying the model to determine whether it has fulfilled the research objectives so that a decision is made regarding the use of the modeling results [42]. This stage evaluates the accuracy value produced by the built model. Accuracy is the percentage value of the accuracy of data records that are classified correctly after testing the classification results [43]. Accuracy can be categorized as follows:

1. Accuracy 0.90 – 1.00 = Excellent classification;
2. Accuracy 0.80 – 0.90 = Good classification;
3. Accuracy 0.70 – 0.80 = Fair classification;
4. Accuracy 0.60 – 0.70 = Poor classification;
5. Accuracy 0.50 – 0.60 = Failure.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 1, March 2023

RESULT AND DISCUSSION

The attributes used in this study only use the Text attribute. Any examples of these attributes are presented as shown in Table 2.

Table 2. Preliminary Dataset

No	Text
1	'RT @julia_m_mac: I made Chat GPT my Virtual Assistant
2	'At least Chat GPT lists out biological facts about it https://t.co/hlXeg2hFNp'
3	'RT @julia_m_mac: I made Chat GPT my Virtual Assistant

The dataset is an example of a tweet regarding Chat GPT, where the data needs to be carried out through the Text-Preprocessing stage to become a usable dataset. The Text-Preprocessing carried out using Chat GPT consists of cleaning text, transform cases, and labeling. The results of the Text-Preprocessing are presented as shown in Figure 4, Figure 5, Figure 6, and Figure 7.

Figure 4. Cleaning Text using Chat GPT

Cleaning text using Chat GPT is done by giving the following instructions "Clean the following sentences from Twitter characters and unnecessary characters without changing the original sentence". After executing the command, Chat GPT will automatically clean up these sentences according to the instructions given.

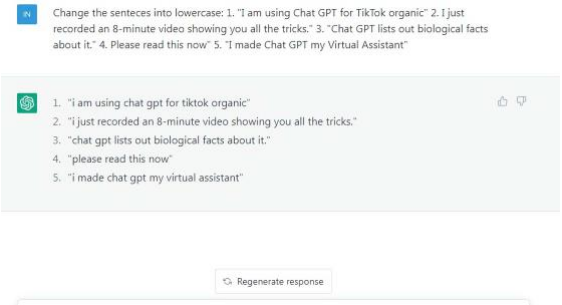


Figure 5. Transform Cases using Chat GPT

The instructions given to make the text in the dataset change to lowercase is to provide the following instruction "Change the sentences into a lowercase". As you can see from the results in Figure 5, all text changes to lowercase.

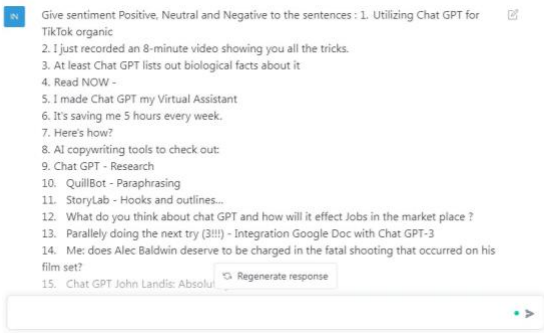


Figure 6. Labeling using Chat GPT

Figure 6 shows the process of giving instructions using Chat GPT to label Positive, Neutral, and Negative text carried out by the Cleaning and Transform Cases processes. The results of this process are presented as shown in Figure 7.

Figure 7. Labeling Results using Chat GPT

Labeling is done on Chat GPT using the “Give label the sentences” instruction. Chat GPT will labeling each sentence given. Each Text-Preprocessing activity using Chat GPT has been carried out. The Text-Preprocessing process will be carried out using Rapidminer, namely Tokenize, Filter Stopwords, Stemming, and Filter Tokens (by length). Chat GPT can be connected to Rapidminer by using the Python Script extension. Its function is to write a Python script that accesses the Chat GPT API and sends a prompt to generate output text. The function is the same as visiting its website. However, after trying it out, there were still errors in responding to the given commands. Therefore, in this study, the method used is to visit the website, then the Text Pre-Processing steps performed by Chat GPT are entered into

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 1, March 2023

Ms.Excel, to be retrieved by Rapidminer for further processing. The process of these stages is presented as shown in Figure 8.

