Journal of
Information Systems Engineering
and Business Intelligence

Vol.9, No.1, April 2023
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Vader Lexicon and Support Vector Machine Algorithm to Detect Customer Sentiment Orientation

Vivine Nurcahyawati [1)*], **Zuriani Mustaffa** [2)]

[1)] *Universitas Dinamika, Indonesia*
Jl. Raya Kedung Baruk 98, Surabaya
[1)]vivine@dinamika.ac.id

[2)] *Universiti of Malaysia Pahang, Malaysia*
Lebuh Persiaran Tun Khalil Yaakob 26300, Pahang
[2)]zuriani@ump.edu.my

Abstract

Background: The concept of customer orientation, which is based on a set of fundamental beliefs that prioritize the interests of the customer, requires companies to detect these interests in order to maintain a high level of quality in their products or services. Furthermore, there are several indicators of customer orientation, and one of them is their opinion or taste, which provides valuable feedback for businesses. With the rapid development of social media, customers can express emotions, thoughts, and opinions about services or products that may not be easily conveyed in the real world.
Objective: The objective of this study is to detect customer orientation towards product or service quality, as expressed in online or social media. Additionally, the study showcases the novelty and superiority of the annotation process used for detecting customer orientation classifications.
Methods: This study employs a method to compare the classification performance of the Vader lexicon annotation process with manual annotation. To accomplish this, a dataset from the Amazon website will be analyzed and classified using the Support Vector Machine algorithm. The objective of this method is to determine the level of customer orientation present within the dataset. To evaluate the effectiveness of the Vader lexicon, the study will compare the results of manual and automatic data annotation.
Results: The results showed that customer orientation towards product or service quality has a predominantly positive value, comprising up to 76% of the total responses analyzed.
Conclusion: The findings demonstrate that using Vader in the annotation process results in superior accuracy values compared to manual annotation. Specifically, the accuracy value increased from 86% to 88.57%, indicating that Vader could be a reliable tool for annotating text. Therefore, future studies should consider using Vader as a classifier or integrating it into the annotation process to further enhance its performance.

Keywords: Classification, Customer, Orientation, Text analysis, Vader lexicon,

Article history: Received 11 February 2023, First decision 22 March 2023, Accepted 20 April 2023, Available online 28 April 2023

I. INTRODUCTION

The concept of customer orientation is a critical psychological asset that can provide significant benefits to salespeople in their professional endeavors. It comprises a wide range of work ethics, attitudes, tendencies, and mindsets that promote engagement and clarity of roles in customer service. By prioritizing customer satisfaction and needs, salespeople can use customer orientation to ensure positive customer experiences during service interactions. The guiding principle of this orientation is a "concern for the customer", which shifts salespeople's focus from self-interests to customer needs and satisfaction. This customer orientation facilitates meaningful, warm, and empathetic relationships with customers [1]. To foster this orientation, there are several approaches that companies can take, including enabling customers to contact salespeople, hiring friendly and solution-focused personnel, encouraging staff autonomy, valuing employees as assets, training team members, leading by example, capturing the voice of the customer through surveys, customer focus groups, data analysis, and creating a customer value proposition, as well as

[*] Corresponding Author

ISSN 2443-2555 (online) 2598-6333 (print) © 2023 The Authors. Published by Universitas Airlangga.

doi: http://dx.doi.org/10.20473/jisebi.9.1.109-119

optimizing processes through follow-ups [2]. Therefore, this study aims to explore customer orientation through a specific approach, which involves analyzing customer opinions and reviews shared on social media platforms.

Customer orientation plays a crucial role in driving high performance for companies, and its impact surpasses that of technology orientation [3]. For instance, in the restaurant industry, online reviews from previous customers can be analyzed to identify customer segments, predict their preferences, and prioritize service quality improvements [4]. This study is conducted across various domains using a machine learning approach that includes text preprocessing, text extraction, topic modeling, regression analysis, and artificial neural networks to analyze 14 variables for customer value [5]. Additionally, it is important to note that customer preferences for hotel features can vary among different customer types [6]. A study has proposed using aspect-level sentiment analysis and the Kano model-based extraction rule to determine customer orientation towards product attributes [7]. To enhance the accuracy of sentiment element recognition and value analysis, it is essential to use semi-supervised learning techniques to develop a better annotation opinion and offer a sufficient training corpus [8].

In today's digital age, the internet and social media have become essential tools for businesses to assess customer orientation towards their services or products. Gathering customer feedback is crucial for companies to improve their offerings [9]. For instance, social media platforms such as Twitter can be used to collect customer responses and feedback, providing valuable insights into how customers perceive a product or service [9], [10]. Moreover, reviews on social media can also be analyzed to determine public opinion towards the latest films and other forms of entertainment products [11], [12]. An entrepreneur in a restaurant can also use customer reviews to improve their services [13]. To conduct a thorough customer orientation analysis, it is crucial to adopt a suitable study model such as Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and others.

Several studies have shown the efficacy of Support Vector Machine (SVM) in customer orientation analysis, outperforming other algorithms. Specifically, when combined with Fuzzy Matching (FM), this algorithm has consistently produced superior results across multiple datasets [14]. Studies have found that SVM with RBF kernel outperforms the Sigmoid kernel in terms of accuracy [15]. Moreover, comparative analyses of SVM and KNN algorithms have shown that SVM is slightly superior in accuracy, with an average accuracy value of 69.27% for SVM and 61.3% for KNN [16].

