International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 3
ISSN: 2088-5334

Customer Satisfaction Assessment System on Transactions E-commerce Product Purchases Using Sentiment Analysis

Amil Ahmad Ilham [a,*], Anugrayani Bustamin [a], Eugenius Wahyudiarto [a]

[a] *Informatics Department, Universitas Hasanuddin, Makassar, 90245, Indonesia*
Corresponding author: [*]amil@unhas.ac.id

***Abstract*— Currently, more products appear, and various services that offer similar products make it difficult for buyers to decide to buy before seeing reviews from other users. The growth of different e-commerce platforms exacerbates this. Users spent more time choosing products on each platform with many alternative considerations, such as looking at ratings, prices, and reviews from other buyers. This study conducted the optimization process of selecting e-commerce products so that users do not have to spend a long time reading every review when they want to buy a product. This research is expected to provide a comprehensive assessment of the purchase transaction of a product from the reviews provided. The data is sourced from product reviews on e-commerce in Indonesia, which are then classified into positive, negative, and neutral sentiments. The data is divided into 10 folds of data using stratified k-fold cross-validation, consisting of training and testing data with ratios of 90% and 10% of the total data. Our research proposed a system that implemented our modified Naive Bayes model to calculate a product's Customer Satisfaction (CSAT) score and compare it with the Google Cloud NLP model. In our model, the log prior and log-likelihood formulas are modified in the algorithm, adding the prefix "NOT_" after the negation words in the preprocessing. This doubled our model's F1 score and increased the accuracy by 32%, from 59% to 91%, when compared to the Naive Bayes algorithm without modification.**

***Keywords*— Customer satisfaction score; product purchase recommendation; sentiment analysis; Naive Bayes.**

Manuscript received 30 Aug. 2022; revised 6 Oct. 2022; accepted 7 Nov. 2022. Date of publication 30 Jun. 2023.

I. INTRODUCTION

Research involving scenarios in the discipline of Natural language processing is becoming popular nowadays, along with the growth of data. However, not a bit of data on the internet is a new source of information that can be utilized in different contexts, such as using a sentiment analysis-based recommendation system. Sentiment analysis is often referred to as opinion mining. This analysis explores the context to identify information from the source material.

This is a matter of recommendation. There are two categories of recommendation techniques, content-based filtering (CBF) and collaborative filtering (CF) [1]. Some combine the two so that it becomes a hybrid. CBF will recommend products based on their similarity to those previously purchased, and CF will recommend products based on those purchased by other entities. Several solutions can be done to overcome the problems above. First, the product search feature will enter text into the tag classification model and perform a query to the database. Second, recommend the best product. To get the best product, it is necessary to do a comparison among the products and also reviews of each product. Sentiment analysis (SA) will be used to determine user attitudes towards the products offered [2]. Several algorithms can be used in SA such as XGBoost, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Logistic Regression (LR) [1], and Naive Bayes [3]. Some even use lexicon-based keywords that are combined with the algorithm [2]. The challenge is preprocessing the raw data that may contain typos from users, Indonesian abbreviations, and emoticons. Emoticons play an important role in SA [3].

Several natural language processing studies with sentiment analysis scenarios have been developed. Alatrash et al. [4] proposed a novel e-learning hybrid recommendation system methodology based on sentiment analysis using a Convolutional Neural Network (CNN). This study recommended appropriate e-learning materials based on learners' preferences and tested using customizable datasets, namely ABHR-1 and two public datasets, with an accuracy of 90,37%. In another research related to the sentiment analysis scenario, a novel unsupervised learning and hierarchical

clustering method is proposed using the Twitter dataset. Accuracy measure (proportion of correct predictions) is used to evaluate the performance of understudied techniques. It is empirically shown that the performance of unsupervised learning techniques is comparable with supervised learning techniques [5].

Different studies have been developed using data from social media such as Twitter, Facebook, Instagram, and others to gain new knowledge, including sentiments from user comments [6]–[8]. Farzadnia and Vanani also conducted sentiment analysis for service reviews in 2022. This study reviews comments from airline passengers and classifies them based on their sentiment level. This process becomes a reference for new approaches to evaluating and improving customer satisfaction [9]. Furthermore, using social network data, sentiment analysis was used in the biomedical aspect to evaluate the conveyed patient's review in clinical outcomes or the impact of a drug and a medical process. [10]. Several challenges to processing social media data come from accustomed to utilizing a set of graphic symbols to express their emotion, namely emoticons. Therefore, Liu et al. examined suitable sentiment analysis methods, including rule-based and classification algorithms, to involve the impact of supplementing emoticons as additional features to enhance the performance of algorithms [11]. The CSAT system has been built using Twitter data containing emoticons and classified using SVM. Social media presents its own set of opportunities to get a positive impact on improving the quality of companies based on user comments. This study showed that the algorithm achieved an accuracy of around 87% [12]. A model for classifying customer satisfaction was also developed in 2019 by acquiring data from Booking.com. This study classifies hotel user reviews to improve service quality and hotel management. Word2vec and the artificial neural network algorithm yield 92.48% accuracy for classifying sentiment in this study [13].

A comparative review shows various ML models that are used in sentiment analysis-based recommendation systems. The result is SVM stands out as more effective in most cases, but results are application dependent, while NB tends to be stable due to the probability of word occurrence in data [14], [15]. The biggest obstacle faced by sentiment analysis-based recommendation systems is data sparsity. Content-based filtering or collaborative filtering cannot solve the problem. A hybrid approach can be a solution, using only the advantages of both techniques effectively [14]. A context-aware recommender system is also proposed to recommend nearby restaurants that match the user's preferences. NLP techniques extracted user preferences for food, then clustered the food names using a semantic approach. Sentiment analysis is used to obtain users' opinions regarding each food, whether it is positive or negative [16].

ML models have been studied to classify sentiment analysis of user reviews about My Indihome; visualization and text association are used to extract and identify the topic and information that users often discuss in the comment section [17]. Another research also studied sentiment analysis of Indonesian tweet using rule-based and ML algorithm approach with bag of words feature extraction. The result shows NB method obtained higher accuracy than the lexicon-based method, and the accuracy is strongly influenced by the word references in the bag of words [18].

Our system is similar to what Kim [19] did. The difference is Kim's system only uses positive and negative keywords to determine sentiments, and there are always neutral sentiments from customers. Our approach considers positive, negative, and neutral sentiments for calculating product scores. The system created by Hanni can compare product specifications [20]. However, since not all products have specifications, we proposed a customer satisfaction assessment score, as a recommendation, based on the results of sentiment analysis.

II. Materials and Method

A. General Overview

The following is a general overview of the system developed in this study, shown in Fig.1.

