Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 3 (2023) 609 - 617 ISSN Media Electronic: 2580-0760

Implementation of n-gram Methodology to Analyze Sentiment Reviews for Indonesian Chips Purchases in Shopee E-Marketplace

M. Eka Purbaya[1], Diovianto Putra Rakhmadani[2], Maliana Puspa Arum[3], Luthfi Zian Nasifah[4]
[1,2,3,4] Department Of Digital Business, Institut Teknologi Telkom, Purwokerto, Indonesia
[1]m.eka@ittelkom-pwt.ac.id, [2]diovianto@ittelkom-pwt.ac.id, [3]maliana@ittelkom-pwt.ac.id, [4] 21111017@ittelkom-pwt.ac.id

Abstract

Chips are a well-known product among Small and Medium Enterprises (SMEs). In order to enhance the quality of chips as an SME product, sentiment analysis is a crucial step. In this research, sentiment analysis of chip purchases on the Shopee E-marketplace was conducted using the Natural Language Processing (NLP) method, utilizing the N-Gram Model and Term Frequent-Inverse Document Frequency (TF-IDF) as feature extraction techniques, and the Support Vector Machine (SVM) algorithm for sentiment classification. The objective of this research is to identify the most suitable feature extraction model and optimal SVM kernel type from the options of Linear, Polynomial degree, Gaussian RBF, and Sigmoid kernels. Results from the experiments indicate that the TF-IDF and unigram feature extraction techniques offer the best performance for SVM classification when utilizing the Linear kernel. By labeling the dataset, it was observed that using a lexicon-based approach for sentiment classification resulted in 84.31% of the total reviews being positive. The words "price", "cheap" and "quality" in unigram have the highest weights above 0.040. In the unigram model, linear kernel accuracy and precision performance values are 88.4% and 87.3%. At the same time, the recall performance values is 88.4%. The results of the F1-Score assessment matrix from Unigram were 86.9%, Bigram was 78.5% and Trigram was 77.4%. Ultimately, the unigram model combined with a linear kernel in the SVM algorithm demonstrates strong potential for application in the development of various systems focused on detecting user reviews in the Indonesian language on the Shopee E-Marketplace.

Keywords: N-gram; sentiment analysis; shopee; support vector machine

1. Introduction

As of early February 2022, the COVID-19 pandemic continued to persist. To mitigate the risk of COVID-19 transmission, the Indonesian government implemented community activity restrictions and urged employees to work from home (WFH) while adhering to health protocols. These changes have impacted Small and Medium Enterprises (SMEs) from both the buyer and seller perspectives. On the buyer side, there has been a decrease in demand and consumer confidence in products. On the seller side, companies have reduced inventory, production material, and workforce, leading to issues in the supply chain. [1], [2]. Under such circumstances, business entities have devised strategies to ensure that marketing and sales activities remain unhampered by adopting digital technologies, such as selling products via e-commerce platforms and e-marketplaces.

Previous research [3], [4] has demonstrated that e-commerce has a positive impact on enhancing the marketing and sales performance of SMEs [1]. An e-marketplace is a form of e-commerce platform that enables buyers and sellers to interact without the need for in-person meetings. In order to make informed purchasing decisions, prospective buyers rely on product reviews available on e-marketplaces to gain insight into the quality of the product. Meanwhile, sellers can leverage product reviews to improve the quality of their products or services. Analyzing reviews entails reviewing the entire review section to obtain an overall understanding of the review's meaning. For a small number of reviews, manual analysis of individual reviews is feasible, but for a large volume of reviews, sentiment analysis [5], [6] is a faster and more efficient option.

"Keripik," known as chips in English, is a popular processed product among SMEs in Indonesia, offering a wide variety of options such as cassava chips, taro chips, banana chips, spinach chips, and many more. These products are not only available in stores but also on the Shopee e-marketplace. In order to sustain or enhance the quality of chips as an SME product during the pandemic, sentiment analysis is critical.

Accepted: 29-11-2023 | Received in revised: 25-03-2023 | Published: 02-06-2023

Sentiment analysis is a branch of Natural Language Processing (NLP) that focuses on detecting and extracting subjective information from textual data, including emotions, opinions, and attitudes. The primary objective of sentiment analysis is to ascertain the overall polarity of a text, determining whether it is positive, negative, or neutral. This is usually accomplished through a combination of natural language processing techniques and machine learning algorithms, which can analyze and categorize large volumes of text data. The outcomes of sentiment analysis can be applied in a variety of contexts, such as marketing research, customer feedback analysis, and social media monitoring. By utilizing the NLP approach [7], [8] to analyze the sentiment of reviews for purchasing chips products, the system can easily classify reviews into positive or negative categories.

Sentiment analysis has been widely discussed since the publication submitted by [9] in 2002. Bo Pang solves the overall sentiment problem in document classification with the objective of ascertaining whether a document review expresses a positive or negative sentiment. In order to carry out classification, the system needs to utilize machine learning which is part of artificial intelligence. Bo Pang initially conducted experiments using various algorithms, including Naive Bayes, maximum entropy classification, and support vector machines, to classify sentiment in text data. The findings of the study showed that the SVM algorithm achieved the highest average accuracy rate of 81.6%.

Siswanto [10] conducted research on sentiment analysis to determine the accuracy of comments on MotoGP on social media using the SVM and NB algorithms. The study found that SVM outperformed Naive Bayes with an accuracy rate of 95.50%. Rahat [11] used the same algorithm as Siswanto and reported that SVM provided higher accuracy (82.48%) than NB. Sitepu [12] utilized the SVM algorithm to analyze customer sentiment on Shopee and found that SVM yielded an accuracy rate of 97.3%. Meanwhile, Xu [13] compared the performance of Linear SVM and Naïve Bayes algorithms for text classification, and SVM was found to be superior based on Precision, Recall, F1-score, and Accuracy metrics.

