International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 6, No. 1, (2022), pp. 134-144

Text Mining for Customer Sentiment Using Naive Bayes and SMOTE Methods on TokopediaCare Twitter

Rico Budiyanto[1], Indah Purnamasari[2], Dedi Dwi Saputra[3]
[1,2,3]Faculty of Information Technology, Universitas Nusa Mandiri

[1]11211942@nusamandiri.ac.id, [2]indah.ihi@nusamandiri.ac.id, [3]dedi.eis@nusamandiri.ac.id

Abstract

At this time, buying and selling online has become part of the lives of the Indonesian people and the world, especially during the pandemic, marketplace users are increasing and slowly replacing traditional markets. Tokopedia as one of the largest marketplaces in Indonesia has the largest users in the 3rd quarter of 2019. Customer complaints to Tokopedia services can be submitted through Social Media such as Twitter and also other media. Complaints submitted via Twitter to the TokopediaCare are still manually identified by Tokopedia customer service so it takes a long time to respond to customer complaints, because customer services need so many time to classified where is complaint or not complaint tweet. Text mining is used to process customer complaint data through text or sentences submitted by tweets using the Naïve Bayes method and the Syntethic Minority Oversampling Technique Method (SMOTE) feature for the implementation of machine learning can help identify the classification of complaints submitted via Twitter automatically. The use of the Naive Bayes method is added with the Synthetic Minority Oversampling Method feature which is considered better for generating predictions on tweets submitted by customers.

Keywords: *Tokopedia, Text Mining, Naive Bayes, SMOTE, Sentiment Analysis*

1. Introduction

To provide customers satisfaction, Tokopedia provides access to customer service through social media, one of which is Twitter. Access to customer service through Twitter makes it easy for customers to submit complaints without having to call customer service, besides that customers who are active on Twitter social media do not need to open an application to make a complaint. They done this to achieve customer satisfaction, in order to create customer loyalty to Tokopedia services. Customer satisfaction is something that is expected by the company when the goods or services have been marketed [1]. Sentiment Analysis or opinion mining according to Liu is based on the broad field of Natural Language processing, text mining and linguistic computation which aims to analyze opinions, sentiments, attitudes, emotions, evaluate someone's judgment, namely the author or speaker regarding a discussion, product, individual, organizational services or certain activities that describe a big problem. There are several names used for Sentiment Analysis such as opinion mining, subjectivity analysis, opinion extraction, emotional analysis, sentiment mining, affect analysis, review mining, emotional analysis, and so on [2].

Data mining is a process to get interesting patterns and knowledge from some data. Data origins can include databases, data warehouses, the Web, repositories and other information, or data that has been dynamically streamed into the system [3]. Currently, complaints submitted via Twitter to Tokopedia services via Twitter TokopediaCare are still being analyzed manually, because there is no tool that can predict a sentence in a tweet, so we need a tool that can help in dividing the classification between complaints and non-complaints. this is due to the limited technology and literature that discusses text

ISSN : 2580-7250

mining in detecting patterns of humanities characteristics, especially sentiment analysis in Bahasa. Naive Bayes as one of the methods in machine learning that applies probability [4] and compares by adding the SMOTE (Synthetic Minority Oversampling Technique Methode) feature to resolve imbalanced data [5] to find out how effective these methods and techniques are in predicting tweets from users, In this case we takes a case study on the TokopediaCare Twitter account.

2. Research Methodology

Following are the stages of research on Text Mining for Sentiment Analysis of Customers Using the Naive Bayes and SMOTE:

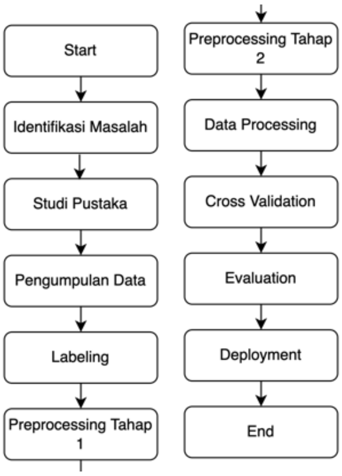

Figure 1. Stage Of Research

2.1. Problem Identification

At this stage, problem identification is carried out based on direct observations or observations on the TokopediaCare Twitter account to find out the problems that exist in the Tokopedia services.