To improve the accuracy of sentiment analysis, it is essential to use semi-supervised learning techniques to create a better annotation corpus for opinion analysis and provide an adequate training dataset. This approach can lead to more effective sentiment value analysis and more precise identification of elements [17], [18]. In lexicon-based systems, the dictionary is a critical component, but updating it manually poses a significant challenge. This issue can be overcome by using a lexicon-based machine-learning approach with automated dictionary updates [19]. As a result, this study employed the Vader lexicon as the initial dataset annotation.

SVM and the Vader lexicon have been used in various domains to predict consumer response and product or service experiences [20]–[24]. Twitter has also been used as a platform for analyzing responses to climate change [25]. The Vader lexicon has proven to be an effective annotation tool, demonstrating high accuracy rates. In a study on predicting customer response from chatbot sources, Vader achieved an accuracy rate of 93.33% [24]. When combined with SVM, the accuracy of Vader increased to 80.3%, surpassing Naive Bayes, Random Forest, Neural Network, Decision Tree, and k-Nearest Neighbour [26]. Furthermore, Vader outperformed other lexicons, such as Textblob, Flair (LSTM), and Flair (BERT), achieving an accuracy value of 72.59% in unsupervised group comparison [21], [25].

This study aims to enhance the process of opinion annotation by merging two different techniques: Support Vector Machine (SVM) and Vader lexicon. The combined approach aims to leverage the strengths of each technique to achieve a higher level of accuracy. Furthermore, the integration is carried out in a sequential manner where the Vader lexicon is used to establish opinion values, and the resulting data is employed as annotation data in SVM. Clustering is also applied to the data based on their attributes to gain further insights into customer orientation [22]. Essentially, this fusion transforms the lexicon-based approach into a mechanism for transferring learning to SVM. The expectation is that the amalgamation of techniques and algorithms will lead to highly accurate opinion annotation.

This study aims to develop a hybrid method that can enhance the accuracy of classification predictions. The method involves the use of classification to identify customer orientation towards a product. This study has the following objectives (RO):

RO1. Determine the overall customer perception of a product or service in the utilized data source.
RO2. Understanding the consumer orientation base on top positive and negative sentences found from samples evaluated by the Vader lexicon and the SVM approach.
RO3. List words that are often conveyed by consumers to express their opinions.
RO4. Clustering consumers based on gender and city for each perception.

II. Methods

Customer orientation has a significant impact on marketing performance in various businesses. This shows companies need to focus not only on marketing performance but also on understanding customer orientation toward their products. To achieve this, a hybrid approach has been developed. Fig. 1 shows the framework proposed in this study [27]. The explanation for each stage is provided below.

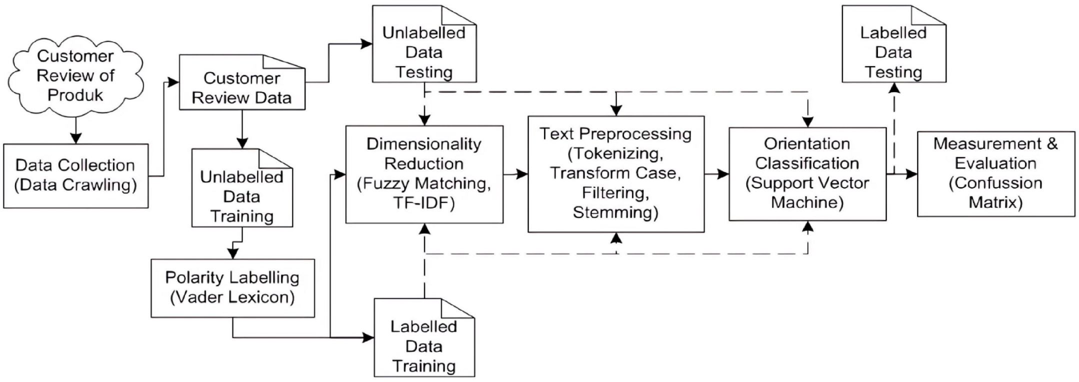
Customer Review of Produk
Data Collection (Data Crawling)
Customer Review Data
Unlabelled Data Testing
Unlabelled Data Training
Polarity Labelling (Vader Lexicon)
Labelled Data Training
Dimensionality Reduction (Fuzzy Matching, TF-IDF)
Text Preprocessing (Tokenizing, Transform Case, Filtering, Stemming)
Orientation Classification (Support Vector Machine)
Labelled Data Testing
Measurement & Evaluation (Confussion Matrix)

Fig. 1 Proposed Hybrid Approach

A. Data Collection

This study focuses on analyzing customer reviews of a specific product, which is sourced from the Amazon dataset. The dataset consists of two types of reviews: positive and negative. To facilitate analysis, the data is stored in a Microsoft Excel file. The raw customer review data contains various characters, symbols, numbers, URLs, mentions, and other information that require preprocessing to make it more uniform and reduce noise. Subsequently, the dataset is divided into 80% training data and 20% test data. Although the acquired dataset has positive and negative annotations, there are instances where the annotations are inappropriate, as shown in Table 1. For instance, the first data is annotated as positive, even though the sentence has a more negative connotation. The same is true for the second data point. These instances have the potential to adversely affect the accuracy of the analysis.