Various algorithms can be utilized for sentiment classification, including [14] K-Nearest Neighbor (KNN), [15], [16] Support Vector Machine (SVM), [17] Neural Network, and [18] Naïve Bayes. According to [18], SVM algorithm performs better than Naïve Bayes with an accuracy rate of 93.65%. Meanwhile, [17] found that for SME product sales reviews in Nias Regency, SVM algorithm has the highest accuracy rate of 92%, precision rate of 95%, and recall rate of 82%, which is higher than other algorithms such as Naïve Bayes, K-Nearest Neighbor, and Neural Network.

To implement machine learning models, feature extraction is a crucial step. The Bag of Words model, which employs Term Frequency - Inverse Document Frequency (TF-IDF), is a feature extraction technique that measures the significance of a word in a document by considering its relationship with other words in the document and assigning a weight to each word [19], [20]. In text mining, TF-IDF is a weighting factor that reflects the importance of a word [21]. The value of TF-IDF increases as the word frequency in the document increases, but it is reduced by the frequency of words in the entire corpus. Prastyo et al [22] experimented using TF-IDF to perform sentiment analysis using the SVM algorithm and four kernel configurations (Polynomial, Sigmoid, RBF and Linear). Using training data of 3200 tweets and 200 tweets for data testing, a comparison of SVM performance is carried out through scenarios using all features of TF-IDF, namely 1000, 2000, 3000 and 4000 features. The results obtained from this experiment are Polynomial kernel when using 1000 features, gives an accuracy value of 92.03%, which decreases when added features. In contrast, using only 2000 features results in the highest accuracy value, which is achieved by utilizing the RBF kernel. From this, it can be concluded that more features do not always indicate better machine learning algorithm performance.

Purbaya et al. [23] conducted experiments to analyze the performance of SVM kernel using the TF-IDF technique. Their results showed that the linear kernel in SVM performed the best in testing with an accuracy and recall value of 89.60% and an F1-Score value of 88.60%. However, their research did not utilize the n-gram technique to assist in sentiment analysis for reviews on purchasing chips.

There have been multiple methods developed to classify text documents, and one of these is known as character n-gram. This technique is known for its efficiency, reliability, and speed, and it is capable of handling textual errors. Essentially, character n-gram involves analyzing all the different n-character substrings that exist within a given string [24]. Nasser [24] conducted research that applied the TF-IDF method along with the addition of the n-gram technique to detect COVID-19 patients. The outcome of this study showed a significant enhancement in the vector space model (TF-IDF) by using a predetermined number of n-grams.

Previous research has suggested that the combination of TF-IDF with n-gram technique and SVM algorithms holds promise as a machine learning model for sentiment analysis. In this study, the researchers will investigate the efficacy of this approach by applying the TF-IDF model with n-gram technique for feature extraction and SVM algorithm for machine learning to analyze sentiment in chip purchase reviews on the Shopee E-Marketplace.

The study aims to determine the optimal n-gram technique in the SVM kernel, including the Linear, Polynomial degree, Gaussian RBF, and Sigmoid kernels. Additionally, the research will identify the main features mentioned in the reviews of "chips" products. The study's anticipated outcomes are recommendations on sentiment categories, prominent features in buying chips reviews on the Shopee e-marketplace, and an evaluation of the classifier's performance in conducting sentiment analysis on such reviews.

2. Research Methods

The research method is carried out through a series of paths as shown in Figure 1.

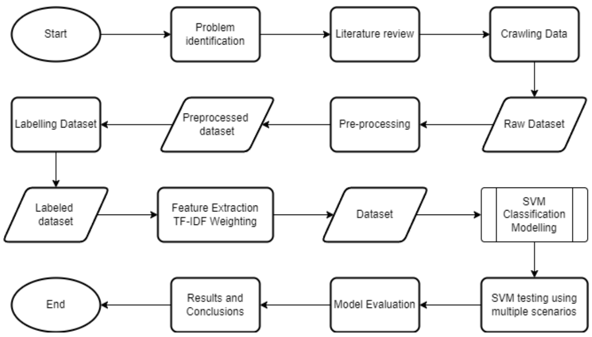

Figure 1. Research Stage

2.1 Data

The data collection process for purchasing chip reviews on the Shopee e-marketplace involves several stages executed in Google Collaboratory using the Python programming language. The product selection was based on the keyword "keripik" entered in the Shopee search, and a total of 30 products were selected based on being the top search results and with sales exceeding 150, to ensure a substantial amount of reviews. This selection process follows the central limit theory and assumes that the data is normally distributed [25]. Table 1 and 2 illustrates the data mining process for purchase review data for each selected product.

Once the product list is obtained, its link is fed into a Web Scraper program developed using Python to extract data on all reviews of the product, and create a raw dataset in CSV file format for each product. To remove any rows with empty reviews and replace null username columns with "NN", a cleaning process is performed. After cleaning is completed for each product, the raw datasets are combined for further preprocessing.