2.2. Literature Review

After identifying the problem, a literature review is also carried out on pre-existing research related to machine learning, data mining, marketing and customer service and others related to the research to be carried out.

3. Results and Discussion

3.1. Collecting Data

That is collecting data through the RapidMiner application using a connection to a Twitter account and operator Retrieve Twitter by connecting the Twitter API through the developer's Twitter account. Twitter data can be accessed through the Twitter REST (Representational State Transfer) API which has been provided by Twitter by first submitting a request to Twitter to obtain data access from Twitter by registering as a developer account.

a) Population and Data Sample
Data population in this study is tweet data from customers Tokopedia marketplace users that mention to the TokopediaCare twitter account from November 2021 to April 2022 with a total sample of 2584 sentences (tweet)

b) Data Type
Data Type used in this study is secondary data (public data) originating from the interaction of Tokopedia users with the TokopediaCare account on Twitter social media.

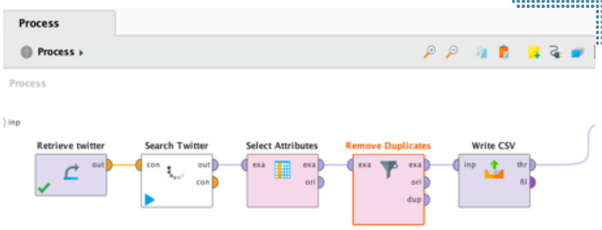

Figure 2. Collecting Data Process

3.2. Labeling

At this stage, the labeling of each sentence in the tweet that has been collected has been carried out so that there are results of classification of complaints and non-complaint categories. Labeling of 2584 datasets by Cityzen or Tokopedia Users and the data result with the classification of "complaint" is 601 data, while for the classification of "not complaint" as many as 1983 data.