TABLE 1
INAPPROPRIATE DATASET ANNOTATION

Product Reviews	Polarity (Raw Data)	Polarity (should be)
!!!!!!!!!!!!!: I never got this product. I recieved an e-mail that mentioned, the product was returned to the whare-house, and my money was re-funded	Positive	Negative
!!!: haven't watched the movie yet as i have not read the book but am definitely looking forward to viewing and glad it completed my series	Negative	Positive

B. Polarity Annotation with Vader Lexicon

In the realm of analyzing opinion data, lexicon analysis is a simpler approach compared to machine learning, which requires more complex algorithms and greater computing power [28]. Previous studies have demonstrated the efficacy of the Vader lexicon, a sentiment analysis tool capable of accurately annotating opinion data [27]. This tool employs a lexicon dictionary comprising 1773 opinion words that are assigned specific weights. The polarity value of each data point is calculated based on the weight or value of each word in a sentence. A sentence is considered positive if the polarity value is greater than 0, negative if it is less than 0, and neutral if it is neither positive nor negative [15]. In this study, the Vader lexicon was used as a dataset annotation process, which can expedite the annotation process and improve the accuracy of the results.

C. Dimensionality Reduction

To simplify features in text data, two common algorithms used are FM and TF-IDF. The FM algorithm is a rule-based approach that enables the detection of predefined keywords and expressions in the text. This technique is capable of identifying expressions that occur within the text, allowing for some errors in the match. Additionally, it can detect matches in the text, even if the words are misspelled or have different suffixes or prefixes. The distance metric in FM is commonly employed to quantify the dissimilarity between two strings [29], [30]. In this FM, a specific approach that involves calculating the ratio of similarity distance values is used through the Levenshtein distance similarity technique [31].

Word weighting is a technique used to index words by assigning weights to each word, based on its significance in the document. This study utilized a word weighting approach that considered three factors including term frequency, inverse document frequency, and document length. To perform the word weighting, the study computed the values for term frequency (TF), document frequency (DF), inverse document frequency (IDF), and TF-IDF values [32], [33].

D.Text Preprocessing Series

Preprocessing is a crucial step in cleaning review data before processing and analysis to ensure accuracy and compliance with requirements. For instance, data in the form of Microsoft Excel files should be recalled for preprocessing to make it more structured and compatible. This typically involves several procedures such as transforming cases, stemming, tokenizing, and others.

E. Orientation Classification

The customer orientation classification stage involves using the SVM algorithm to distribute the training data, while testing is performed using K-Fold Cross Validation with various values of k. To evaluate the performance of the model, the study employed Cross Validation, a statistical method that splits the data into two subsets, including test and training data. K–Fold Cross Validation is a specific type of cross-validation, where the data is divided into k equal-sized parts or folds. During each fold, one part is used as test data while the remaining k – 1 parts are used as training data. This process is repeated k times, ensuring that every part is used as both test and training data [34].

SVM is a robust machine learning system that uses a high-dimensional feature space represented as a linear function to train on parameters and optimize learning through statistical bias. The system works to maximize the distance between classes to find the optimal hyperplane, that can effectively separate two classes for classification in a higher-dimensional space. To achieve this, SVM uses kernel tricks that transform the data into a higher-dimensional space to separate data linearly. Several kernel functions are commonly used, including Linear, Polynomial, and Radial Basis Functions (RBF) [15].

After following dimensionality reduction and text preprocessing, the test and training data are used in the classification process through the SVM algorithm. The SVM algorithm is used to predict the class of the test data, whether it is positive, negative, or neutral.

F. Measurement and evaluation

During the evaluation stage, it is crucial to conduct system testing to assess the performance of the classification results. This is typically achieved by calculating various metrics such as accuracy, precision, recall, and f-measure values. Performance evaluation is an important parameter used to gauge the accuracy and effectiveness of a method. In this evaluation process, a confusion matrix is utilized during the system development process [27]. The accuracy value is used to determine the proportion of correctly predicted outcomes in real-world scenarios. Precision indicates the degree of accuracy of the test results, while recall serves as a gauge for the percentage of outcomes where the correct value was found. The precision and recall values are combined to calculate the F1-Score, which represents the overall performance of the model. The equations (1), (2), (3), and (4) shown below represent the formulas used to calculate the aforementioned values:

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

$$\text{F1} - \text{Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \tag{4}$$

information:
TP (True Positive) = indicates a successful detection of a positive outcome
TN (True Negative) = represents a positively determined bad outcome
FP (False Positive) = displays negative results for positive detection
FN (False Negative) = positive outcomes are shown to be negative

III. RESULTS

A total of 300,000 data will be collected as training data, which is manually annotated as either negative or positive reviews based on Table 3. The annotation process involves linguists who determine the sentiment orientation of each sentence, with 1 indicating a negative review and 2 for a positive one. It is essential to note that this process is both time-consuming and costly [14]. Some data points (examples in Table 1) may not be suitable for manual annotation, as mentioned in Section 2. To address these issues, the lexicon approach has been proposed to facilitate a quicker and more accurate annotation process. Table 2 shows an example of annotation using this approach.