Table 1. List product of object analysis sentiment

No	Seller	List Product
1	cemilansukaku	Keripik Kaca Kemasan Box 500ml / Keripik Beling Vi…
2	fmjaya_collection	Kripik pinoh / Keripik Ubi / Kripik singkong…
3	kia.rayra	KERIPIK KITA | MAKANAN RINGAN | CEMILAN | JAJANAN …
4	mostwantedite m	Keripik Ubi Pedas Polos Oleh Oleh Khas Manado 100 …
5	kia.rayra	[DIST] CEMILAN | KERIPIK KITA | KRIPIK | ASIN | PE…
6	albarrsnack0056	keripik kaca 50 gr…
7	addarsnack	Keripik Basreng Bumbu Basah Daun Jeruk Halal Netto…
8	mandirisukses. officialstore	KERIPIK BUAH APEL / MANGGA / TAPE / RAMBUTAN ~ KER…
9	duo_bocil_snack	Keripik Tahu Bulat Gurih Enak 250 Gram - Duo Bocil…
10	snackcadaz	KERIPIK BAYAM 250 gram / KRIPIK PEYEK DAUN BAYAM C…
11	keripik.kentang .mama.hani	Keripik Kentang Ebi 450 Gram Homemade Enak Berkual…
12	snackbilqis	Zanana Chips 80 gram Keripik Pisang Makanan Sehat …
13	cemilankitayu mmy	Cemilan keripik jingkar kita - makanan ringan - sn…
14	makaronihaho	Kripik setan / kripik pedas haho versi ecer 60gr…
15	keripikkentang garutnastra	Keripik Kentang keju /Termurah / Harga Pabrik / Na…

Table 2. List product of object analysis sentiment (continuation)

No	Seller	List Product
16	rubybeauty30	Kripik bawang Dimar / Kripik Renyah/ Keripik Gurih…
17	ngabret.id	KERIPIK KACA / KERIPIK BELING / ELOD BANDUNG…
18	tokokuesumber mas	KERIPIK USUS 250gr / KRIPIK USUS AYAM / KERIPIK US…
19	nsfood_snacks	KERIPIK KACA PEDAS ORIGINAL DAUN JERUK KRIPCA CEMI…
20	tokokuesumber mas	SINGKONG BALADO TES 250gr / KERIPIK SINGKONG PEDAS…
21	jogjamushroom	kripik jamur tiram. kancing dan merang khas jejamu…
22	cemilankitayum my	CEMILAN | KERIPIK KITA | KRIPIK | PEDAS | ASIN | J…
23	arkajaya_keripi k	KERIPIK PISANG LAMPUNG BERAT 500 gram ENAK & TERMU…
24	yunita_snack	Kulit ayam crispy pedas daun jeruk kemasan pouch 1…
25	nsfood_snacks	COMRING KERIPIK COMRING COMRO KERING MAKANAN KHAS …
26	trendifashion	Keripik pisang coklat lampung . oleh oleh khas lam…
27	golden.cakery	Keripik Pisang Nangka KOIN - Ekonomis 170 gr…
28	widia_krisna_w ati	KERIPIK SAYUR MIX TOPLES 800 ML(wortel,buncis,edam…
29	cemilan_bizee	KERIPIK KACA VIRAL | EXTRA PEDAS DAUN JERUK | KRIP…
30	keripik.kita	(AGEN) CEMILAN KITA KERIPIK KERUPUK ASIN PEDAS DAU…

DOI: https://doi.org/10.29207/resti.v7i3.4726
Creative Commons Attribution 4.0 International License (CC BY 4.0)

2.2 . Stage of pre-processing the data

The pre-processing stage involves preparing text data by removing unnecessary information to enable sentiment analysis [26].

This study employs several techniques, including:

Case folding refers to the process of converting the characters in a word to their basic form, which typically involves converting all characters to lowercase to achieve consistency in the shape of the letters.

Cleansing is the process of removing irrelevant components from the documents that do not contribute to the information contained in the document, such as URLs, hashtags (#), usernames (@usernames), emails, emoticons (:@, :*, :D), punctuation marks like commas (,), periods (.) and other punctuation marks.

Tokenizing is the process of dividing text data into multiple tokens. This technique breaks down a sequence of characters in a text into individual words, taking into account certain characters that can be treated as word delimiters or not.

Filtering is the process of eliminating insignificant words from the tokenization phase. Stopwords, such as adverbs, exclamations, pronouns, prepositions, and conjunctions, will be removed. Additionally, regular expressions are used to remove punctuation marks such as “\.|,|:|;|!|\?|\(|\)|\||\+|'|"|‘|’|“|”|'|\’|…|\-|–|—|\$|&|*|>|<” [27].

Stemming is the stage where the root word of each word obtained from the filtering process is identified. It involves converting different word forms to their basic forms by removing prefixes and suffixes.

Normalization is a stage where non-standard words are altered to conform to the standard spelling. It involves the use of a dictionary comprising standard and non-standard words, which is created based on the comments data used in this study.

2.3. Feature Extraction

Once the pre-processing stage is completed using NLP approach, the next step involves computing the weight of words in a review. The Term Frequency Inverse Document Frequency (TF-IDF) algorithm is applied to determine the weight [28]. TF-IDF provides a weight to a word or term based on its relationship with the entire review. The frequency of the word in a review shows its significance in the review [16], and which reviews contain that word, allowing classification into positive and negative reviews. The TF-IDF computation is done using formula 1.

$$W_{x,y} = tf_{x,y} \,.\log\left(\frac{N}{dfx}\right) \quad (1)$$

Formula 1, shows that the weight of a term (t_y) against documents (d_x) is represented by W_x. The number of occurrences of the term (t_y) in documents (d_x) is represented by $tf_{x,}$. N is the total number of documents in the dataset, and df_x is the number of documents containing the term (t_y), at least once.