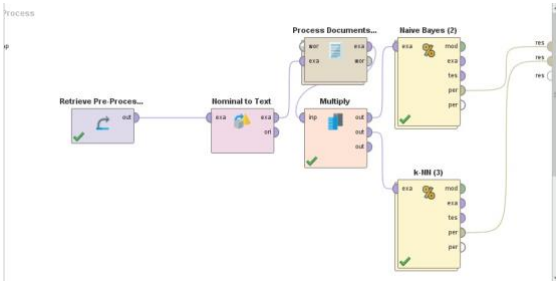


Figure 8. Text-Preprocessing using Rapidminer

The dataset generated from Text-Preprocessing using Chat GPT and stored in Ms.Excel is retrieved using the Retrieve operator in Rapidminer, then connected to several operators including the Process Documents operator which contains other operators Tokenize, Filter Stopwords, Stemming, and Filter Tokens (by length). Each stage of the Text-Preprocessing has been carried out. To see a clear picture of the results of each of these processes, they are presented in tabular form, as shown in Table 3.

Table 3. Text-Preprocessing Results

Preliminary Dataset	'RT @julia_m_mac: I made Chat GPT my Virtual Assistant
Cleaning Text	I made Chat GPT my Virtual Assistant
Tranform Cases	i made chat gpt my virtual assistant
Tokenizing	["i", "made", "chat", "gpt", "my", "virtual", "assistant"]
Filter Stopword	["made", "chat", "gpt", "virtual", "assistant"]
Stemming	["made", "chat", "gpt", "my", "virtual", "assist"]
Text-Preprocessing Results	made chat gpt virtual assist

Based on the results of the Text-Preprocessing, it can be seen that the distribution of sentiments given to GPT Chat exists. The results of the sentiment distribution can be seen as shown in Figure 9.

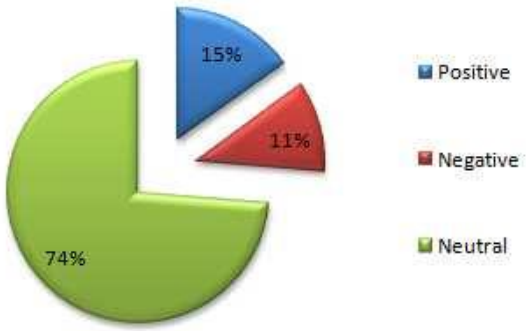


Figure 9. Percentage Rates

The dataset shows that Twitter users give 15% positive sentiment, 74% neutral sentiment, and 11% negative sentiment. Datasets that have gone through the Text-Preprocessing stage can be visualized using Word Cloud. Word Cloud is a visual representation of the frequency of occurrence of words in a text. The size of the letter determines the frequency of occurrence of a word, so the larger the font size, the greater the event of the word. Conversely, the smaller the font size, the lower the frequency of occurrence of the word [44], while the Word Cloud for this dataset is presented as shown in Figure 10.

Figure 10. Word Cloud

Word Cloud shows that the words that often appear in tweets are "Chat" and "GPT," which shows that most tweets regarding GPT Chat have neutral sentiments.

The next stage is model testing using the K-Nearest Neighbor and Naïve Bayes Algorithms. Algorithm performance testing is carried out using K-Fold Cross Validation [45] with a value of K = 10 and using Stratified, Linear, and Shuffle sampling methods. The results of model testing are presented as shown in Table 4.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 1, March 2023

Table 4. Modelling Results

Algorithms	Sampling	Accuracy
K-Nearest Neighbour	Stratified	73.39%
	Linear	**73.57%**
	Shuffle	71.96%
Naïve Bayes	Stratified	52.39%
	Linear	50.54%
	Shuffle	56.07%

Table 4 shows the results of the accuracy of each algorithm. The K-Nearest Neighbor Algorithm has the best accuracy using the Linear Sampling method, which is 73.57%, while the Naïve Bayes Algorithm has the best accuracy using Shuffle Sampling, which is 56.07%.