Fig. 1 CSAT assessment system design

The stages of how the system in Fig.1 works described as follows.

1) Web Scrapers: At this stage, the review data is taken from the Tokopedia website [21]. The metadata taken from

this stage is shown in Fig. 2. There are three tables to accommodate raw data from Tokopedia, namely pages, products, and reviews. Note that the JSON data type is used in the value column to dynamically store data [22].

Fig. 2 Scraper database structure

2) Data: At this stage, the data that has been stored is carried out by a query process to be processed at the next stage.

3) Preprocessing: On the data that has been selected, a preprocessing operation is applied so that it can be used to create a sentiment analysis model [23].

4) Analyze sentiment from each review: In this stage, a sentiment analysis model is formed, and then a model is also formed to detect the topic of each review. There are two alternative sentiment models used, namely:

- Pretrained Model (Google Cloud Natural Language / GCloud NL): using GCloud NL to analyze sentiment [24].
- Model from scratch (Naive Bayes): using a self-made Naive Bayes model to analyze sentiment [25].

5) Result: The result from the previous stage generates the sentiment of each review.

6) Score Calculation: The sentiment results of each review are processed to produce a product recommendation score.

7) Database: The existing scores are then stored as a new table in the database to be used in the viewer application.

8) Web Viewer: The CSAT score results are taken from the database to be displayed on the web page.

B. Data Collection

To retrieve review data from Tokopedia, a web scraper is required. This web scraper was created using the Python programming language. The data structure formed from this process is shown in Fig.2. The data retrieval step was carried out with the steps described in the form of a flowchart in Fig. 3 as the main scrapper process, Fig. 4 to get the URL index of each product category, Fig. 5 to get the URL of product detail, and Fig. 6 to get the product detail data.

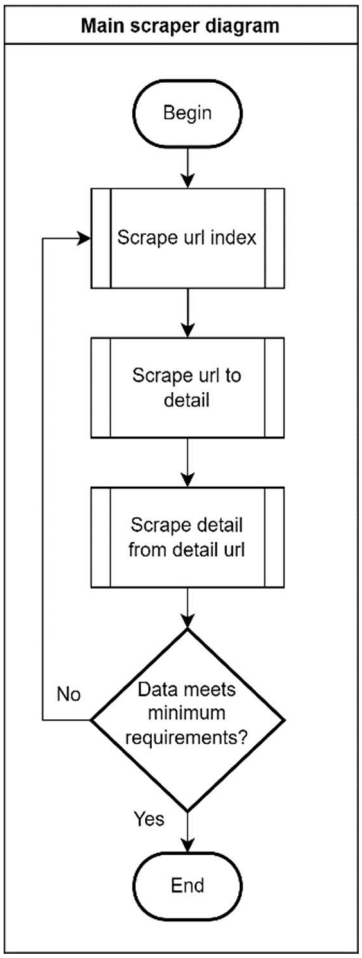

Fig. 3 Main scraper process

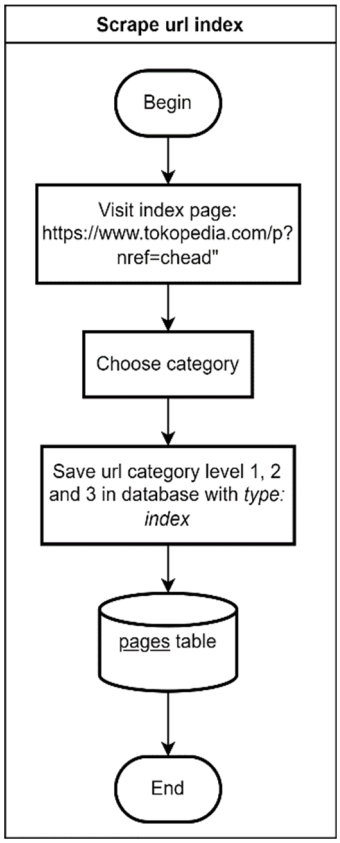

Fig. 4 Index scraping process

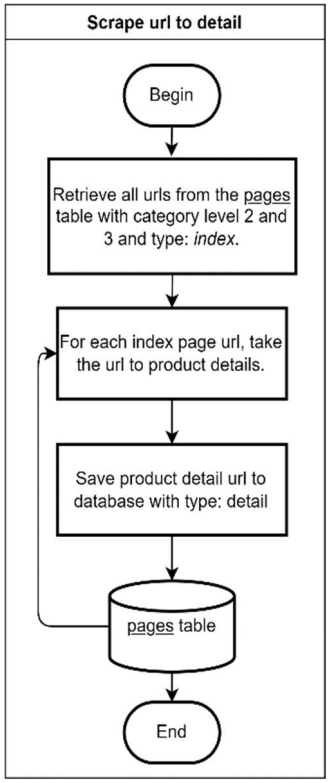

Fig. 5 URL to detail scraping process

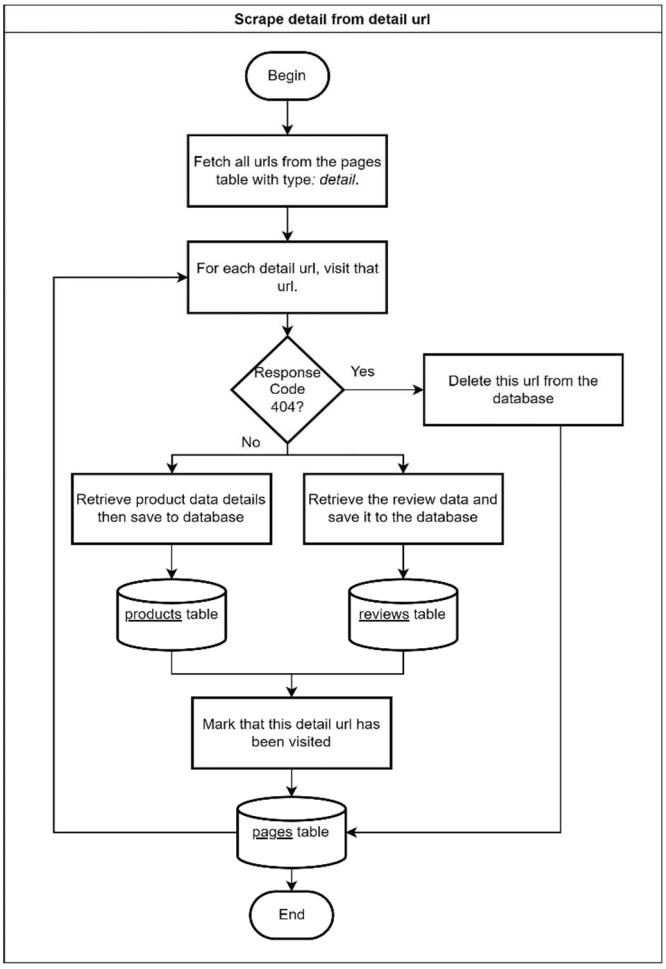

Fig. 6 Detail data scraping process

C. Data Preprocessing and Labelling

The main concept of preprocessing in reviews is taking important words and emojis and separating phrases/sentences. In addition, preprocessing is done to reduce the number of unique words in the review data. Preprocessing is carried out in 2 stages: preprocessing I is carried out for labeling preparation, and preprocessing II is for preparing sentiment analysis model training.