TABLE 2
RAW DATA OF REVIEW PRODUCT AND ANNOTATION MANUALY FOR TRAINING DATA

Product Reviews	Annotation
!!!: Simply Horrible.If you need an explnation see the movie.If you don't agree, you deserve this tripe of a film.	1
": It was okay. i dont like old movies though. i just wanted to see it because i liked the new version. The action was good.	1
" BUY THIS CD ": THIS IS " REALLY " THE BEST OF RANDY'S SONGS ... YOU WILL BE PLEASED !!!!!! ... I HAD BOUGHT OTHER CD'S .... I LIKE THIS ONE	2
"A Beautiful Love Story": A lovely romantic story. I was delighted to find this movie on DVD format available now here at Amazon.A romantic movie to watch again and again.	2

Section 2 shows that the use of automatic annotation using the Vader lexicon can improve classification performance. In comparison to manual annotation, the dataset will be annotated automatically with Vader before entering the preprocessing stage. The Vader lexicon consists of over 7500 lexical features with validated valence scores indicating sensory polarity (positive/negative) and feeling intensity on a scale of -4 (negative) to +4 (positive), with 0 representing neutrality. For instance, the words 'okay', 'for', 'bad', and 'sick' has a score of 0.9, 3.1, -2.5, and -1.5 respectively. Vader evaluates any given text and generates a positive, negative, or neutral score for each lexical feature. These scores are then added together to form a compound score, which is a matrix normalizing all scores from -1 to +1. A composite score greater than 0.05 is considered positive, a score less than -0.05 is considered negative, and a composite score between -0.05 and 0.05 is considered neutral [35]. The sentence polarity is determined based on the compound score of each existing word. Table 3 provides an example of how the Vader lexicon is scored.

TABLE 3
SCORING WITH VADER LEXICON

Review	Score	Annotation
!: Apostrophe is a great second or first …	4,231	Positive
' A CLASSIC DANZIG CD " ! !: MANY PEOPLE …	- 0,359	Negative
- a googleplex rating: I dont even …	1,667	Positive
I am sorry to burst the bubble but to see …	0,923	Positive
- for purists only: The entire recording…	1,308	Positive

The success of this search concept hinges on the ability to determine whether a searched string is similar to a string contained in the dictionary, even if the character arrangement is not an exact match. This determination of "similarity" is achieved using a function known as the Levenshtein Distance similarity. To identify similar words, each word is compared with every other word. Unique similar words are stored to prevent duplicate words, as shown in Fig. 2.

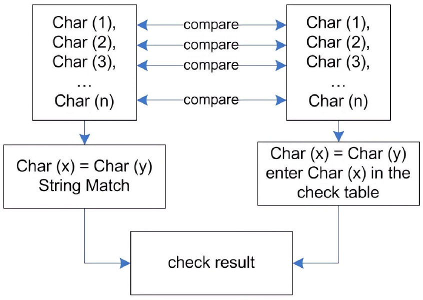

Fig. 2 Distance Measurement for Fuzzy Matching

The subsequent stage involves assigning a weight to each vocabulary in the document using the term frequency-inverse document frequency calculations (TF-IDF). Table 4 shows the results of calculating the term frequency and TF-IDF.

TABLE 4
WORD VECTORS WITH TF-IDF

No	Annotation	a	able	….	yours	zappa	zero
1	Positive	0.0097	0.0	0.0	0.0	0.3448	0.0
2	Positive	0.0295	0.0	0.0	0.0	0.0	0.0
3	Negative	0.0120	0.0	0.0	0.1063	0.0	0.0
4	Negative	0.0152	0.0896	0.0	0.0	0.0	0.0
5	Negative	0.0139	0.0	0.0	0.0	0.0	0.0

Following the annotation of the data, the next step is preprocessing, which involves several essential techniques. Table 5 shows that the initial text is in the form of raw data consisting of customer reviews of a product. Data cleaning is the first technique employed, which involves the removal of noise and all non-letter characters. The case of all letters is transformed to lowercase to ensure consistency. Subsequently, stop word removal is implemented to eliminate unimportant words from the document based on a predefined stop word list. The next step is stemming, which enables the identification of the root form of a word and replaces it with the correct grammatical structure in English. Finally, the tokenization process is carried out by breaking the documents into token pieces, achieved by separating sentences based on space characters.

TABLE 5
PREPROCESSING RESULT

Preprocessing Stage		Sample Result
Initial text	:	": It was okay. i dont like old movies though. i just wanted to see it because i liked the new version. The action was good.
Cleaning	:	It was okay i dont like old movies though i just wanted to see it because i liked the new version The action was good
Tranform Case	:	it was okay i dont like old movies though i just wanted to see it because i liked the new version the action was good
Stopword Removal	:	okay dont like old movies though just wanted see because liked new version action good
Stemming	:	okay dont like old movies though just want see because like new version action good
Tokenizing	:	okay, dont, like, old, movies, though, just, want, see, because, like, new, version, action, good

A. RO1: Determine the overall customer perception of a product or service in the utilized data source.

Following the TF-IDF calculation, the subsequent step involves developing a classification model using a Support Vector Machine (SVM) algorithm. In this phase, the classification process is conducted using SVM. The distribution of the test and training data is performed by using Cross Validation with a fold value of 10. Based on the results of the classification using SVM, the composition is shown in Fig. 3. The findings showed that the majority of customer orientation is positive, with a percentage of 76%, while negative and neutral feedback constitute 22% and 2% respectively.

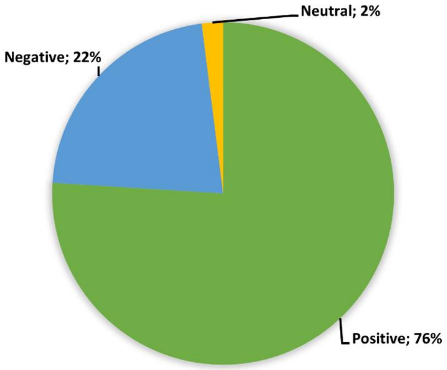

Fig. 3 Customer Orientation Classification

B. RO2: Understanding the consumer orientation based on top positive and negative sentences found from samples evaluated by the Vader lexicon and the SVM approach.