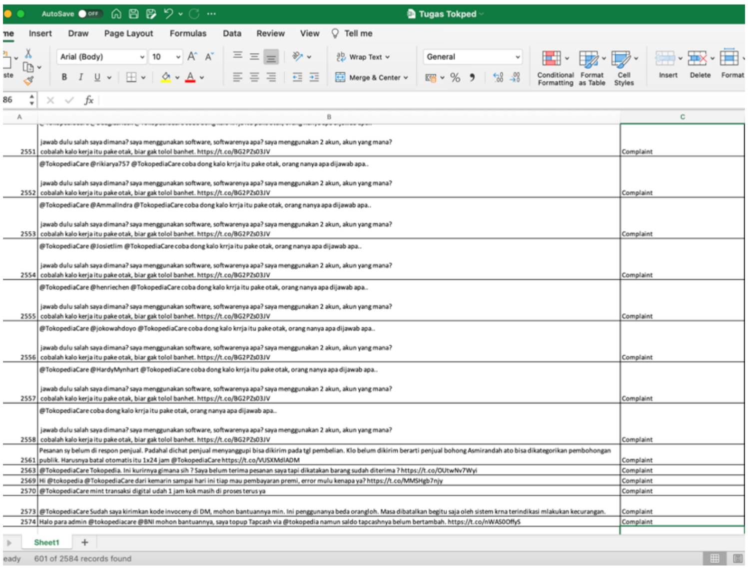

Figure 3. Labeling Complaint Sentences

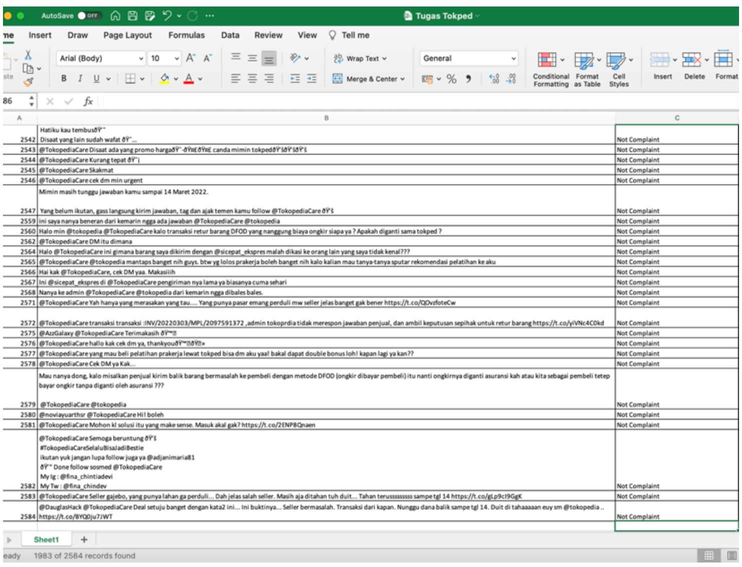

Figure 4. Labeling Not Complaint Sentences

3.3. Preprocessing Phase 1

At this stage, data processing will be done through the Gata Framework web application, there are 5 processes [6], that is:

Figure 5. Preprocessing Phase 1 Process

a) Annotation Removal
On the Gata Framework website, it is done to remove the mention or @ sign in the tweet sentence so that the resulting sentence does not have the @ sign.

b) Transformation Remove URL
In this process, the URL is removed in the sentence contained in the tweet on the Gata Framework web.

c) Tokenizasion Regular Expression (Regexp)
At this stage, the sentences on a dataset are broken into words on the Gata Framework website.

d) Indonesian Stemming
In this process, the sentence that contains the affix is removed so that the word that contains the affix becomes the base word on the Gata Framework website.

e) Indonesian Stop Word Removal
The last process in preprocessing phase 1 is to do Indonesian Stopword removal on the Gata Framework website.

Tabel 1. Preprocessing Result using Gata Framework Website

No	Text	Status	@Anotation Removal	Transformation: Remove URL	Regexp	Indonesian Stemming
1.	@TokopediaCare Siap ka, tolong dibantu ya terima kasih	Not Complaint	siap ka, tolong dibantu ya terima kasih	siap ka, tolong dibantu ya terima kasih	siap ka tolong dibantu ya terima kasih	siap ka tolong bantu ya terima kasih
2.	@tokopedia Pearl pink, warnanya manis bangett kaya minto @TokopediaCare	Not Complaint	pearl pink, warnanya manis bangett kaya minto care	pearl pink, warnanya manis bangett kaya minto care	pearl pink warnanya manis bangett kaya minto care	pearl pink warna manis bangett kaya minto care
….	….	….	…..	…….	……	……
2584	@DauglasHack @TokopediaCare Deal setuju banget dengan kata2 ini... Ini buktinya... Seller bermasalah. Transaksi dari kapan. Nunggu dana balik sampe tgl 14. Duit di tahaaaaan euy sm @tokopedia .. https://t.co/BYQ0ju7JWT	Not Complaint	deal setuju banget dengan kata2 ini... ini buktinya... seller bermasalah. transaksi dari kapan. nunggu dana balik sampe tgl 14. duit di tahaaaaan euy sm .. https://t.co/byq0ju7jwt	deal setuju banget dengan kata2 ini... ini buktinya... seller bermasalah. transaksi dari kapan. nunggu dana balik sampe tgl 14. duit di tahaaaaan euy sm ..	deal setuju banget dengan kata ini ini buktinya seller bermasalah transaksi dari kapan nunggu dana balik sampe tgl duit di tahaaaaan euy sm	deal tuju banget dengan kata ini ini bukti seller masalah transaksi dari kapan nunggu dana balik sampe tgl duit di tahaaaaan euy sm

International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 6, No. 1, (2022), pp. 134-144

3.4. Preprocessing Phase 2

At this stage we use Rapid Miner Software, several functions, methods and operators are carried out in data processing, namely read excel, which is reading data that has been carried out in preprocessing phase 1, then process Nominal to Text , namely converting nominal attributes into string attributes, Process Document from data in which there are several operators to process the existing data set, SMOTE Upsampling to overcome the imbalance data and without SMOTE upsampling as a comparison of the level of accuracy, Precission, Recall, and AUC of the data set owned, then the process is carried out using the Cross Validation operator.

a) Read Excel
This stage is carried out by importing data that has been preprocessed in phase 1 into the Rapid Miner application using the read excel operator.