After obtaining the accuracy results using the Text-Preprocessing method using Chat GPT, the next step is to compare it using the Text Pre-Processing method using the operators available in Rapidminer with the same classification method, namely the K-Nearest Neighbor and Naïve Bayes Algorithms.

The first step is Text Pre-Processing to perform Text Cleaning and Labeling using Rapidminer. The Process Model constructed is presented in the form of an image, as shown in Figure 11.

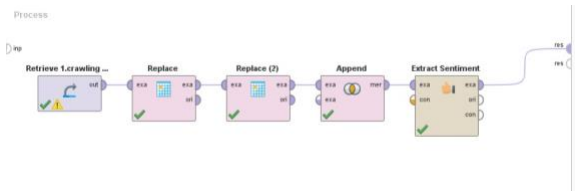
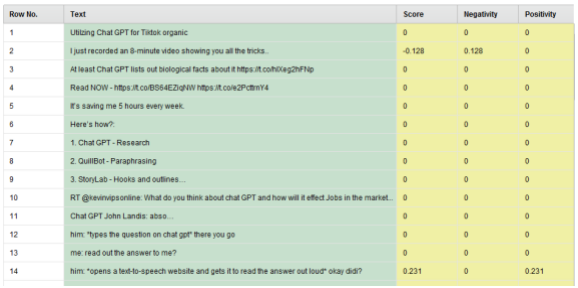


Figure 11. Text-Preprocessing Using Rapidminer

The retrieve operator is used to retrieve the existing dataset, which is then connected to the Replace operator to clean the dataset where its function is to replace parts of the values of selected nominal attributes matching a specified regular expression by a specified replacement. Then, it is connected to the Append operator which function is to build a merged ExampleSet from two or more compatible ExampleSets by adding all examples into a combined set. Finally, it is connected to the Extract Sentiment operator which function is to create a sentiment score by applying either open source sentiment dictionaries or proprietary API methods on an existing text attribute. There are options to expose additional results depending on the chosen method. For its Text Score, it uses the Valence Aware Dictionary and Sentiment Reasoner (VADER) lexicon and rule-based sentiment. VADER is specifically attuned to sentiments expressed in social media and produces scores based on a dictionary of words. This operator calculates and then exposes the sum of all sentiment word scores in the text [46]. The result of the previous Process Model is presented in the form of an image, as shown in Figure 12.



Row No.	Text	Score	Negativity	Positivity
1	Utilzing Chat GPT for Tiktok organic	0	0	0
2	I just recorded an 8-minute video showing you all the tricks..	-0.128	0.128	0
3	At least Chat GPT lists out biological facts about it https://t.co/hlXeg2hFNp	0	0	0
4	Read NOW - https://t.co/BS64EZlqNW https://t.co/e2PctmrY4	0	0	0
5	It's saving me 5 hours every week.	0	0	0
6	Here's how?:	0	0	0
7	1. Chat GPT - Research	0	0	0
8	2. QuillBot - Paraphrasing	0	0	0
9	3. StoryLab - Hooks and outlines...	0	0	0
10	RT @kevinvipsonline: What do you think about chat GPT and how will it effect Jobs in the market...	0	0	0
11	Chat GPT John Landis: abso...	0	0	0
12	him: *types the question on chat gpt* there you go	0	0	0
13	me: read out the answer to me?	0	0	0
14	him: *opens a text-to-speech website and gets it to read the answer out loud* okay didi?	0.231	0	0.231

Figure 12. Cleaning and Labeling Text Result

The text cleaning process is completed, however, in automatic labeling using the Extract Sentiment operator, it only produces two labels, which are Positive and Negative, while in this study, three labels are used, namely Neutral. Therefore, the step taken is to add labels by looking at the values in the Positive and Negative column. If both columns have the same value, then the sentiment given is considered as Neutral.