Table 6 shows consumer reviews with the highest scores based on positive and negative annotation. The review is characterized as positive if the word scores are above 0, while they are classified as negative if the word scores are below 0. These outcomes provide valuable insights for service managers or product owners to evaluate their business processes.

TABLE 6
TOP POSITIVE AND NEGATIVE SENTENCES

Review	Score	Annotation	Scoring String
!!Thoroughly enjoyable!!: This book starred one of the most delicious heroes I've seen in ages! Lucas is heroic (in the very best sense of the word), sexy through and through, and downright easy to fall for. Watching two best friends discover their love for one another was a joy to read. Heartwarming, funny, charming, emotional, richly rewarding, and beautifully written. A keeper!	10,949	Positive	enjoyable (0.49) delicious (0.69) heroes (0.59) heroic (0.67) best (0.82) sexy (0.62) easy (0.49) best (0.82) friends (0.54) love (0.82) joy (0.72) heartwarming (0.54) funny (0.49) charming (0.72) emotional (0.15) richly (0.49) rewarding (0.62) beautifully (0.69)
!!!!!expired perfume!!!!!: I bought this 2 years or so ago, and I am STILL upset about it. it arrived and was a dark orange-ish color, and smelled rank and well just old. I would never ever buy from here again. I waited 15 years to have this fragrance because I could never afford it, only to be let down terribly. maybe they didnt know it was expired and had went bad..i dunno....in any case, bad bad experience.	-2,718	Negative	upset (-0.41) well (0.28) terribly (-0.67) bad (-0.64) bad (-0.64) bad (-0.64)

C. RO3: List words that are often conveyed by consumers to express their opinions.

Fig. 4 illustrates common words that customers frequently use in their feedback about products. The word that occurs most frequently in the dataset is displayed with a larger font size to emphasize its significance. For instance, the word "Great" is the most commonly used term by customers to express positive feedback, followed by "Good", "Album", and others. Words such as "Sound" and "Give" are frequently used in negative feedback annotations.

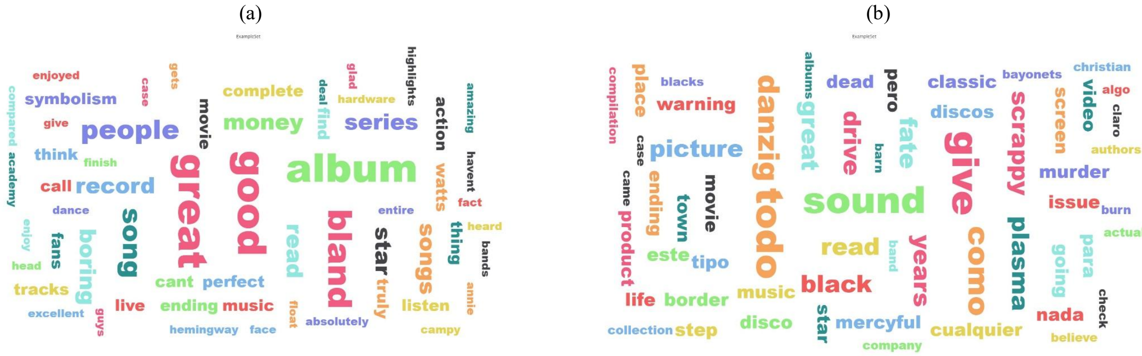

Fig. 4 The Word Most Commonly Used in Positive (a) and Negative (b) Perception

D. RO4: Clustering consumers based on gender and city for each perception.

The analysis also includes a categorization of consumers by gender and city. As showed in Fig. 5, a greater proportion of female consumers gave positive reviews compared to their male counterparts. The dataset includes reviews from different cities: California, Colorado, Kansas, Kentucky, and New York. Interestingly, consumers from New York contributed the highest number of positive reviews.

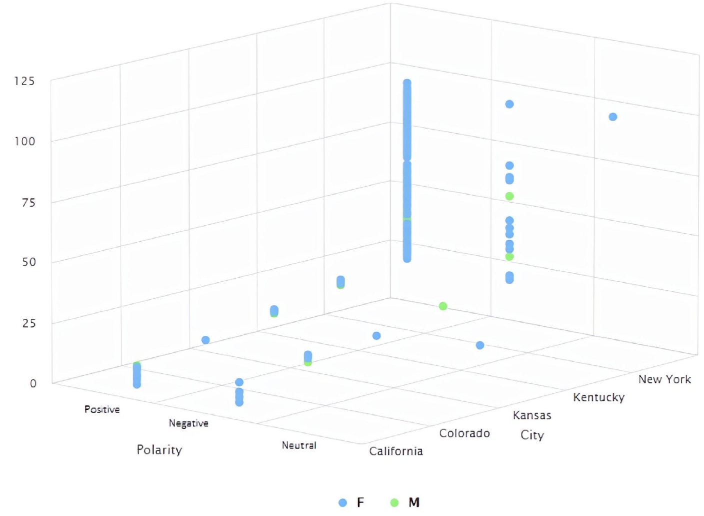

Fig. 5 Clustering based on Gender and City for Each Perception

The study involved a comparison between classifications obtained through manual and automatic annotations using the Vader lexicon. Results presented in Table 7 and Fig. 6 show that Vader's annotation outperformed the manual aspect across several key metrics, including accuracy, precision, recall, and f-score.