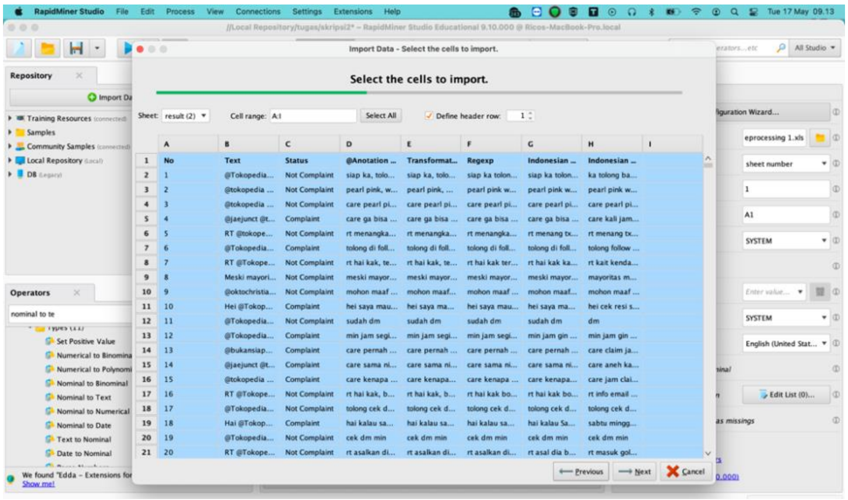

Figure 6. Import data from excel process using read excel Operator

b) Nominal To Text
This operator functions to convert the selected nominal attribute type to text and maps all attribute values to the appropriate string

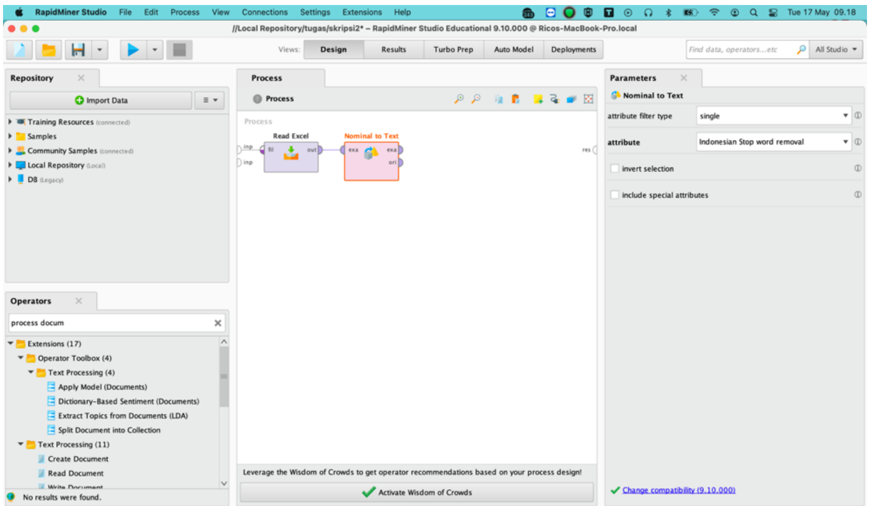

Figure 7. Nominal To Text operator

c) Process Document From Data
Several processes are used to clean the data so that it becomes a vector that can be used as an algorithm calculation, including:

1. Transform Case on the parameter tab we change to transform to lower case is to change the character to lowercase.
2. Tokenize to break sentence into word
3. Filter Tokens (by Length) to select words to be processed into words with a minimum of 4 character and 25 characters maximum
4. Filter Stopwords (Dictionary) to remove words that have low information from a text.
5. Steming (dictionary) to remove affixes in the form of confixes, prefixes, or suffixes in a word so that it returns to the basic word.