1) Preprocessing I: The following are the steps carried out in preprocessing I.

- Lowercasing, which is changing uppercase letters to lowercase letters [26]. For example, the word "Terima" is changed to "terima". This is done so that the words "Terima" and "terima" can be grouped as the same word.
- Punctuation removal, namely the removal of symbols [21]. Special omission for apostrophes and double quotes (‘, “, ', "). Code ' is the numeric character representation of the apostrophe and " is the numeric character representation of double quotes in ISO 10646 [25], [26]. Additionally, multiple symbol transformations (+, @, #, $, %, ^, *, (,), _, =, [,], {, }, <, >, :, `, ~, ”, “ , …) to a space “ “ is also done. This is done so that if there is a word that is immediately followed by the symbol, it can be separated from other words. For example, in the sentence "barang(produk) ini bagus." after being transformed into “barang produk ini bagus.”. If the transformation objective is empty, it becomes “barangproduk ini bagus.” which is undesirable because the words “barang” and “produk” are combined. As an illustration, see Fig. 7.

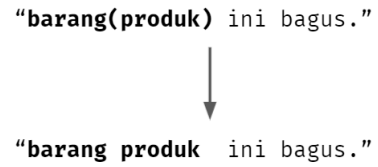

Fig. 7 Example of transforming symbols to spaces

Remember that the main purpose is to minimize the number of unique tokens.

- Emoji reduction, which is the reduction of the same type of emoji from repeating to a single one [27]. See Fig. 8 for the illustration of this stage.

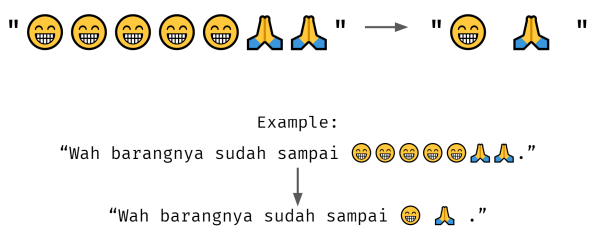

Fig. 8 Example of reducing emoji.

- Character reduction, namely the reduction of characters that are repeated more than 2 times in a word as shown in Fig. 9 [28].

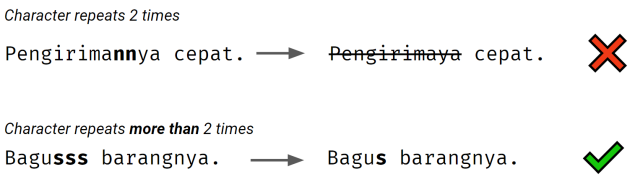

Fig. 9 Character reduction in words

- Punctuation replacement, namely the transformation of some symbols into other words/symbols. The following Table 1 describes some input symbols and their transformation results in the output column.

TABLE I
SYMBOL TRANSFORMATION MAPPING

Input	Output
"&amp"	" dan "
"&"	" dan "
"/"	" atau "
"½"	" setengah "
"-"	" - "

Note that the result of the transformation will add a single space at the beginning and end of the output destination. This is done so that this transformation does not merge with other words.

- Text augmentation, in the form of separation between sentences/phrases in each review. This is done so that the points in the review can be known for their sentiments. Separation is done using these symbols: (period (.), question mark (?), exclamation mark (!), newline (\n), and comma (,)) [29]. The use case is shown in Fig. 10.

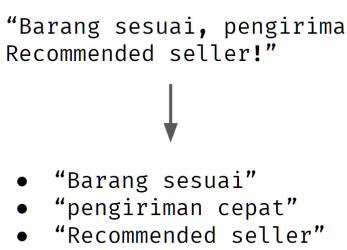

Fig. 10 Example of sentence/phrase separation

- Numbers to words, namely the transformation of numbers into their representation words. The use case is shown in Fig. 11.

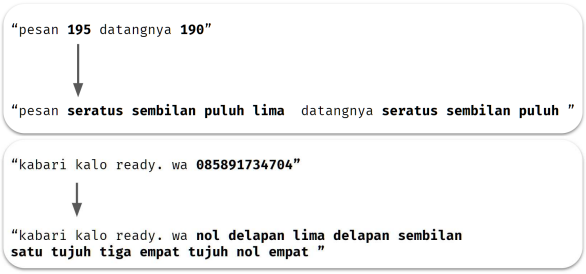

Fig. 11 Numbers transformation to words

See the flowchart in Fig. 12 for the process.

2) Data Labelling: The Sentiment is obtained in 2 ways: using the sentiment classification from GCloud NL and developing a Naive Bayes model. The Naive Bayes model is trained using manually labeled data. The composition of the data after this process can be found in the following Table 2.

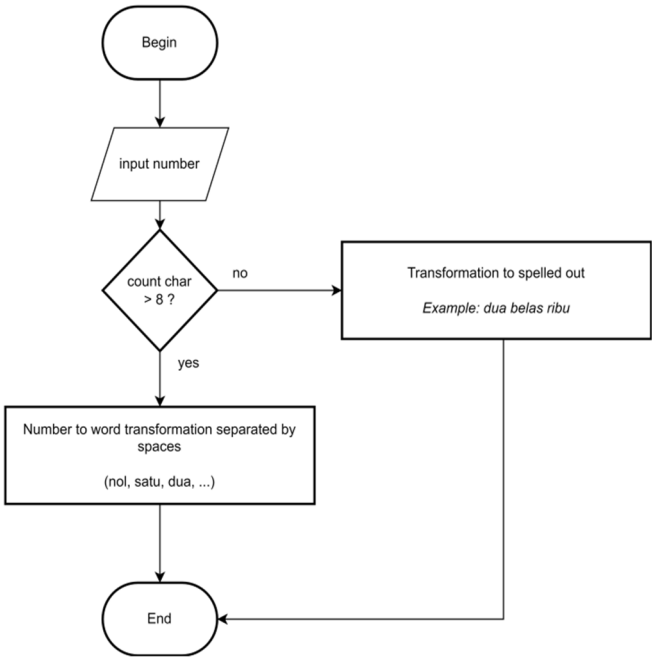

Fig. 12 Number transformation flowchart

TABLE II
DATA COMPOSITION

	Sentiment GCloud NL	Manual Labeling
Total positive sentiment	392539	388159
Total neutral sentiment	13627	5043
Total negative sentiment	35810	48774
Positive ratio/total data	89%	88%
Neutral ratio/total data	3%	1%

For the GCloud NL model, there is no need for labeling because the model has been trained. The flowchart for getting sentiment on reviews using GCloud NL is in Fig. 13. Note that using GCloud NL is solely for a practical solution to directly get sentiment results that will be used by this CSAT assessment. In Fig. 13, the score and magnitude values are limited by a threshold greater than 0.25 by considering the suggestion from the Gcloud NL API service that the threshold value depends on the use case.