In the labeling process, using Chat GPT is the most efficient solution compared to using Rapidminer, when using three labels, which are Positive, Negative, and Neutral. On the other hand, if only using two labels, it is not a significant obstacle to use Rapidminer. Adapun hasil pemberian sentiment, disajikan sebagaimana tampak pada Tabel 5.

Table 5. Labeling Using Rapidminer Results

Text	Negativity	Positivity	Sentiment
Utilzing Chat GPT for Tiktok organic	0.000	0.000	Neutral
I just recorded an 8-minute video showing you all the tricks..	0.128	0.000	Negative
At least Chat GPT lists out biological facts about it	0.000	0.000	Neutral
Read NOW -	0.000	0.000	Neutral
It's saving me 5 hours every week.	0.000	0.000	Neutral
Here's how?:	0.000	0.000	Neutral
Chat GPT - Research	0.000	0.000	Neutral
QuillBot - Paraphrasing	0.000	0.000	Neutral
StoryLab - Hooks and	0.000	0.000	Neutral

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 1, March 2023

Text	Negativity	Positivity	Sentiment
outlines… kevinvipsonline: What do you think about chat GPT and how will it effect Jobs in the market place ?	0.000	0.000	Neutral
Chat GPT John Landis: abso…	0.000	0.000	Neutral
him: *types the question on chat gpt* there you go	0.000	0.000	Neutral
me: read out the answer to me?	0.000	0.000	Neutral
him: *opens a text-to-speech website and gets it to read the answer out loud* okay didi?	0.000	0.231	Positive

The next step is to conduct an experiment to compare the results of automatic labeling between using Chat GPT and Rapidminer. The method used is to take the same five texts to see whether there are differences in the sentiment part or not. The comparison is presented as shown in Table 6.

Table 6. Labeling Using Rapidminer Results

No	Text	Sentiment Chat GPT	Sentiment Rapidminer
1	Utilzing Chat GPT for Tiktok organic	Neutral	Neutral
2	**I just recorded an 8-minute video showing you all the tricks.**	**Positive**	**Negative**
3	At least Chat GPT lists out biological facts about it	Neutral	Neutral
4	Read NOW -	Neutral	Neutral
5	**It's saving me 5 hours every week.**	**Positive**	**Neutral**

Table 4 shows that there are differences in No 2 and 5. If analyzed linguistically, the sentiment provided by Chat GPT is more accurate compared to using Rapidminer. This indicates that automatic labeling using Chat GPT can produce better sentiment than labeling using Rapidminer.

The final stage is to build a Process Model to see the accuracy generated using K-Nearest Neighbour and Naïve Bayes algorithms. The Process Model built is presented in the form of a picture, as shown in Figure 13.

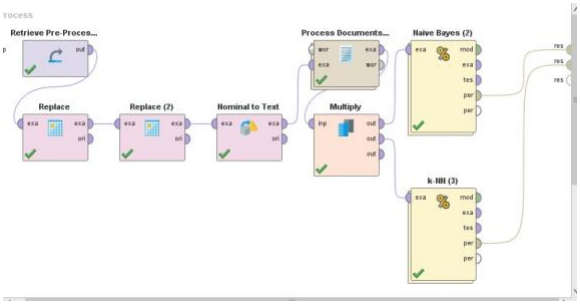


Figure 13. Model Proses Using Rapidminer

Figure 13 shows the Process Document Operator which contains three other operators, namely Tokenize, Filter Stopword, Stemming, and Filter (token by length). Then, from these operators, it is connected to two algorithms, K-Nearest Neighbour and Naïve Bayes, with the same validation method and accuracy calculation as done in the Model Process using Chat GPT. The results are presented in tabular form, as shown in Table 7.