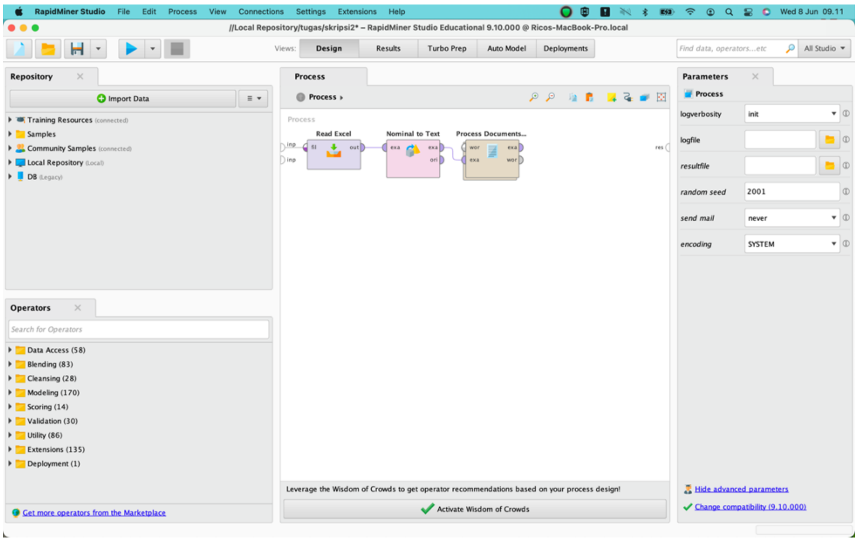

Figure 8. Process Document From Data operator

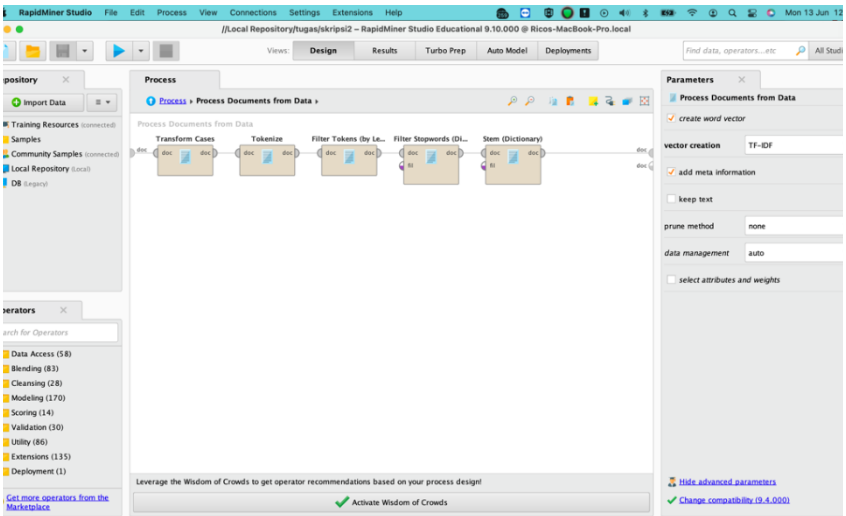

Figure 9. Inside Process Document From Data operator

3.5. Evaluation

At this stage an evaluation of the performance vector of machine learning is carried out with reference to the Confusion Matrix in the form of accuracy, precision, recall, AUC (Optimistic), AUC, AUC (pessimistic) and stemming words that have no value or reduce the level of performance in performance. vectors. The following is an explanation of:

a) Accuracy

The ratio of predictions to the true class (“complaint” and not “complaint”) to the entire data. Accuracy can answer questions about the percentage of complaints and not complaints submitted via twitter to the TokopediaCare account.

Accuracy = (TP+TN) / (TP + FP + FN + TN)

b) Precission

The ratio of true positive prediction to the overall positive predicted outcome. Precision will answer the question of what percentage of "complaint" are submitted from all customers who are predicted to "complaint".

Precisson = (TP) / (TP + FP)

c) Recall (Sensitivity)

the ratio of true positive predictions compared to the overall true positive data. Recall can answer the question "what percentage of customers have been predicted as "complaint" compared to all customers who actually "complaint" [7].

Recall = (TP) / (TP + FN)

d) AUC (Area Under The Curve)

AUC (Area Under The Curve) serves to make it easier to compare one model with another, AUC is the area under the ROC (Receiver Operating Characteristics) curve or as an integral of the ROC [8].