1) Preprocessing II: After labeling the data, several stages were carried out to form training, validation, and test data.

- Added completeness of other columns such as url, and category of each review for CSAT score aggregation.
- Specifically for modeling and internal evaluation, data is divided into training data, validation data, and test data with a ratio of 80%, 10%, and 10%, respectively, of the total data and from each sentiment class [30].

Note, this step is not required in model testing using Stratified K-Fold Cross Validation (SKF).

D. Model Training and Testing

1) Model Training: We used dataset to train the Naive Bayes model. Naive Bayes calculations are done in log space,

to avoid underflow and increase speed. The equation is generally expressed in the following (1) [31].

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \log P(c) \quad + \sum_{i \in positions} \log P(w_i|c) \quad (1)$$

Prior or *P(c)* is the percentage of each sentiment class c in the training data. To determine it, the following equation (2) is used.

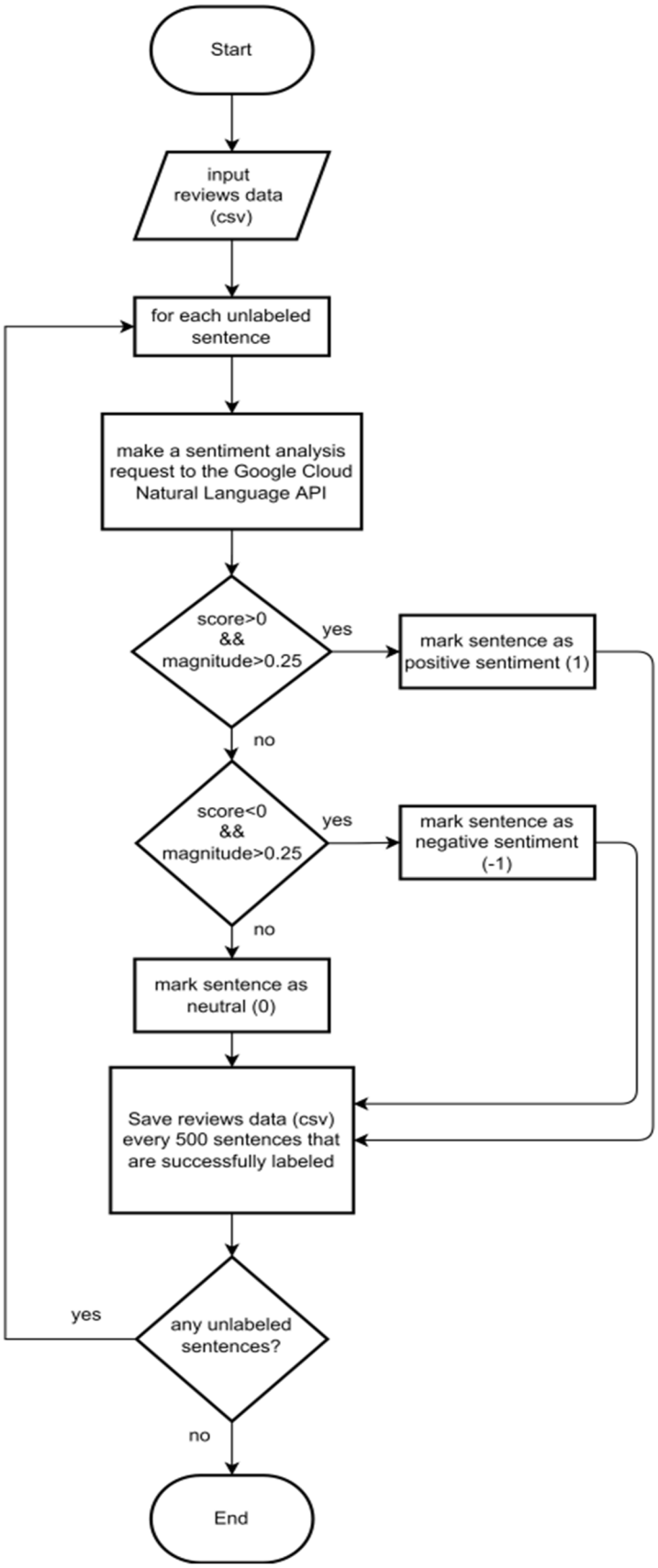

Fig. 13 Flowchart of sentiment prediction using GCloud NL

$$P(c) = \frac{N_c}{N_{doc}} \quad (2)$$

Where N_c is the number of training data with sentiment class c and N_{doc} is the number of training data. Likelihood or $P(w_i|c)$ is assumed to be the ratio between the number of occurrences of a word in the document to the number of words in class *c*, which can be calculated by equation (3) below.

$$P(w_i|c) = \frac{count(w_i,c)}{\sum_{w' \epsilon V} count(w',c)} \quad (3)$$

Where *V*= word of all class c, $count(w_i,c)$ = the number of occurrences of each word with class c, for example: "fantastic"=1, and $count(w',c)$ = the number of all words in class c, for example: 100 positive words out of 1000 words (1/10). But if a word does not appear in the training data, for example the word "fantastic", then the nominee will be worth 0 so that the entire equation is worth 0.

$$P("fantastic"|positive) = \frac{count("fantastic",positive)}{\sum_{w' \epsilon V} count(w',positive)} = 0 \quad (4)$$

This is a problem because Naïve Bayes will generate all likelihoods features together so that if one is 0 it will make the likelihoods of other features 0 as well. This can be solved by adding 1 (Laplace Smoothing), as in equation (4) below,

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

To solve the problem of negation of sentiment, we add the word "NOT_" in front of each word after the word negation. The use of stop words does not improve the performance of the model, so the Naïve Bayes algorithm does not include stop words in Fig. 14 below [32].