Table 7. Modelling Results

Algorithms	Sampling	Accuracy
K-Nearest Neighbour	**Stratified**	**75.33%**
	Linear	75.00%
	Shuffle	75.33%
Naïve Bayes	Stratified	15.00%
	Linear	19.33%
	Shuffle	16.67%

Table 7 shows that the most optimal result is indicated by the K-Nearest Neighbour algorithm using the Stratified Sampling method. The accuracy result is 75.33%. To see the difference between the accuracy results using Text-Preprocessing Chat GPT and Rapidminer, they are presented in table form, as shown in Table 8.

Table 8. Comparison of Accuracy Results

Algorithms	Sampling	Accuracy Chat GPT	Accuracy Rapidminer
K-Nearest Neighbour	Stratified	73.39%	**75.33%**
	Linear	73.57%	75.00%
	Shuffle	71.96%	75.33%
Naïve Bayes	Stratified	52.39%	15.00%
	Linear	50.54%	19.33%
	Shuffle	**56.07%**	16.67%

The comparison results indicate that the Text Pre-Processing method using Rapidminer produces the best accuracy value through the K-Nearest-Neighbour algorithm with Stratified and Shuffle Sampling method, which is 75.33%. It has a narrow difference with Chat GPT, which is 73.57%.

The opposite result is shown through the Naïve Bayes Algorithm. The Text-Preprocessing method using Chat GPT has an optimal accuracy value through the Shuffle Sampling method, which is 56.07%. This is a significant difference compared to the Naïve Bayes accuracy value of the Text Pre-Processing using Rapidminer, which is 19.33% via the Linear Sampling method. This indicates that using the Naïve Bayes Algorithm, the accuracy value of Chat GPT is better than Rapidminer.

CONCLUSION

The conclusion of this study shows that the use of OpenAI Chat GPT can be an alternative in the Text-Preprocessing sentiment analysis process. Chat GPT can provide insight directly based on our instructions in a natural language like we chat with humans. The model test results also show the best value using the K-Nearest Neighbor Algorithm and the Linear Sampling method with an accuracy of 73.57%. This value is included in the category of Fair Classification.

If compared to the accuracy of Text Pre-Processing using Rapidminer, Chat GPT indeed shows a lower value. However, the values produced by both are not far apart, only a difference of 1.76% when using the K-Nearest Neighbour algorithm and still in the category of Fair Classification.

The opposite result is shown in the Naïve Bayes algorithm, where the accuracy of Chat GPT Text-Preprocessing can exceed Rapidminer's accuracy by a significant margin of 36.74%. Although both are in the category of Failure Classification when using the Naïve Bayes algorithm, optimizing this model will result in better accuracy values.

This study focuses on using Chat GPT in Text-Preprocessing, which has several limitations. First, using limited comparison algorithms using K-Nearest Neighbors and Naïve Bayes, future research can add other algorithms to study. Second, improving accuracy performance is not carried out. In the future, you can use optimization methods such as Feature Selection or Feature Extraction.

ACKNOWLEDGMENTS

Authors wishing to acknowledge Institut Teknologi Garut that support and funds this research publication.

REFERENCES

[1] Patmawati and M. Yusuf, “Analisis Topik Modelling Terhadap Penggunaan Sosial Media Twitter oleh Pejabat Negara,” *Build. Informatics, Technol. Sci.*, vol. 3, no. 3, pp. 122–129, 2021, doi: 10.47065/bits.v3i3.1012.

[2] Junadhi, Agustin, M. Rifqi, and M. K. Anam, “Sentiment Analysis Of Online Lectures Using K-Nearest Neighbors Based On Feature Selection,” *Janapati*, vol. 11, no. 3, pp. 216–225, 2022.