The N-Gram model is a statistical model used in natural language processing to predict the probability that a word will appear based on the previous word order in a text or sentence. In the N-Gram model, text or sentences are divided into a series of n-grams, consisting of n consecutive words. For example, if you use n = 3, then the n-gram will consist of three words that are close together in a text or sentence. Then, the model will calculate the frequency of occurrence of n-grams in a text or large corpus and use this information to predict the next possible word in a sentence or text. This model can be used for various tasks such as sentiment analysis, text classification and speech recognition [29]. The classification of N-Gram in this research is categorized into three types. The first type is Unigram, which consists of a single word token. The second type is Bigram, which comprises of two-word tokens. The third type is Trigram, which is made up

2.4 Learning, Training and Testing Model

As stated by Mutawwali et al. [30], Support Vector Machine (SVM) is a method of classifying data that uses previous data and supervised modeling. SVM is a type of binary, linear, and non-probabilistic classifier. The decision boundary used by SVM determines the classification of the training data and enables the formation of an optimal linear or hyperplane model to classify the data.

In real-world scenarios, there is often a challenge because linear classification is rarely separable. As a solution to this non-linear problem, SVM has been modified by incorporating Kernel function [31]. These functions help to transform data points to a higher-dimensional space, improving the algorithm's ability to identify the optimal hyperplane to separate data points of different classes. There are various Kernel functions that can be employed in SVM, including:

The SVM Linear Kernel is a flexible margin that aims to identify a hyperplane with a linear pattern while also allowing for one or more data misclassifications. Even though it permits certain errors, the linear kernel still attempts to identify a line that maximizes margins and minimizes misclassifications. The accuracy of the hyperplane is influenced by the number of tolerable misclassifications given. The linear kernel is the most basic kernel function, which is the dot product of two vectors as seen in formula 2.

$$K(x, z) = x.z + C \quad (2)$$

DOI: https://doi.org/10.29207/resti.v7i3.4726
Creative Commons Attribution 4.0 International License (CC BY 4.0)

Polynomial Kernel in SVM, if the data cannot be separated by straight lines, kernel polynomials are needed. These can create nonlinear decision boundaries. The kernel polynomials generate new features by combining existing features using polynomials as seen in formula 3.

$$K(x,z) = (\gamma x.z + C)^d, d > 0 \quad (3)$$

Radial Basis Function Kernel in SVM, in training datasets with the RBF kernel, two parameters must be taken into account: C and gamma. The C parameter is used to specify the amount of error to be avoided when classifying training data. A higher C value leads to a lower classification error for the training data. On the other hand, the gamma parameter determines the influence of a single training data sample. A smaller gamma value indicates a greater distance between the data point being calculated and the training data as seen in formula 4.

$$K(x,z) = exp(-\gamma\|x - z\|^2), \gamma > 0 \quad (4)$$

Sigmoid Kernel in SVM, the sigmoid kernel is a type of kernel function used in support vector machines (SVM) for classification problems. It is a non-linear kernel that transforms the data into a higher dimensional space to enable the creation of non-linear decision boundaries. The sigmoid kernel takes the form of a sigmoid function, which is a mathematical function that maps input values to a continuous range between 0 and 1. In SVM, the sigmoid kernel is often used for binary classification problems and can be useful for problems where the data is not easily separable by linear decision boundaries. However, the sigmoid kernel can be sensitive to the choice of its parameters and may not perform as well as other kernel functions such as the radial basis function (RBF) kernel as seen in formula 5.

$$K(x,z) = tanh(\gamma x.z + C) \quad (5)$$

2.5 Model Evaluation

The evaluation of performance is a crucial step to assess the effectiveness of a proposed model and draw conclusions from the conducted study. One of the commonly used methods to measure the error of a predictive model is K-fold cross-validation. This method is similar to repeated random subsampling, but it ensures that there is no overlap between any two test sets. In K-fold cross-validation, the training data is divided into k subsets of equal size.

In this context, the term "fold" refers to the number of divisions that the training data is separated into. The subsets are created randomly from the training data without any duplication. One of the k-1 subsets is used as the validation set for training the model. The remaining subset is used as the test set to evaluate the model's performance. This procedure is repeated until each of the k subsets has been utilized as the validation set [32].

An important challenge that can be encountered is when the data is imbalanced, meaning that one class may have a significantly higher number of observations than the other, which can negatively impact the accuracy of the model. The accuracy of the model may be biased towards the majority class. To overcome this challenge, a useful tool is the confusion matrix, which helps to visually represent the performance of a classification algorithm based on the data included in the matrix. [33].

The Confusion Matrix is a method used to evaluate the performance of machine learning algorithms by comparing their predicted values with the actual conditions of the data. Its purpose is to measure the accuracy, precision, and recall of a classification method.

Table 3 presents a confusion matrix for binary classification that compares the distribution of the actual data and the predicted data generated by the model. There are four types of confusion matrices that can be used to calculate performance metrics such as True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN).

Table 3. Confusion Matrix

Actual	Prediction	
	Positive	Negative
Positive	(TP) True Positive	(FN) False Negative
Negative	(FP) False Positive	(TN) True Negative

True Positive (TP) represents the proportion of predicted positive data that is also positive in actual values. False Positive (FP) is the percentage of data that is incorrectly predicted by the system as positive when the actual values are negative. False Negative (FN) shows how much of the data is predicted as negative by the system when it is actually positive. True Negative (TN) indicates the percentage of negative data predicted by the system whose actual values are negative. These components are then used to calculate the classification performance metrics such as accuracy, precision, recall, and F1-Score. Accuracy is the ratio of correct system predictions to the total number of prediction results. Formula 6 shows the formula for computing the accuracy value.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (6)$$

Precision is a metric that calculates the ratio of the correctly predicted positive values (TP) to the overall positive predictions (TP and FP). In other words, it measures the accuracy of positive predictions made by the system. The formula to calculate precision is shown in formula 7.