2) Model Testing: This model was tested using a Stratified K-Fold Cross Validation (SKF) of 10 Fold. SKF will evaluate the confusion matrix model with metric recall, precision, accuracy, and F1 score. Each fold will produce its metric so the final evaluation is to take the average of the ten folds [32].

function TRAIN NAIVE BAYES(D, C) **returns** log $P(c)$ and log $P(w|c)$

for each class $c \in C$ # Calculate $P(c)$ terms
  N_{doc} = number of documents in D
  N_c = number of documents from D in class c
  $logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$
  $V \leftarrow$ vocabulary of D
  $bigdoc[c] \leftarrow$ **append**(d) **for** d $\in$ D **with** class c
  for each word w in V # Calculate $P(w|c)$ terms
    $count(w,c) \leftarrow$ # of occurrences of w in $bigdoc[c]$
    $loglikelihood[w,c] \leftarrow \log \frac{count(w,c) + 1}{\sum_{w' \text{ in } V} (count(w',c) + 1)}$
return *logprior*, *loglikelihood*, *V*

function TEST NAIVE BAYES(*testdoc*, *logprior*, *loglikelihood*, C, V) **returns** best c

for each class $c \in C$
  $sum[c] \leftarrow logprior[c]$
  for each position i in *testdoc*
    $word \leftarrow testdoc[i]$
    if $word \in V$
      $sum[c] \leftarrow sum[c] + loglikelihood[word,c]$
return $\text{argmax}_c\ sum[c]$

Fig. 14 Naive Bayes algorithm [29]

E. Customer Satisfaction Assessment

The sentiment analysis model is then used to predict sentiment in a review group and then aggregated for the calculation of scores as described in the Fig. 15 as main flowchart, and Fig. 16 to aggregate sentiment for each product.