[3] O. P. Zusrotun, A. C. Murti, and R. Fiati, “Sentimen Analisis Belajar Online Di Twitter Menggunakan Naïve Bayes,” *JANAPATI*, vol. 11, no. 3, pp. 310–320, 2022.

[4] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, “Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19),” *J. Media Inform. Budidarma*, vol. 5, no. 2, pp. 406–414, 2021, doi: 10.30865/mib.v5i2.2835.

[5] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, “Twitter Social Media Sentiment Analysis Against Bitcoin Cryptocurrency Trends Using Rapidminer,” *J. Tek. Inform.*, vol. 3, no. 5, pp. 1183–1187, 2022.

[6] I. T. Julianto, “Analisis Sentimen Terhadap Sistem Informasi Akademik Institut Teknologi Garut,” *J. Algoritm.*, vol. 19, no. 1, pp. 449–456, 2022, doi: 10.33364/algoritma/v.19-1.1112.

[7] M. Murali, B. Duraisamy, and J. Vankara, “Measurement : Sensors Independent component support vector regressive deep learning for sentiment classification,” *Meas. Sensors*, vol. 26, no. December 2022, pp. 1–8, 2023, doi: 10.1016/j.measen.2023.100678.

[8] J. Sangeetha and U. Kumaran, “A hybrid optimization algorithm using BiLSTM structure for sentiment analysis,” *Meas. Sensors*, vol. 25, no. December 2022, pp. 1–7, 2023, doi: 10.1016/j.measen.2022.100619.

[9] M. Dowling and B. Lucey, “ChatGPT for (Finance) research: The Bananarama Conjecture,” *Financ. Res. Lett.*, no. 103662, pp. 1–20, 2023, doi: 10.1016/j.frl.2023.103662.