DOI: https://doi.org/10.29207/resti.v7i3.4726
Creative Commons Attribution 4.0 International License (CC BY 4.0)

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

The next metric is called recall, which aims to measure the ratio of predicted positive data to all actual positive data. Formula 8 shows the mathematical expression used to compute the recall value.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

The F1-Score is a metric used to evaluate the performance of a classification algorithm and is a measure of the balance between precision and recall. It is the harmonic mean of precision and recall, as shown in Formula 9.

$$F1 - Score = \frac{2 \times precission \times recall}{precission+recall} \quad (9)$$

3. Results and Discussions

3.1 Labelling Dataset

After crawling and cleaning the data, a raw dataset containing 7757 reviews was obtained. The raw reviews were pre-processed to remove redundant words and facilitate feature extraction using applications compatible with Orange Data Mining. The dataset was then labeled using a lexicon-based technique, using an Indonesian sentiment dictionary developed by Fajri [34]. A sample of the labeled data is presented in Table 4.

Table 4. The example of the dataset

Tagging	Review
Negative	"Keripiknya sudah sampai, rasanya enak cuma kurang berasa anggurnya, tapi not bad lah, pengiriman juga cepat"
Negative	"Enak nyesel beli dikit besok2 beli lagi yang banyak terima kasih yah"
Positive	"Produk sesuai dengan pesanan, masih belum di coba, semoga seperti yang diharapkan."
Positive	"Rasa enak kualitas ok harga terjangkau seller ramah pengiriman cepat"

The labeled dataset provides insights into the sentiment classification results as illustrated in Figure 2. The figure shows that 84.31% of reviews are positive, while 15.69% are negative.

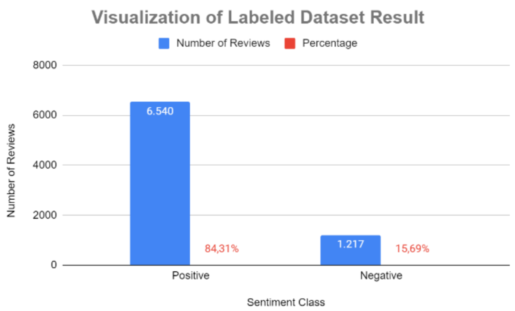

Figure 2. Visualization of labeled dataset results

Following the pre-processing phase, the relevant words were visualized through a wordcloud, as depicted in Figures 3, 4 and 5. The most frequently used word in the reviews of buying chips on Shopee e-marketplace in Indonesian is "harga" or price in English Language, followed by "enak" or tasty, and "kualitas" or quality. These findings were obtained from the meaningful words after the pre-processing stage. As for the results of the bigram feature, the top-ranking factor is a low price, indicating that customers consider the product's price as the primary factor. Moreover, the speed of delivery and taste are also essential factors that customers prioritize in their reviews.

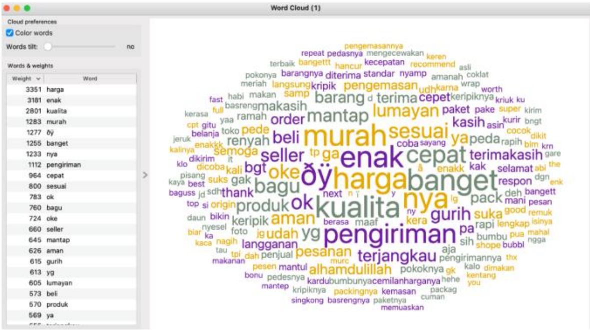

Figure 3. Unigram wordcloud result

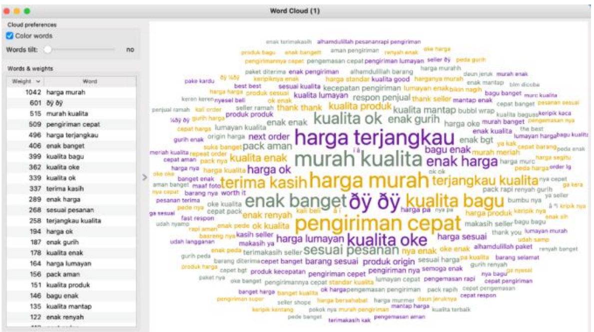

Figure 4. Bigram wordcloud result

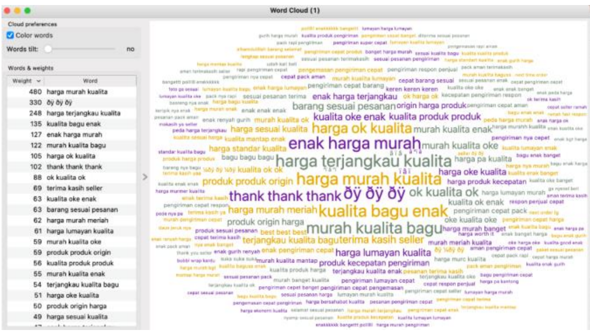

Figure 5. Trigram wordcloud result

3.2 TF-IDF and N-Gram Method for Feature Extraction

In this phase, the relationship between the words/terms and the document is determined by assigning a weight to each word. The significance of a word in a document is evaluated using the TF-IDF method. To identify the most suitable feature extraction model and

Creative Commons Attribution 4.0 International License (CC BY 4.0)

obtain the best results, the N-Gram method is used, incorporating Unigram, Bigram, and Trigram. The feature extraction process generates the word order and frequency of occurrences throughout the dataset. The TF-IDF and N-Gram feature extraction results are shown in Figure 6-8 using Orange Data Mining tools. The figure indicates that the words "price," "cheap," and "quality" carry the highest weights above 0.040. There is not much difference between the bigram and trigram features as they hold the same meaning for the words "price," "cheap," and "quality." Through figure 6 , 7 and 8, it can also be seen that the weight of words in unigrams is much higher than the weights in bigrams and trigrams. Based on this, a hypothesis arises that the classification accuracy results from unigram feature extraction will have a better value than the results from bigram and trigram feature extraction.