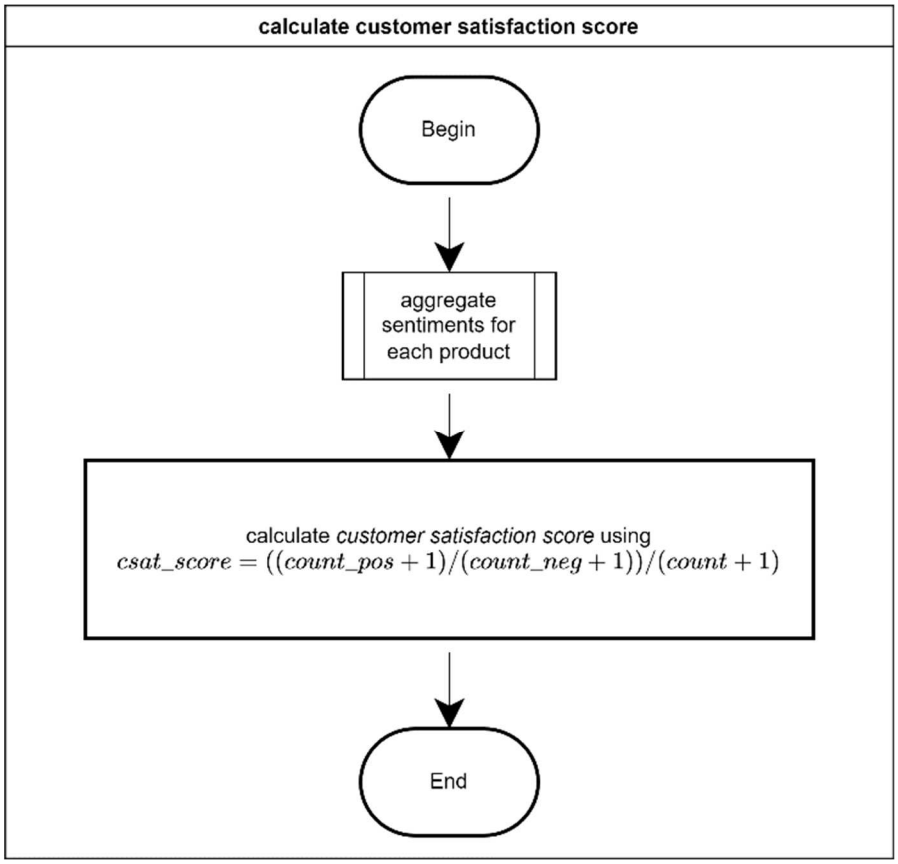

Fig. 15 CSAT assessment main flowchart

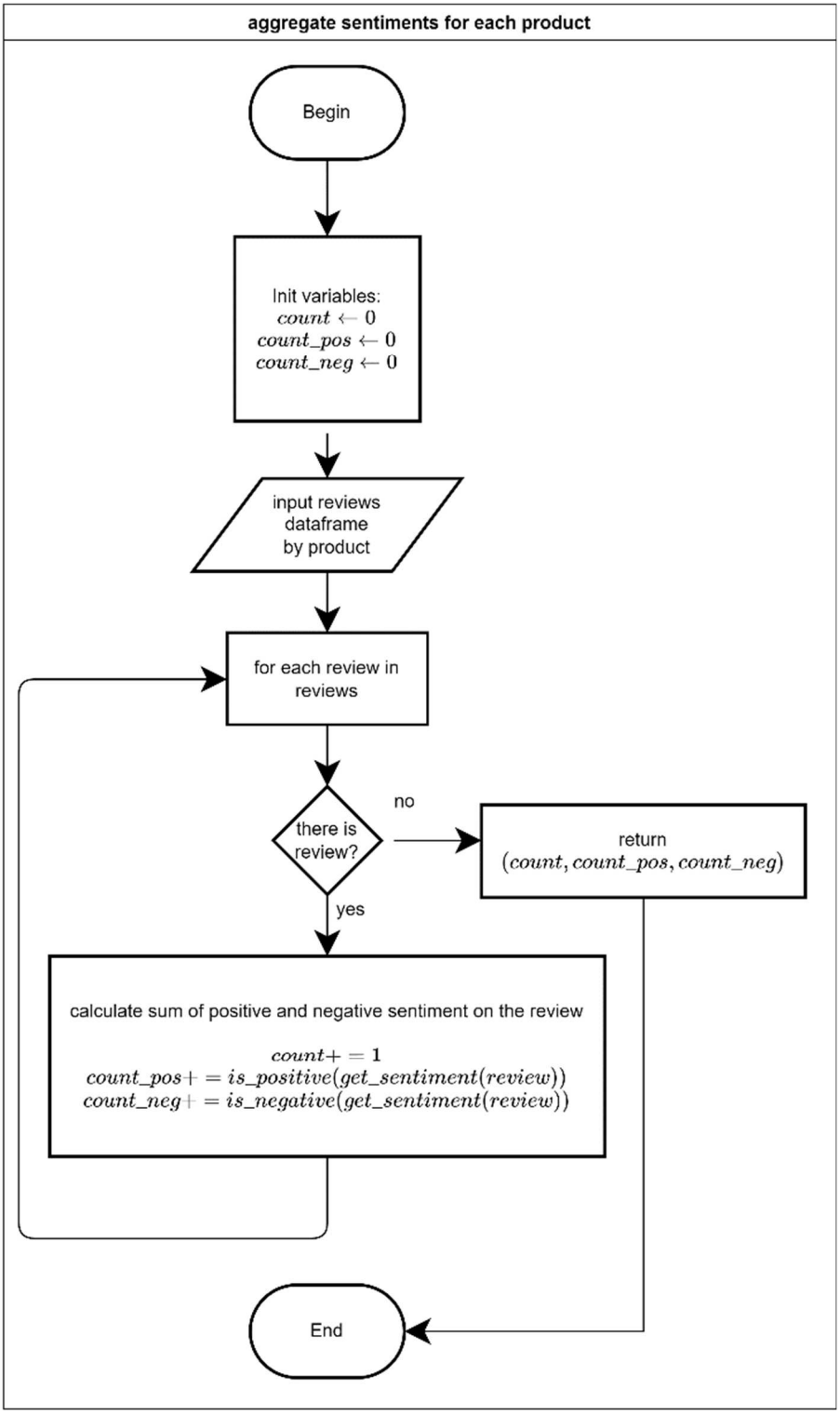

Fig. 16 CSAT assessment sentiment aggregation flowchart

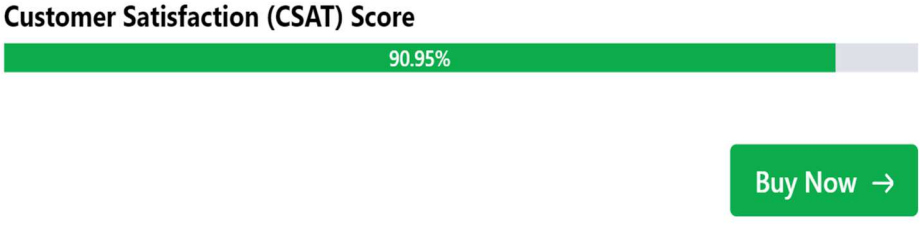

Fig. 17 CSAT assessment summary

The results of this assessment are then stored in a database and displayed on the website. The website that used the Naive Bayes model is at this address: https://nb-recommerce.ardenov.com. The website that used the GCloud NL model at this address: https://gcloud-recommerce.ardenov.com. The following Fig. 17 is a view of the CSAT score from a product.

III. Results and Discussion

A. Model Improvement Experiment

Several experiments were carried out in improving the model:

- In model 1 (with stop words), stop words are used in preprocessing using Indonesian stop words from nltk.org except for the words "tidak", "ada", and the addition of the word "nya".
- In another model 1, no stop words are used at all but still use the same algorithm.
- In model 2, the log prior and loglikelihood equations are modified to the algorithm in Fig. 18. So, the new algorithm becomes:

function TRAIN NAIVE BAYES(D, C) **returns** log *P(c)* and log *P(w|c)*

for each class $c \in C$ # Calculate $P(c)$ terms

N_c = number of documents from D in class c

N_{c_denom} = number of documents from D in class other than c summed

$logprior[c] \leftarrow log \frac{N_c}{N_{c\_denom}}$

$bigdoc[c] \leftarrow$ **append**(d) **for** $d \in D$ **with** class c

$V \leftarrow$ vocabulary of D

for each word $w \in V$ # Calculate $P(w|c)$ terms

for each class $c \in C$

$count(w, c) \leftarrow$ # of occurences of w in $bigdoc[c]$

$prob(w, c) \leftarrow \frac{count(w,c)+1}{(\Sigma_{w' \in V} count(w',c))+|V|}$

for each class $c \in C$

$prob\_w\_denom$ = ratio from $prob(w, c)$ other than c summed

$loglikelihood[w, c] \leftarrow log \frac{prob(w,c)}{prob\_w\_denom}$

return $logprior, loglikelihood, V$

Fig. 18 Update train algorithm on sentiment model

- In model 3, the same algorithm is used as in model 2 with the addition of the prefix "NOT_" on one word after the negation words in the preprocessing [31]. The negation words used are 'tidak', 'nggak', 'not', 'no', 'didnt', dan 'g'.

The results of this experiment are in the following Table 3.

TABLE III
PERFORMANCE COMPARISON BETWEEN SENTIMENT MODELS

	F1	**accuracy**	**recall**	**precision**
model 1 (with stop words)	40%	59%	40%	40%
model 1	43%	61%	42%	43%
model 2	55%	90%	58%	52%
model 3[a]	**57%**	**91%**	**60%**	**55%**
GCloud NL[b]	81%	95%	75%	88%

[a]the final model from scratch used in the CSAT assessment. [b]the pretrained model used in the CSAT assessment.

The comparison of model 3 and GCloud NL confusion matrix can be seen in the Table 4 and Table 5, added with macro average precision and recall metrics [31], [33].

TABLE IV
CONFUSION MATRIX MODEL 3

actual / prediction	**negative**	**neutral**	**positive**	recall	macro average precision
negative	3317.0	2.9	1557.5	68%	
neutral	180.9	4.5	318.9	1%	55%
positive	1812.9	11.9	36991.1	95%	
recall	62%	23%	95%	F1	57%
macro average precision		60%		Accuracy	91%

As seen in Table 4 and Table 5, model 3 has a 1% more precise detection capability on the negative sentiment than the GCloud NL model.

TABLE V
CONFUSION MATRIX GCLOUD NL

actual / prediction	**negative**	**neutral**	**positive**	recall	macro average precision
negative	3257.8	612.5	1007.1	67%	
neutral	0.0	504.3	0.0	100%	88%
positive	323.2	245.0	38246.8	99%	
recall	91%	37%	97%	F1	81%
macro average precision		75%		Accuracy	95%

More detailed information about the 10-fold confusion matrix of both models can be seen in Table 6 and 7.