[10] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue," *openai.com*, 2022. https://openai.com/blog/chatgpt/.
[11] S. Demir and B. Topcu, "Graph-based Turkish text normalization and its impact on noisy text processing," *Eng. Sci. Technol. an Int. J.*, vol. 35, pp. 1–13, 2022, doi: 10.1016/j.jestch.2022.101192.
[12] M. O. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari, and A. Hilal, "Preprocessing Arabic text on social media," *Heliyon*, vol. 7, no. 2, pp. 1–15, 2021, doi: 10.1016/j.heliyon.2021.e06191.
[13] A. E. Budiman and A. Widjaja, "Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 3, pp. 475–488, 2020, doi: 10.28932/jutisi.v6i3.2892.
[14] S. Sugriyono and M. U. Siregar, "Preprocessing kNN algorithm classification using K-means and distance matrix with students' academic performance dataset," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 4, pp. 311–316, 2020, doi: 10.14710/jtsiskom.2020.13874.
[15] V. V. Nhlabano and P. E. N. Lutu, "Impact of Text Pre-processing on the Performance of Sentiment Analysis Models for Social Media Data," *2018 Int. Conf. Adv. Big Data, Comput. Data Commun. Syst.*, pp. 1–6, 2018.
[16] L. G. Irham, A. Adiwijaya, and U. N. Wisesty, "Klasifikasi Berita Bahasa Indonesia Menggunakan Mutual Information dan Support Vector Machine," *J. Media Inform. Budidarma*, vol. 3, no. 4, p. 284, 2019, doi: 10.30865/mib.v3i4.1410.
[17] F. Syah, H. Fajrin, A. N. Afif, M. R. Saeputra, D. Mirranty, and D. D. Saputra, "Analisa Sentimen Terhadap Twitter IndihomeCare Menggunakan Perbandingan Algoritma Smote, Support Vector Machine, AdaBoost dan Particle Swarm Optimization," *urnal JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 7, no. 1, pp. 54–58, 2023.
[18] A. H. Anshor and A. Safuwan, "Analisis Sentimen Opini Warganet Twitter Terhadap Tes Screening Genose Pendeteksi Virus Covid-19 Menggunakan Metode Naïve Bayes Berbasis Particle Swarm Optimization," *JINTEKS (Jurnal Inform. Teknol. dan Sains)*, vol. 5, no. 1, pp. 170–178, 2023.
[19] A. P. Nardilasari, A. L. Hananto, S. S. Hilabi, and B. Priyatna, "Analisis Sentimen Calon Presiden 2024 Menggunakan Algoritma SVM," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 7, no. 1, pp. 11–18, 2022.
[20] B. Kurniawan Rachmat, A. Suwarisman, I. Afriyanti, A. Wahyudi, and D. D. Saputra, "Analisis Sentimen Complain dan Bukan Complain pada Twitter Telkomsel dengan SMOTE dan Naïve Bayes," *J. Teknol. Inf. dan Komunikasi)*, vol. 7, no. 1, pp. 107–113, 2023, [Online]. Available: https://doi.org/10.35870/jti.
[21] M. Fahmi, Y. Yuningsih, and A. Puspita, "Sentiment Analysis Of Online Gojek Transportation Services On Twitter Using The Naïve Bayes Method," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 8, no. 2, pp. 84–90, 2023, doi: 10.33480/jitk.v8i2.4004.
[22] M. R. Qisthiano, I. Ruswita, and P. Armilia, "Implementasi Metode SVM dalam Analisis Sentimen Mengenai Vaksin dengan Menggunakan Python 3," *J. Ilm. Sist. Inf.*, vol. 13, no. 1, pp. 1–7, 2023.
[23] D. Setiyawati and N. Cahyono, "Analisa Sentimen Pengguna Sosial Media Twitter Terhadap Perokok di Indonesia," *Indones. J. Comput. Sci.*, vol. 12, no. 1, pp. 262–272, 2023.
[24] Alfandi Safira and F. N. Hasan, "Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode Naive Bayes Classifier," *Zo. J. Sist. Inf.*, vol. 5, no. 1, pp. 59–70, 2023, doi: 10.31849/zn.v5i1.12856.
[25] M. T. Anwar, D. Riandhita, A. Permana, P. Sistem, I. Industri, and J. Pusat, "Analisis Sentimen Masyarakat Indonesia Terhadap Produk Kendaraan Listrik Menggunakan VADER," *J. Tek. Inform. dan Sist. Inf.*, vol. 10, no. 1, pp. 783–792, 2023.
[26] I. P. Rahayu, A. Fauzi, and J. Indra, "Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Naive Bayes Dan Support Vector Machine," *J. Sist. Komput. dan Inform. Hal 296−*, vol. 301, no. 2, pp. 25–38, 2022.
[27] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, "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.
[28] H. Andriana, S. S. Hilabi, and A. Hananto, "Penerapan Metode K-Nearest Neighbor pada Sentimen Analisis Pengguna Twitter Terhadap KTT G20 di Indonesia," *JURIKOM (Jurnal Ris. Komputer)*, vol. 10, no. 1, pp. 60–67, 2023, doi: 10.30865/jurikom.v10i1.5427.
[29] A. Pebdika, R. Herdiana, and D. Solihudin,

"Klasifikasi Menggunakan Metode Naive Bayes Untuk Menentukan Calon Penerima PIP," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 452–458, 2023.

[30] M. K. Insan, U. Hayati, and O. Nurdiawan, "Analisis Sentimen Aplikasi Brimo Pada Ulasan Pengguna Di Google Play Menggunakan Algoritma Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 478–483, 2023.

[31] M. Dennis, F. Zoromi, and M. K. Anam, "Penerapan Algoritma Naïve Bayes Untuk Pengelompokkan Predikat Peserta Uji Kemahiran Berbahasa Indonesia," *J. Media Inform. Budidarma*, vol. 6, no. 2, pp. 1183–1190, 2022, doi: 10.30865/mib.v6i2.3956.