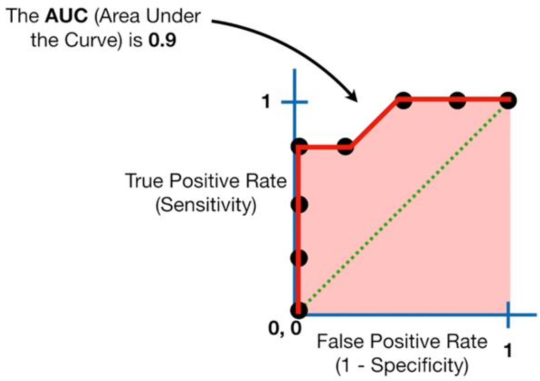

Figure 10. AUC Sample

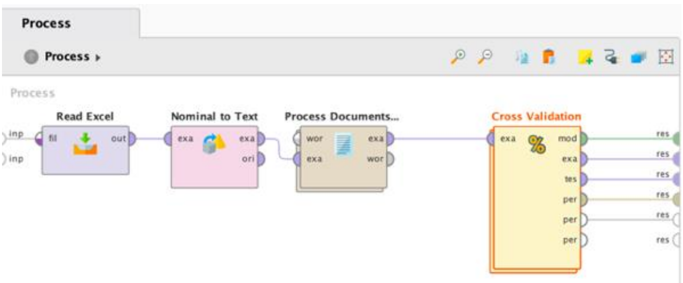

Figure 11. Cross Validation Operator

We use tthis operator is to perform training and also testing on data sets in which there is a Naive Bayes algorithm operator for training data for evaluation process. This use for testing the data to determine the level of accuracy, performance, recall and AUC (Area Under Curve) in the dataset. Inside cross validation operator there is training and testing section we use operator Naive Bayes to training the data and apply model to test Naive Bayes Algorithm and we got the performance

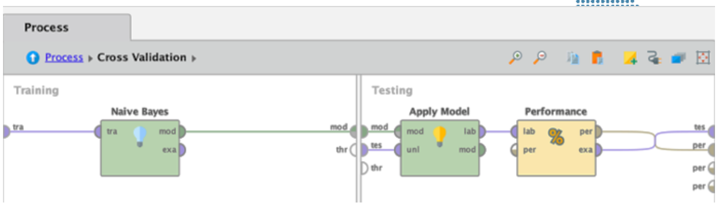

Figure 12. Inside Cross Validation Operator

The use of the Naive Bayes Algorithm in this study resulted in the following data:

1. Accuracy : 83.13%
2. Precission : 65%
3. Recall : 60.56%
4. AUC : 0.801%

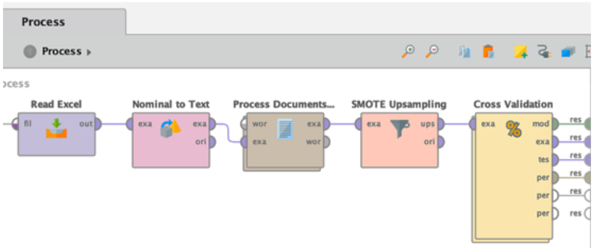

Figure 13. Added SMOTE Upsampling Operator

And we use SMOTE for comparation of Naive Bayes Algorithm and result is:

1. Accuracy : 81.59%
2. Precission : 82.73%
3. Recall : 84.17%
4. AUC : 0.852%

3.6. Deployment

At the deployment stage, database input is carried out from the dataset that has been preprocessed in stages 1 and stage 2, deployment is also the application of machine learning into the php programingfrom the Gata Framework so that it can directly predict tweets addressed to the TokopediaCare Twitter account. Depolyment based on the results of the evaluation of the process of testing the model between the Naive Bayes algorithm and the Naive Bayes algorithm model added with the Syntethic Minority Over Sampling Technique Method (SMOTE) feature. Because this weight will be used in predicting a sentence or tweet in the application