TABLE 7
RESULTS OF TESTING THE USE OF VADER LEXICON FOR ANNOTATION

Evaluation	Manual Annotation (%)	Vadel Lexicon Annotation (%)
Accuracy	86	88,57
Precision	86,08	89,71
Recall	95,43	96,32
F-Score	90,51	92,89

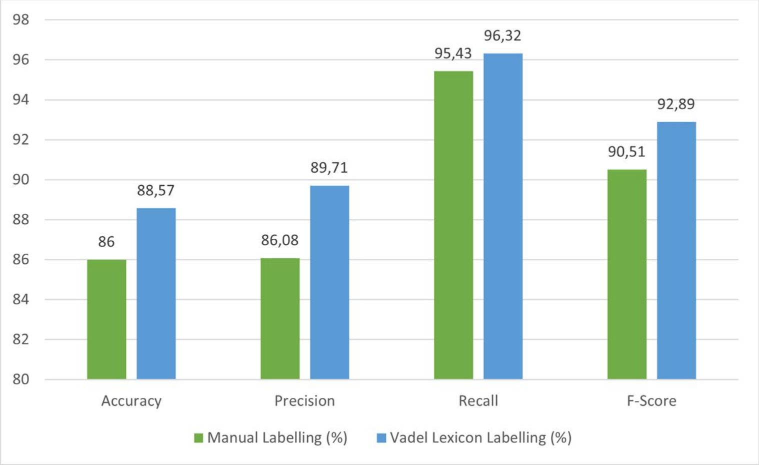

Fig. 6 Comparison Results Using the Vader Lexicon for Annotation

Throughout the evaluation phase, several tests were conducted to assess the effectiveness of various techniques, including text preprocessing, factorization machines (FM), and term frequency-inverse document frequency (TF-IDF). To assess the performance of these techniques, SVM classification was implemented using both manual and individual data annotations with the Vader lexicon.

IV. DISCUSSION

This study focused on two primary objectives: detecting customer orientation towards a product or service and evaluating the performance of lexicon-based opinion analysis. However, due to the non-standardized nature of customer opinions expressed on social media or online platforms, the data required initial processing to enhance its structure. Several preprocessing techniques were employed, including data cleaning, case transformation, stop word removal, stemming, and tokenization. Afterward, dimensionality reduction was performed using FM and TF-IDF techniques. The classification stage employed the SVM algorithm due to its demonstrated high performance in previous studies [14]. Analysis through these stages showed that customer orientation tended to be positive with a value of 76%.

In the second analysis, an automatic annotation process was conducted using the Vader lexicon. The results of this analysis showed that the performance of the lexicon was superior to that of the manual annotation process. Specifically, the accuracy value increased from 86% to 88.57%, representing a 2.57% improvement. This level of performance surpassed that of previous studies, which reported an increase from 70% to 88.57%.

The proposed model demonstrated superior performance compared to the benchmark aspect [36], which achieved a maximum accuracy of 70%. In contrast, the experiment achieved an accuracy value of 88.57%. Fig. 7 shows that the proposed model outperforms the benchmark aspect in terms of accuracy. It is important to note that the data used in this study is substantially larger than that of the benchmark. In the benchmark study, annotations were determined based on the conversion of rating values. Numbers 4 & 5, 1 & 2, and 3 were considered positive, negative, and neutral annotations respectively. This approach had a weakness, as the rating value did not always reflect the actual sentiment expressed in the review. For instance, a review with a high rating may still contain negative feedback. This study employed the Vader lexicon for annotation, which assigns weights to individual words and determines the overall sentiment of the review. The assigned weights provide a score that determines whether the sentiment is positive, negative, or neutral. This method has been shown to increase accuracy.

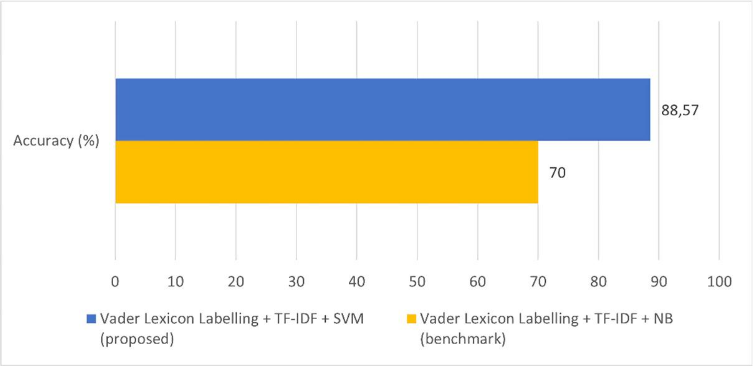

Fig. 7 Performance Comparison with Benchmark Study

The study has shown limitations in the effectiveness of the Vader lexicon for annotation due to the incompleteness of word dictionaries and applied word weights. Although the word underwent normalization, not all normalized words are included in the lexicon. This makes some words that should have weighted values to produce an opinion not to be detected, resulting in no weighted value. For instance, the sentence "not good" conveys a negative sentiment, but the word "not" does not weigh the lexicon, while "good" has a high weight representing a positive opinion. In this case, the phrase "not good" is considered a positive opinion, despite its negative connotation. To achieve maximum results in the annotation stage, it is necessary to expand the lexicon dictionary by maximizing words starting with "not" to have a weight that tends to represent negative opinions.

V. CONCLUSIONS

The proposed study introduces a novel approach to customer orientation classification by combining two powerful techniques: Support Vector Machine (SVM) and Vader lexicons. The SVM algorithm's performance is enhanced by implementing dimensionality reduction techniques such as the FM and TF-IDF algorithms. The primary objective of this study is to determine customers' perceptions of products and services, which constitutes a significant contribution to the field. Data analysis results reveal frequently used positive and negative words, providing valuable insights into customer behavior. Moreover, this study contributes practical techniques for data extraction, including tokenization, stop word deletion, case transformation, and stemming. After experimentation, the Vader lexicon-SVM emerges as the most promising approach, achieving an impressive accuracy rate of 88.57%. Therefore, it can be concluded that this approach can effectively produce the best opinion analysis performance.