[32] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, "Comparison Of Classification Algorithm And Feature Selection in Bitcoin Sentiment Analysis," *JUTIF*, vol. 3, no. 3, pp. 739–744, 2022.

[33] D. S. Utami and A. Erfina, "Analisis Sentimen Pinjaman Online di Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *SISMATIK (Seminar Nas. Sist. Inf. dan Manaj. Inform.*, vol. 1, no. 1, pp. 299–305, 2021.

[34] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval (2nd edition)*. Cambridge: Cambridge University Press, 2009.

[35] Han and Kamber, *Data Mining Concepts and Technique*. San Francisco: Diane Cerra, 2006.

[36] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Practical Machine Learning Tools and Technique*. San Francisco: Morgan Kaufmann, 2011.

[37] L. K. Harsono, Y. Alkhalifi, Nurajijah, and W. Gata, "Analisis Sentimen Stakeholder atas Layanan haiDJPb pada Media Sosial Twitter Dengan Menggunakan Metode Support Vector Machine dan Naïve Bayes," *J. Ilmu-ilmu Inform. dan Manaj.*, vol. 14, no. 1, pp. 36–44, 2020.

[38] A. Ahmad and W. Gata, "Sentimen Analisis Masyarakat Indonesia di Twitter Terkait Metaverse dengan Algoritma Support Vector Machine," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 6, no. 4, pp. 548–555, 2022, doi: 10.35870/jtik.v6i4.569.

[39] G. Feng, M. Fan, and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," *IEEE Access*, vol. 10, pp. 19558–19571, 2022, doi: 10.1109/ACCESS.2022.3151652.

[40] A. Y. Pratama, Y. Umaidah, and A. Voutama, "Analisis Sentimen Media Sosial Twitter Dengan Algoritma K-Nearest Neighbor dan Seleksi Fitur Chi-Square (Kasus Omnibus Law Cipta Kerja)," *Sains Komput. Inform.*, vol. 5, no. 2, pp. 897–910, 2021, [Online]. Available: https://tunasbangsa.ac.id/ejurnal/index.php/jsakti/article/view/386/365.

[41] K. Ayuningsih, Y. A. Sari, and P. P. Adikara, "Klasifikasi Citra Makanan Menggunakan HSV Color Moment dan Local Binary Pattern dengan Naïve Bayes Classifier," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 3, no. 4, pp. 3166–3173, 2019.

[42] Yunitasari, H. S. Hopipah, and R. Mayasari, "Optimasi Backward Elimination untuk Klasifikasi Kepuasan Pelanggan Menggunakan Algoritme k-nearest neighbor (k-NN) and Naive Bayes," *Technomedia J.*, vol. 6, no. 1, pp. 99–110, 2021, doi: 10.33050/tmj.v6i1.1531.

[43] D. Nurlaela, "Penerapan Adaboost untuk Meningkatkan Akurasi Naive Bayes Pada Prediksi Pendapatan Penjualan Film," *Inti Nusa Mandiri*, vol. 14, no. 2, pp. 181–188, 2020.

[44] R. Parlika, S. I. Pradika, A. M. Hakim, and K. R. N. M, "Analisis Sentimen Twitter Terhadap Bitcoin dan Cryptocurrency Berbasis Python TextBlob," *J. Ilm. Teknol. Inf. dan Robot.*, vol. 2, no. 2, pp. 33–37, 2020.

[45] D. Kurniadi, E. Abdurachman, H. L. H. S. Warnars, and W. Suparta, "Predicting student performance with multi-level representation in an intelligent academic recommender system using backpropagation neural network," *ICIC Express Lett. Part B Appl.*, vol. 12, no. 10, pp. 883–890, 2021, doi: 10.24507/icicelb.12.10.883.

[46] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, "Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile," *PETIR J. Pengkaj. dan Penerapan Tek. Inform.*, vol. 15, no. 2, pp. 264–275, 2022.