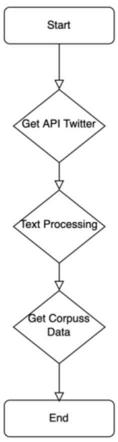

Figure 14. insert preprocessing result into database

International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 6, No. 1, (2022), pp. 134-144

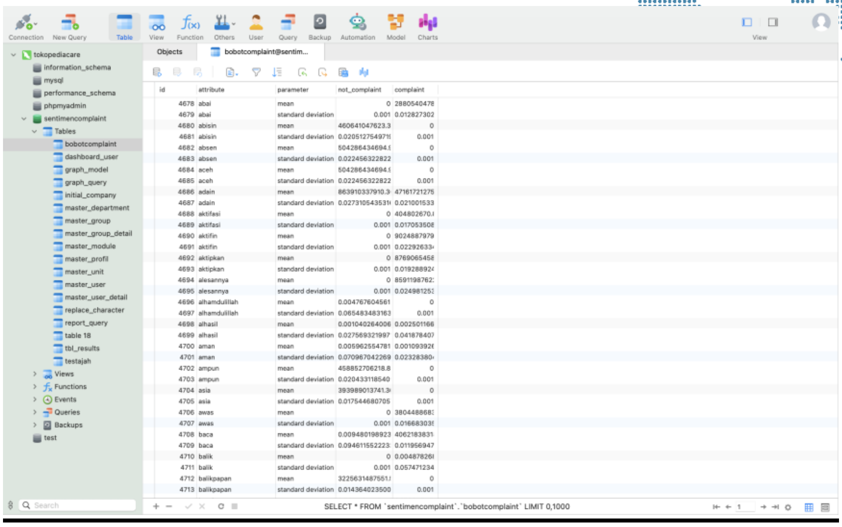

Figure 15. Insert Text Mining Result into Database

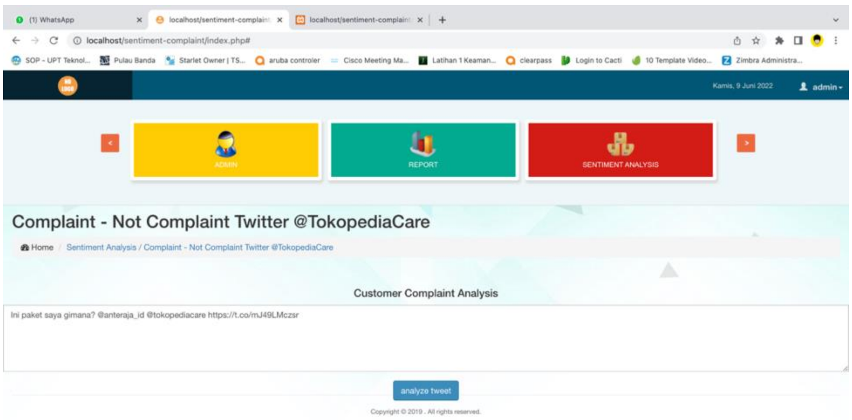

Figure 16. Get API Twitter

Deployment results to retrieve data from the @Tokopedia Care Twitter account using an API (Application Programming Interface). Then the next step, when you press the Analyze Tweet button, it will go to the text processing stage.