Word	TF-IDF
harga	0.061
murah	0.053
kualita	0.044
ok	0.043
enak	0.040
oke	0.032
terjangkau	0.031
banget	0.029
mantap	0.028
gurih	0.026
pengiriman	0.024
cepat	0.023
sesuai	0.023
ðÿ	0.022
nya	0.021
lumayan	0.020
bagu	0.020

Figure 6. Feature extraction using TF-IDF (Unigram)

Word	TF-IDF
harga murah	0.046
harga terjangkau	0.026
kualita ok	0.020
kualita oke	0.019
enak banget	0.012
harga ok	0.012
pengiriman cepat	0.012
murah kualita	0.010
harga lumayan	0.009
kualita bagu	0.009
ðÿ ðÿ	0.008
terima kasih	0.008
enak gurih	0.008
kualita mantap	0.008
enak harga	0.008
sesuai pesanan	0.007

Figure 7. Feature extraction using TF-IDF (Bigram)

Word	TF-IDF
harga murah kualita	0.009
harga terjangkau kualita	0.005
ðÿ ðÿ ðÿ	0.004
enak harga murah	0.004
kualita bagu enak	0.004
harga murah meriah	0.003
murah kualita bagu	0.003
harga ok kualita	0.003
terima kasih seller	0.003
ok kualita ok	0.002
produk produk origin	0.002
kualita produk produk	0.002
murah kualita enak	0.002
terjangkau kualita bagu	0.002
produk kecepatan pengiriman	0.002

Figure 8. Feature extraction using TF-IDF (Trigram)

3.3 Model Evaluation

The Orange Data Mining tools are utilized to conduct the research, and the widgets or components in the system are tailored to align with the scenarios outlined in Figure 9. Following testing, an evaluation of the model is required to assess the performance of each kernel in the SVM. The research's overall process is carried out with the aid of Orange Data Mining tools. The widgets or components within the system are modified to match the scenarios previously planned in Figure 9. Post-testing, an evaluation of the model is essential to gauge the performance of each kernel within the SVM.

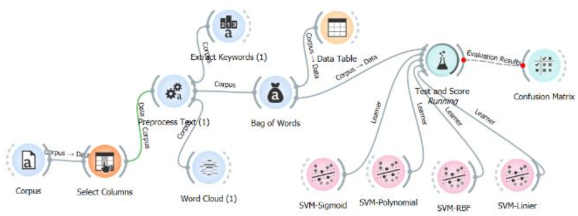

Figure 9. Widget component

Following the formation of features using TF-IDF, the SVM algorithm is employed to model the features through four kernel usage scenarios. The objective is to forecast whether the purchase reviews of chips belong to a positive or negative label. The data modeling procedure utilizes the stratified 10-fold Cross Validation mechanism, which separates the training and testing data into ten groups with comparable data class proportions to the original dataset. The employment of stratified 10-fold cross-validation is expected to produce the highest level of performance according to the proposed SVM algorithm.