Author Contributions: *Nurchyawati*: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Visualization, Writing – Original Draft. *Mustaffa*: Supervision, Methodology, Validation, Writing - Review & Editing.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] H. Park and W.-M. Hur, "Customer Showrooming Behavior, Customer Orientation, and Emotional Labor: Sales Control as a Moderator," *Journal of Retailing and Consumer Services*, vol. 72, pp. 1–10, 2023, doi: https://doi.org/10.1016/j.jretconser.2023.103268.

[2] M. R. Jalilvand, "The Effect of Innovativeness and Customer-Oriented Systems on Performance in The Hotel Industry of Iran," *Journal of Science and Technology Policy Management*, vol. 8, no. 1, pp. 43–61, 2017, doi: 10.1108/JSTPM-08-2016-0018.

[3] R. T. Frambach, P. C. Fiss, and P. T. M. Ingenbleek, "How Important is Customer Orientation for Firm Performance? A Fuzzy Set Analysis of Orientations, Strategies, and Environments," *J Bus Res*, vol. 69, no. 4, pp. 1428–1436, Apr. 2016, doi: 10.1016/j.jbusres.2015.10.120.

[4] M. Nilashi *et al.*, "Big Social Data and Customer Decision Making in Vegetarian Restaurants: A Combined Machine Learning Method," *Journal of Retailing and Consumer Services*, vol. 62, Sep. 2021, doi: 10.1016/j.jretconser.2021.102630.

[5] W. Kwon, M. Lee, and K. J. Back, "Exploring the Underlying Factors of Customer Value in Restaurants: A Machine Learning Approach," *Int J Hosp Manag*, vol. 91, Oct. 2020, doi: 10.1016/j.ijhm.2020.102643.

[6] Y. Bian, R. Ye, J. Zhang, and X. Yan, "Customer Preference Identification from Hotel Online Reviews: A Neural Network Based Fine-Grained Sentiment Analysis," *Comput Ind Eng*, vol. 172, Oct. 2022, doi: 10.1016/j.cie.2022.108648.

[7] J. Zhang, A. Zhang, D. Liu, and Y. Bian, "Customer Preferences Extraction for Air Purifiers based on Fine-Grained Sentiment Analysis of Online Reviews," *Knowl Based Syst*, vol. 228, p. 107259, Sep. 2021, doi: 10.1016/j.knosys.2021.107259.

[8] J. Zhang, X. Lu, and D. Liu, "Deriving Customer Preferences for Hotels based on Aspect-Level Sentiment Analysis of Online Reviews," *Electron Commer Res Appl*, vol. 49, Sep. 2021, doi: 10.1016/j.elerap.2021.101094.

[9] P. Ray and A. Chakrabarti, "Twitter Sentiment Analysis for Product Review Using Lexicon Method," in *International Conference on Data Management, Analytics and Innovation*, 2017, pp. 211–216. doi: 10.1109/ICDMAI.2017.8073512.

[10] J. Mahilraj, G. Tigistu, and S. Tumsa, "Text Preprocessing Method on Twitter Sentiment Analysis using Machine Learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 12, pp. 233–240, Sep. 2020, doi: 10.35940/ijitee.K7771.0991120.

[11] D. H. Abd, A. R. Abbas, and A. T. Sadiq, "Analyzing Sentiment System to Specify Polarity by Lexicon-Based," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 283–289, Feb. 2021, doi: 10.11591/eei.v10i1.2471.

[12] A. Alsayat, "Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model," *Arab J Sci Eng*, vol. 47, no. 2, pp. 2499–2511, Feb. 2022, doi: 10.1007/s13369-021-06227-w.

[13] M. Işik and H. Dağ, "The Impact of Text Preprocessing on the Prediction of Review Ratings," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 3, pp. 1405–1421, 2020, doi: 10.3906/elk-1907-46.

[14] V. Nurcahyawati and Z. Mustaffa, "Improving Sentiment Reviews Classification Performance using Support Vector Machine-Fuzzy Matching Algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 3, pp. 1817–1824, Jun. 2023, doi: 10.11591/eei.v12i3.4830.

[15] R. H. Muhammadi, T. G. Laksana, and A. B. Arifa, "Combination of Support Vector Machine and Lexicon-Based Algorithm in Twitter Sentiment Analysis," *Jurnal Ilmu Komputer dan Informatika*, vol. 8, no. 1, pp. 59–71, 2022, doi: https://doi.org/10.23917/khif.v8i1.15213.

[16] F. Firmansyah *et al.*, "Comparing Sentiment Analysis of Indonesian Presidential Election 2019 with Support Vector Machine and K-Nearest Neighbor Algorithm," in *International Conference on Computing, Engineering, and Design*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ICCED51276.2020.9415767.

[17] S. Shevira, I. M. A. D. Suarjaya, and P. W. Buana, "Lexicon and Naive Bayes Algorithms to Detect Mental Health Situations from Twitter Data," *Journal of Information Systems Engineering and Business Intelligence*, vol. 8, no. 2, pp. 142–148, 2022, doi: 10.20473/jisebi.8.2.