TABLE VI
EACH FOLD RESULT FROM SKF OF MODEL 3

Iteration	**precision (%)**	**recall (%)**	**accuracy (%)**	**F1**	**negative, negative**	**negative, neutral**	**negative, positive**	**neutral, negative**	**neutral, neutral**	**neutral, positive**	**positive, negative**	**positive, neutral**	**positive, positive**
1	54.82	52.72	91.31	53.753375	4	1499	173	0	331	1819	14	36983	
2	54.93	62.9	91.29	58.643350	3	1525	178	4	322	1816	6	36994	
3	53.57	65.01	91.35	58.743108	2	1768	151	5	348	1545	7	37264	
4	54.74	60.93	91.26	57.673309	3	1565	193	5	307	1783	12	37021	
5	54.87	61.23	91.2	57.873323	2	1552	182	6	317	1820	15	36981	
6	54.69	63.27	91.01	58.673328	3	1546	187	4	314	1917	5	36894	
7	55.02	62.55	91.26	58.553346	1	1530	195	6	303	1823	13	36980	
8	54.53	55.81	91.15	55.163314	4	1559	173	2	329	1834	14	36968	
9	54.88	61.03	91.09	57.793333	2	1542	175	6	323	1883	15	36918	
10	55.28	60.18	91.18	57.633384	5	1489	202	7	295	1889	18	36908	
var	0.21	13.51	0.01	2.78 5994.44	1.43	6077.61	210.1	4.5	234.99	10538.99	19.21	10875.43	
mean	54.73	60.56	91.21	57.453317	2.9	1557.5	180.9	4.5	318.9	1812.9	11.9	36991.1	
std	0.46	3.68	0.11	1.67 77.42	1.2	77.96	14.49	2.12	15.33	102.66	4.38	104.29	

TABLE VII
EACH FOLD RESULT FROM SKF OF GCLOUD NL

Iteration	precision (%)	recall (%)	accuracy (%)	F1	negative, negative	negative, neutral	negative, positive	neutral, negative	neutral, neutral	neutral, positive	positive, negative	positive, neutral	positive, positive
1	88.879	75.331	95.163	81.546	3324	587	967	0	504	0	336	248	38232
2	88.247	75.132	94.977	81.163	3230	604	1044	0	504	0	325	247	38244
3	88.314	75.408	95.09	81.352	3234	620	1024	0	504	0	288	238	38290
4	88.38	75.108	95.05	81.206	3247	627	1003	0	505	0	311	247	38258
5	88.319	74.936	94.984	81.079	3241	623	1013	0	505	0	338	243	38235
6	88.278	75.129	95.002	81.175	3233	615	1029	0	505	0	330	235	38251
7	88.446	75.165	95.034	81.266	3259	601	1017	0	504	0	331	246	38239
8	88.387	74.815	95.004	81.037	3251	637	989	0	504	0	333	249	38234
9	88.796	75.351	95.201	81.523	3307	614	956	0	504	0	294	257	38265
10	88.378	75.032	94.975	81.16	3252	597	1029	0	504	0	346	249	38220
var	0.047	0.035	0.006	0.03	1026.4	231.167	811.433	0	0.233	0	372.178	37.656	405.511
mean	88.443	75.141	95.048	81.251	3257.8	612.5	1007.1	0	504.3	0	323.2	245.9	38246.8
std	0.217	0.186	0.08	0.173	32.037	15.204	28.486	0	0.483	0	19.292	6.136	20.137

A Customer Satisfaction system based on sentiment analysis has been developed in this study using several Naive Bayes models, which are then compared with Google Cloud NLP. This study begins with data acquisition through web scrapping techniques and then continues with text preprocessing. In this stage, the labeling process takes a long time because it is done manually to ensure the data to be classified and can be properly validated. A performance comparison is carried out on four models, which can be seen in Table III. In model 1, testing is carried out in two schemes using a stop word and without a stop word. The results obtained in the second scheme without a stop word give better results. This is because some stop words are quite meaningful in the sentiment class, such as the use of the word "no." Indonesian people often use the word in product reviews, which determines the classification of their sentiments.

Reviews of the first model are still being carried out to improve performance. The strategy used is to modify the equations on log prior and loglikelihood. When this technique is applied, there is a two-fold increase in performance on the F1 score, and accuracy reaches 90%, which can be seen in model 2. Regarding research results obtained, limitations are encountered because this model still needs to improve predicting sentiment if it contains negative sentences. To deal with this problem in model 2, we add a "NOT_" prefix for each word just after the negation word. The negation words are "tidak", "nggak", "not", "didn't", dan "g." The results of this model have a better quality of 91% and can predict sentiment for reviews containing the word negation.

IV. CONCLUSION

Based on the results and discussion of the previous experiment, it was found that the third sentiment model (model 3) has the best performance than the other models but has not been able to exceed the performance of GCloud NL. The Naive Bayes model can be improved because log prior and loglikelihood are modified, stop words are not used, and the "NOT_" prefix is added after each review's negation word. This model must go through 2 stages of preprocessing, namely before and after data labeling. The new Naive Bayes model is selected to make customer satisfaction assessments. This system provides 91% accuracy and an F1 score of 57% in predicting sentiment. In addition, using a self-developed model can better detect negative sentences so that customer dissatisfaction from an e-commerce product purchase transaction is easier to detect.

REFERENCES

[1] Z. Shahbazi and Y.-C. Byun, "Product Recommendation Based on Content-based Filtering Using XGBoost Classifier," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 4, pp. 6979-6988, 2020.

[2] G. Khanvilkar and Prof. Deepali Vora, "Sentiment Analysis for Product Recommendation Using Random Forest," *Int. J. Eng. Technol.*, vol. 7, no. 3.3, pp. 87-89, Jun. 2018, doi: 10.14419/ijet.v7i3.3.14492.

[3] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Bangalore, India, 2016, pp. 416–419. doi: 10.1109/ICATCCT.2016.7912034.

[4] R. Alatrash, R. Priyadarshini, H. Ezaldeen, and A. Alhinnawi, "Augmented language model with deep learning adaptation on sentiment analysis for E-learning recommendation," *Cognitive System Research*, vol. 75, pp. 53-69, Sep. 2022, doi: 10.1016/j.cogsys.2022.07.002.

[5] M. Bibi, W. A. Abbasi, A. Aziz, S. Khalil, M. Uddin, C.Iwendi, and T. R. Gadekallu, "A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis," *Pattern Recognition Letters*, vol. 158, pp. 80-86, June 2022, doi: 10.1016/j.patrec.2022.04.004.

[6] H. Deng, D. Ergu, F. Liu, Y. Cai and B. Ma, "Text sentiment analysis of fusion model based on attention mechanism," *Procedia Computer Science*, vol 199, pp. 741-748, 2022, doi: 10.1016/j.procs.2022.01.092.

[7] A.R. Rahmanti, C.Chien, A.A.Nursetyo, A. Husnayain, B. S. Wiratama, A. Fuad, H- Yang and Y.J. Li, "Social media sentiment analysis to monitor the performance of vaccination coverage during the early phase of the national COVID-19 vaccine rollout," *Computer Methods and Programs in Biomedicine*, vol 221, June 2022, doi: 10.1016/j.cmpb.2022.106838.