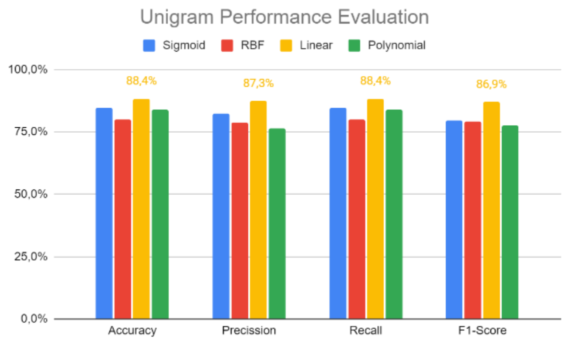

Figure 10. Unigram performance evaluation

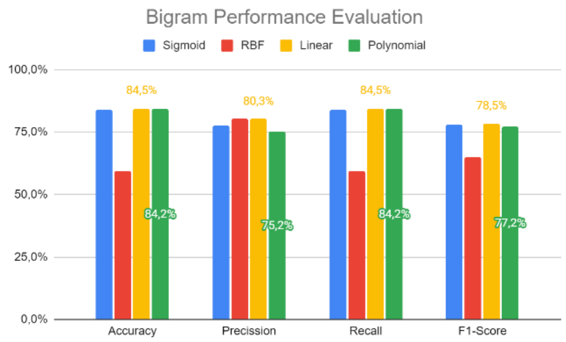

Figure 11. Bigram performance evaluation

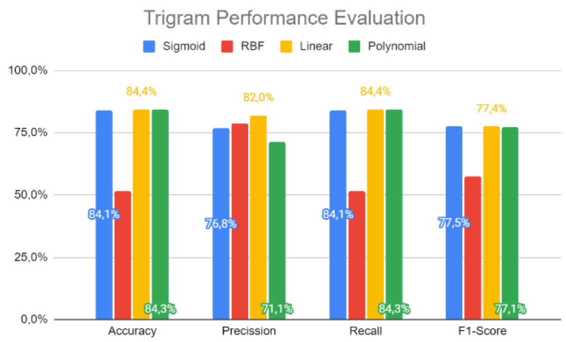

Figure 12. Trigram performance evaluation

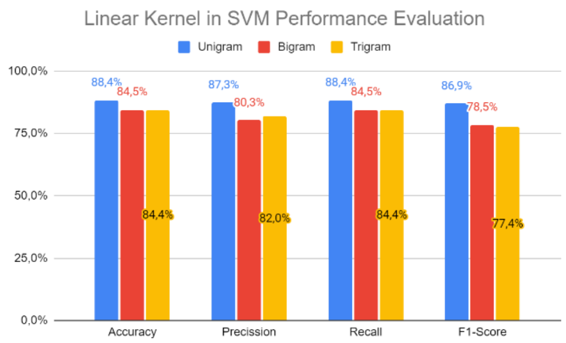

Figure 13. Linear Kernel performance evaluation

Figures 10, 11, 12, and 13 illustrate the findings of the comparison of the proposed model's performance in this study. The Linear kernel in SVM has shown the most outstanding accuracy, precision, recall, and F1-Score evaluation metrics in the Unigram, Bigram, and Trigram environments for the sentiment analysis of chips purchase reviews on Shopee's E-Marketplace. Figure 13 reveals that the Unigram model is the most effective feature extraction method, enhancing the Linear Kernel's performance compared to the Bigram and Trigram models with respective accuracy, precision, recall, and F1-Score values of 88.4%, 87.3%, 88.4%, and 86.9%.

4. Conclusion

This study conducted a comparison experiment using kernel support vector machines to analyze the sentiment of chip purchase reviews. The dataset used in this study was collected using a scrapping technique on the chip purchase reviews from the Shopee E-Marketplace, with 84.31% of the 7757 reviews categorized as positive. The SVM algorithm was the primary focus of this research and performed well when modeled using the Linear kernel. Feature extraction using the TF-IDF method and N-Gram model indicated that the Unigram technique outperformed the Bigram and Trigram methods, with performance values of 88.4% accuracy, 87.3% precision, 88.4% recall, and 86.9% F1-Score.

Based on the findings of this research, it is evident that the Linear kernel in the SVM algorithm, when coupled with the TF-IDF feature extraction technique using the Unigram model, shows great potential for developing various systems related to detecting user reviews on the Indonesian-language Shopee E-Marketplace. Kernel functions play a crucial role in solving non-linear problems and assist SVM in identifying the optimal hyperplane. Further experiments could involve parameter optimization of SVM in analyzing the sentiment of purchase reviews on the Shopee E-Marketplace.

References

[1] W. laura Hardilawati, "Strategi Bertahan UMKM di Tengah Pandemi Covid-19," *Jurnal Akuntansi dan Ekonomika*, vol. 10, no. 1, pp. 89–98, 2020, doi: 10.37859/jae.v10i1.1934.

[2] D. A. I. dan E. Y. Devin Ananda D. S, "Kajian Strategi Pengembangan UMKM Dalam Menghadapi Era Digital (Studi Kasus UMKM Keripik Apel Delicious Kota Batu)," vol. 1, no. 1, pp. 19–27, 2021.

[3] S. Sandri and W. laura Hardilawati, "Model Pemasaran Hubungan Pelanggan, Inovasi Dan E-Commerce Dalam Meningkatkan Kinerja Pemasaran Ukm Di Pekanbaru," *Jurnal Akuntansi dan Ekonomika*, vol. 2, no. 1, pp. 20–42, 2019.

[4] D. Setyorini, E. Nurhayaty, and R. Rosmita, "PENGARUH TRANSAKSI ONLINE (e-Commerce) TERHADAP PENINGKATAN LABA UMKM (Studi Kasus UMKM Pengolahan Besi Ciampea Bogor Jawa Barat)," *Jurnal Mitra Manajemen*, vol. 3, no. 5, pp. 501–509, May 2019, doi: 10.52160/ejmm.v3i5.228.

[5] E. H. Muktafin, K. Kusrini, and E. T. Luthfi, "Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing," *Jurnal Eksplora Informatika*, vol. 10, no. 1, pp. 32–42, 2020, doi: 10.30864/eksplora.v10i1.390.

DOI: https://doi.org/10.29207/resti.v7i3.4726
Creative Commons Attribution 4.0 International License (CC BY 4.0)

[6] P. Gentsch, *AI Business: Framework and Maturity Model*. 2019. doi: 10.1007/978-3-319-89957-2_3.

[7] R. Kibble, "Introduction to natural language processing Undergraduate study in Computing and related programmes," *Roeper Rev*, vol. 1, no. 2, p. 26, 2013.

[8] L. T. Vo, *Mining Social Media - Finding Stories in Internet Data*. Wlliam Pollock, 2020.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002*. pp. 79–86, 2002.

[10] Siswanto, Y. P. Wibawa, W. Gata, G. Gata, and N. Kusumawardhani, "Classification Analysis of MotoGP Comments on Media Social Twitter Using Algorithm Support Vector Machine and Naive Bayes," in *2018 International Conference on Applied Information Technology and Innovation (ICAITI)*, 2018, pp. 96–101. doi: 10.1109/ICAITI.2018.8686751.

[11] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," *Proceedings of the 2019 8th International Conference on System Modeling and Advancement in Research Trends, SMART 2019*, pp. 266–270, 2020, doi: 10.1109/SMART46866.2019.9117512.

[12] M. B. Sitepu, I. R. Munthe, and S. Z. Harahap, "Implementation of Support Vector Machine Algorithm for Shopee Customer Sentiment Anlysis," *Sinkron*, vol. 7, no. 2, pp. 619–627, 2022, doi: 10.33395/sinkron.v7i2.11408.

[13] H. Xu and Y. Lv, "Mining and Application of Tourism Online Review Text Based on Natural Language Processing and Text Classification Technology," *Wirel Commun Mob Comput*, vol. 2022, 2022, doi: 10.1155/2022/9905114.

[14] Y. M. Febrianti, I. Indriati, and A. W. Widodo, "Analisis Sentimen Pada Ulasan 'Lazada ' Berbahasa Indonesia Menggunakan K-Nearest Neighbor (K-NN) Dengan Perbaikan Kata Menggunakan Jaro Winkler Distance," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 10, pp. 3689–3698, 2018.