[18] N. M. Sham and A. Mohamed, "Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches," *Sustainability (Switzerland)*, vol. 14, no. 8, Apr. 2022, doi: 10.3390/su14084723.

[19] W. Zhao *et al.*, "Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis," *IEEE Trans Knowl Data Eng*, vol. 30, no. 1, pp. 185–197, Jan. 2018, doi: 10.1109/TKDE.2017.2756658.

[20] A. Borg and M. Boldt, "Using VADER Sentiment and SVM for Predicting Customer Response Sentiment," *Expert Syst Appl*, vol. 162, Dec. 2020, doi: 10.1016/j.eswa.2020.113746.

[21] M. S. Hossain, M. F. Rahman, M. K. Uddin, and M. K. Hossain, "Customer Sentiment Analysis and Prediction of Halal Restaurants Using Machine Learning Approaches," *Journal of Islamic Marketing*, 2022, doi: 10.1108/JIMA-04-2021-0125.

[22] V. M. C. Sagarino, J. I. M. Montejo, and A. M. Ceniza-Canillo, "Sentiment Analysis of Product Reviews as Customer Recommendations in Shopee Philippines Using Hybrid Approach," in *International Conference on Information Technology and Digital Applications*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICITDA55840.2022.9971379.

[23] A. G. Budianto, B. Wirjodirdjo, I. Maflahah, and D. Kurnianingtyas, "Sentiment Analysis Model for KlikIndomaret Android App During Pandemic Using Vader and Transformers NLTK Library," in *IEEE International Conference on Industrial Engineering and Engineering Management*, IEEE Computer Society, 2022, pp. 423–427. doi: 10.1109/IEEM55944.2022.9989577.

[24] J. N. Mindoro, M. A. F. Malbog, M. D. S. Nipas, J. A. B. Susa, A. G. Acoba, and J. S. Gulmatico, "Sentiment Analysis in Customer Experience in Philippine Courier Delivery Services using VADER Algorithm Thru Chatbot Interviews," in *International Conference on Power, Control and Computing Technologies*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICPC2T53885.2022.9777007.

[25] D. Effrosynidis, A. I. Karasakalidis, G. Sylaios, and A. Arampatzis, "The Climate Change Twitter Dataset," *Expert Syst Appl*, vol. 204, Oct. 2022, doi: 10.1016/j.eswa.2022.117541.

[26] D. Marutho, Muljono, S. Rustad, and Purwanto, "Sentiment Analysis Optimization Using Vader Lexicon on Machine Learning Approach," in *International Seminar on Intelligent Technology and Its Applications: Advanced Innovations of Electrical Systems for Humanity*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 98–103. doi: 10.1109/ISITIA56226.2022.9855341.

[27] M. K. Bashar, "A Hybrid Approach to Explore Public Sentiments on COVID-19," *SN Comput Sci*, vol. 3, no. 3, pp. 1–19, May 2022, doi: 10.1007/s42979-022-01112-1.

[28] M. S. Hossen and N. R. Dev, "An Improved Lexicon Based Model for Efficient Sentiment Analysis on Movie Review Data," *Wirel Pers Commun*, vol. 120, no. 1, pp. 535–544, Sep. 2021, doi: 10.1007/s11277-021-08474-4.

[29] M. S. Oliveira, A. Mourthe, and M. C. Duque, "Extracting events from Daily Drilling Reports using Fuzzy String Matching," *The APPEA Journal*, vol. 62, no. 2, pp. S158–S161, May 2022, doi: 10.1071/aj21118.

[30] H. Kyung Yu and J. Gon Kim, "Indoor Positioning by Weighted Fuzzy Matching in Lifi Based Hospital Ward Environment," *J Phys Conf Ser*, vol. 1487, no. 1, Apr. 2020, doi: 10.1088/1742-6596/1487/1/012010.

[31] S. Abdulmalek, S. AL-Hagree, M. Alsurori, M. Hadwan, A. Aqlan, and F. Alqasemi, "Levenstein's Algorithm On English and Arabic: A Survey," in *International Conference of Technology, Science and Administration (ICTSA)*, Taiz, Yemen: IEEE, 2021. doi: 10.1109/ICTSA52017.2021.9406547.

[32] D. N. de Oliveira and L. H. de C. Merschmann, "Joint Evaluation of Preprocessing Tasks with Classifiers for Sentiment Analysis in Brazilian Portuguese Language," *Multimed Tools Appl*, 2021, doi: 10.1007/s11042-020-10323-8.

[33] E. Araslanov, E. Komotskiy, and E. Agbozo, "Assessing the Impact of Text Preprocessing in Sentiment Analysis of Short Social Network Messages in the Russian Language," in *International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ICDABI51230.2020.9325654.

[34] M. K. Delimayanti, R. Sari, M. Laya, M. R. Faisal, Pahrul, and R. F. Naryanto, "The Effect of Pre-Processing on the Classification of Twitter's Flood Disaster Messages using Support Vector Machine Algorithm," in *Proceedings of ICAE 2020 - 3rd International Conference on Applied Engineering*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ICAE50557.2020.9350387.

[35] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *International AAAI Conference on Weblogs and Social Media,* 2014, pp. 216–225. doi: https://doi.org/10.1609/icwsm.v8i1.14550.

[36] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, "Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam Menganalisis Sentimen Data Ulasan PLN Mobile," *PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika*, vol. 15, no. 2, pp. 264–275, Nov. 2022, doi: 10.33322/petir.v15i2.1733.

Publisher's Note: Publisher stays neutral with regard to jurisdictional claims in published maps and institutional affiliation