[8] I. N. K. Bayu, I. M. A. Suarjaya and P. W. Buana, "Classification of Indonesian Population's Level Happiness on Twitter Data Using N-Gram, Naïve Bayes, and Big Data Technology," *Int. J on Adv Scie Eng Inf Tech*, vol. 12, no. 5, pp. 1944-1949, 2022.

[9] S. Farzadnia and I. R. Vanani, "Identification of opinion trends using sentiment analysis of airlines passengers' reviews," *Journal of Air Transport Management,* vol. 103, August 2022, doi: 10.1016/j.jairtraman.2022.102232.

[10] G. Hanane and N. El Habib, "Semisupervised neural biomedical sense disambiguation approach for aspect-based sentiment analysis on social networks," *Journal of Biomedical Inf*, vol. 135, 2022, doi: 10.1016/j.jbi.2022.104229.

[11] L. Chucu, F. Fan, L. Xu, C. Tie, T. Xu, L. Jianguo, and L. Xin, "Improving sentiment analysis accuracy with emoji embedding," *Journal of Safety Scie and Resilience,* vol. 2, no. 4, pp. 246-252, 2021.

[12] A. Shaha, A. Allulo, A. Asma, A. Amany, A. Sara, A. Nada & A. Aisha, "Customer Satisfaction Measurement using Sentiment Analysis." *International Journal of Advanced Computer Science and Applications*, vol. 9, no 10, 2018, doi:14569/IJACSA.2018.090216.

[13] L. R. Shreyas, "Sentiment Analysis of Customer Satisfaction using Deep Learning," *International Research Journal of Computer Science,* vol.6, no.12, 2019.

[14] A. Nair, C. Paralkar, J. Pandya, Y. Chopra, and D. Krishnan, "Comparative Review on Sentiment analysis-based Recommendation

system," in *6th Int. Conf. Converg. Technol. I2CT 2021*, pp. 2–7, 2021, doi: 10.1109/I2CT51068.2021.9418222.

[15] T. Hariguna and V. Rachmawati, "Community Opinion Sentiment Analysis on Social Media Using Naive Bayes Algorithm Methods," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 2, no. 1, pp. 33–38, 2019, doi: 10.47738/ijiis.v2i1.11.

[16] E. Asani, H. Vahdat-Nejad, and J. Sadri, "Restaurant recommender system based on sentiment analysis," *Mach. Learn. with Appl.,* vol. 6, no. July, p. 100114, 2021, doi: 10.1016/j.mlwa.2021.100114.

[17] M. D. Prasetio, R. Y. Xavier, H. Rachmat, W. Wiyono, and D. S. E. Atmaja, "Sentiment analysis on myindihome user reviews using support vector machine and naïve bayes classifier method," *Int. J. Ind. Optim.*, vol. 2, no. 2, p. 141, 2021, doi: 10.12928/ijio.v2i2.4449.

[18] M. Ahmad, M. Octaviansyah, A. Kardiana, and K. Prasetyo, "Sentiment Analysis System of Indonesian Tweets using Lexicon and Naïve Bayes Approach." in *Fifth International Conf. on Informatics and Computing,* pp.1-4, 2020.

[19] J. Kim, D. Hwang, and H. Jung, "Product recommendation system based user purchase criteria and product reviews," *Int. J. Electr. Comput. Eng. IJECE*, vol. 9, no. 6, p. 5454, Dec. 2019, doi: 10.11591/ijece.v9i6.pp5454-5462.

[20] A. R. Hanni, M. M. Patil, and P. M. Patil, "Summarization of customer reviews for a product on a website using natural language processing," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, India, Sep. 2016, pp. 2280–2285. doi: 10.1109/ICACCI.2016.7732392.

[21] P. Thota and E. Ramez, "Web Scraping of COVID-19 News Stories to Create Datasets for Sentiment and Emotion Analysis," in *the 14th PErvasive Technologies Related to Assistive Environments Conference*, New York, NY, USA, Jun. 2021, pp. 306–314. doi: 10.1145/3453892.3461333.

[22] L. Irshad, L. Yan, and Z. Ma, "Schema-Based JSON Data Stores in Relational Databases," *J. Database Manag. JDM*, vol. 30, no. 3, pp. 38–70, Jul. 2019, doi: 10.4018/JDM.2019070103.

[23] H.-T. Duong and T.-A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput. Soc. Netw.*, vol. 8, no. 1, p. 1, Jan. 2021, doi: 10.1186/s40649-020-00080-x.

[24] F. Alzazah, X. Cheng, and X. Gao, "Predict Market Movements Based on the Sentiment of Financial Video News Sites," in *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, Jan. 2022, pp. 103–110. doi: 10.1109/ICSC52841.2022.00022.

[25] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Oct. 2019, pp. 1–5. doi: 10.1109/ICIC47613.2019.8985884.

[26] N. H. Khun and H. A. Thant, "Visualization of Twitter Sentiment during the Period of US Banned Huawei," in *2019 International Conference on Advanced Information Technologies (ICAIT)*, Nov. 2019, pp. 274–279. doi: 10.1109/AITC.2019.8921014.

[27] Z. Yang, S. Vijlbrief, and N. Okazaki, "TokyoTech_NLP at SemEval-2019 Task 3: Emotion-related Symbols in Emotion Detection," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA, Jun. 2019, pp. 350–354. doi: 10.18653/v1/S19-2061.

[28] H. M. Keerthi Kumar and B. S. Harish, "Classification of Short Text Using Various Preprocessing Techniques: An Empirical Evaluation," in *Recent Findings in Intelligent Computing Techniques*, Singapore, 2018, pp. 19–30. doi: 10.1007/978-981-10-8633-5_3.

[29] H.-H. Chen, "The Contextual Analysis of Chinese Sentences with Punctuation Marks," *Lit. Linguist. Comput.*, vol. 9, no. 4, pp. 281–289, Jan. 1994, doi: 10.1093/llc/9.4.281.

[30] M. Nabil, M. Aly, and A. Atiya, "LABR: A Large-Scale Arabic Sentiment Analysis Benchmark," *Computing and Language*, May 2015. doi: 10.48550/arXiv.1411.6718.

[31] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3d Edition draft. web.stanford.edu, 2020. Accessed: Aug. 02, 2021. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/.

[32] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement," *ACM SIGKDD Explor. Newsl.*, vol. 12, no. 1, pp. 49–57, Nov. 2010, doi: 10.1145/1882471.1882479.

[33] Y. Huang and L. Li, "Naive Bayes classification algorithm based on small sample set," in *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, Sep. 2011, pp. 34–39. doi: 10.1109/CCIS.2011.6045027.