[15] S. Yuan, A. Pratiwi, and S. R. Nudin, "Analisis Sentimen terhadap Facebook Marketplace Menggunakan Metode Lexicon Based dan Support Vector Machine," vol. 3, pp. 9–15, 2021.

[16] B. Pamungkas, M. E. Purbaya, and D. J. A.K, "Analisis Sentimen Twitter Menggunakan Metode Support Vector Machine (SVM) pada Kasus Benih Lobster 2020," *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, vol. 3, no. 2, pp. 10–20, 2021, doi: 10.20895/inista.v3i2.243.

[17] M. Harahap, B. P. A. Sihombing, O. A. F. Laia, B. T. Saragih, and K. Dharma, "Analisis Sentimen Review Penjualan Produk Umkm Pada Kabupaten Nias Dengan Komparasi Algoritma Klasifikasi Machine Learning," *METHOMIKA Jurnal Manajemen Informatika dan Komputerisasi Akuntansi*, vol. 5, no. 2, pp. 147–154, 2021, doi: 10.46880/jmika.vol5no2.pp147-154.

[18] A. A. Lutfi, A. E. Permanasari, and S. Fauziati, "Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine," *Journal of Information Systems Engineering and Business Intelligence*, vol. 4, no. 2, p. 169, 2018, doi: 10.20473/jisebi.4.2.169.

[19] A. F. Rochim, K. Widyaningrum, and D. Eridani, "Comparison of Linear , Radial Base Function , and Polynomial Kernel Function Support Vector Machine Method Towards COVID-19 Sentiment Analysis".

[20] A. Nurkholis, D. Alita, and A. Munandar, "Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 2, pp. 227–233, 2022, doi: 10.29207/resti.v6i2.3906.

[21] B. P. Zen, I. Susanto, and D. Finaliamartha, "TF-IDF Method and Vector Space Model Regarding the Covid-19 Vaccine on Online News," *SinkrOn*, vol. 6, no. 1, pp. 69–79, 2021, doi: 10.33395/sinkron.v6i1.11179.

[22] P. H. Prastyo, I. Ardiyanto, and R. Hidayat, "Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF," *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020*, 2020, doi: 10.1109/ICDABI51230.2020.9325685.

[23] M. E. Purbaya, D. P. Rakhmadani, M. P. Arum, and L. Z. Nasifah, "Comparison of Kernel Support Vector Machines in Conducting Sentiment Analysis Review of Buying Chips on the Shopee E- Marketplace in Indonesian," in *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2022, pp. 435–440. doi: 10.1109/ICIMCIS56303.2022.10017546.

[24] N. Nasser, L. Karim, A. el Ouadrhiri, A. Ali, and N. Khan, "n-Gram based language processing using Twitter dataset to identify COVID-19 patients," *Sustain Cities Soc*, vol. 72, Sep. 2021, doi: 10.1016/j.scs.2021.103048.

[25] M. Mether, "The history of the central limit theorem," *Sovelletun Matematiikan erikoistyöt*, vol. 2, no. 1, p. 08, 2003, doi: 10.1007/978-0-387-87857-7.

[26] A. S. Widagdo, B. S. W.A, and A. Nasiri, "Analisis Tingkat Kepopuleran E-Commerce Di Indonesia Berdasarkan Sentimen Sosial Media Menggunakan Metode Naïve Bayes," *Jurnal Informa : Jurnal Penelitian dan Pengabdian Masyarakat*, vol. 6, no. 1, pp. 1–5, 2020, doi: 10.46808/informa.v6i1.159.

[27] S. I. Nurhafida and F. Sembiring, "Analisis Text Clustering Masyarakat di Twiter Mengenai Mcdonald'sxbts Menggunakan Orange Data Mining," *Seminar Nasional Sistem Informasi dan Manajemen Informatika*, vol. 1, no. 1, pp. 28–35, 2021.

[28] Imamah and F. H. Rachman, "Twitter sentiment analysis of Covid-19 using term weighting TF-IDF and logistic regresion," *Proceeding - 6th Information Technology International Seminar, ITIS 2020*, pp. 238–242, 2020, doi: 10.1109/ITIS50118.2020.9320958.

[29] H. E. Wynne and Z. Z. Wint, "Content based fake news detection using N-gram models," *ACM International Conference Proceeding Series*, 2019, doi: 10.1145/3366030.3366116.

[30] L. Mutawalli, M. T. A. Zaen, and W. Bagye, "Klasifikasi Teks Sosial Media Twitter Menggunakan Support Vector Machine (Studi Kasus Penusukan Wiranto)," *Jurnal Informatika dan Rekayasa Elektronik*, vol. 2, no. 2, p. 43, 2019, doi: 10.36595/jire.v2i2.117.

[31] A. S. Nugroho, A. Budi Witarto, and D. Handoko, "Support Vector Machine," *IlmuKomputer.com*. p. 11, 2003. doi: 10.1109/CCDC.2011.5968300.

[32] D. Berrar, "Cross-Validation," in *Encyclopedia of Bioinformatics and Computational Biology - Volume 1*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Elsevier, 2019, pp. 542–545. doi: 10.1016/b978-0-12-809633-8.20349-x.

[33] M. Awad and R. Khanna, *Efficient Learning Machines*. Berkeley, CA: Apress, 2015. doi: 10.1007/978-1-4302-5990-9.

[34] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," *Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017*, vol. 2018-Janua, no. December, pp. 391–394, 2018, doi: 10.1109/IALP.2017.8300625.

DOI: https://doi.org/10.29207/resti.v7i3.4726
Creative Commons Attribution 4.0 International License (CC BY 4.0)