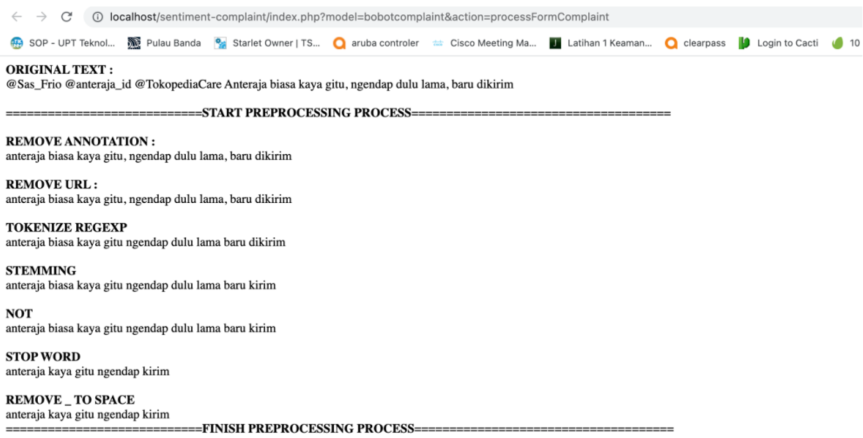

Figure 17. Preprocessing Result on application

After the tweet data is retrieved, proceed with preprocessing and cleansing the tweet, using the Remove @annotation, Remove URL, Tokenize Regexp, Stemming, Not Transformation Negative, Stopword, Remove _ to Space techniques.

```
==========================FINISH PREPROCESSING PROCESS==================================

================================PROBABILITY OF WORD=============================================

0 anteraja

1 kaya
kaya Not Complaint : 0.00823279995568006, Complaint : 0.011780314690692314
kaya Not Complaint : 0, Complaint : 0.00673997608950266

2 gitu

3 ngendap

4 kirim
kirim Not Complaint : 0.021534167761419094, Complaint : 0.09014604229356253
kirim Not Complaint : 0.0022669019149307, Complaint : 0
================================PROBABILITY OF WORD=============================================

================================================================================
SUMMARY WEIGHT OF WORD POSITIVE or NEGATIVE

Not Complaint : 0.03203386963203
Complaint : 0.10866633307376
================================================================================
KESIMPULAN: Complaint
```

Figure 18. Word Weight Calculation

In Figure 15 stages of word weight assessment in tweets addressed to the @TokopediaCare account, the results of the calculation produce a Complaint conclusion because the complaint weight of the words in the tweet is greater than Not Complaint.

4. Conclusion

The results of the research that has been carried out regarding customer sentiment towards Tokopedia services carried out via Twitter on the @TokopediaCare account, there are the following conclusions, the use of the Naive Bayes Algorithm in the training and testing process of the processed dataset obtains an accuracy value of 83.13%, which is higher than the addition of the Syntethic Minority Over Sampling Technique Method (SMOTE) feature. In the implementation of Machine Learning to predict tweet sentences used at the time of deployment using the Naive Bayes algorithm and the addition of the Syntethic Over-Sampling Technice Method (SMOTE) feature, it is more effective in overcoming data imbalances because the resulting weight can reach above 80%, namely accuracy 81.59 %, Precision 82.73%, Recall 84.17%, and AUC of 0.852% while the use of the Naive Bayes algorithm that does not add SMOTE features is 83.13% accuracy, Precision 65.00%, Recall 60.56%, and AUC is 0.801%.

References

[1] N. Ruhyana and D. Rosiyadi, “Klasifikasi Komentar Instagram untuk Identifikasi Keluhan Pelanggan Jasa Pengiriman Barang dengan Metode SVM dan Naïve Bayes Berbasis Teknik Smote,” *Fakt. Exacta*, vol. 12, no. 4, p. 280, 2020, doi: 10.30998/faktorexacta.v12i4.4981.
[2] B. Liu, “Sentiment Analysis and Opinion Mining,” no. May, 2012.
[3] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.
[4] Mochammad Haldi Widianto, “Algoritma Naive Bayes,” *Https://Binus.Ac.Id*, 2019. https://binus.ac.id/bandung/2019/12/algoritma-naive-bayes/ (accessed Apr. 09, 2022).
[5] Arwan, V. Ardiana, L. Reza Ariana, F. Samuel, D. Ramdani, and Aditya, “Synthetic Minority Over-sampling Technique (SMOTE) Algorithm For Handling Imbalanced Data,” *binus.ac.id*, 2018.
[6] D. D. Saputra *et al.*, “Optimization Sentiments of Analysis from Tweets in

myXLCare using Naïve Bayes Algorithm and Synthetic Minority over Sampling Technique Method," *J. Phys. Conf. Ser.*, vol. 1471, no. 1, 2020, doi: 10.1088/1742-6596/1471/1/012014.

[7] R. Arthana, "Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning," *Medium.com*, 2019. https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8.

[8] R. Arifin, "Memahami ROC dan AUC," *Medium.com*, 2019.