Exceeding Manual Labeling: VADER Lexicon as an Accurate Alternative to Automatic Sentiment Classification

Vivine Nurcahyawati
Department of Information System
Universitas Dinamika, Indonesia
vivine@dinamika.ac.id

Zuriani Mustaffa
Faculty of Computing, Universiti Malaysia
Pahang Al-Sultan Abdullah, Malaysia
zuriani@umpsa.edu.my

Mohammed Khalaf
Department of Computer Science
University of Al-Maarif, Iraq
m.i.khalaf@uoa.edu.iq

Abstract: *The number of internet users worldwide has increased dramatically, resulting in a surge of content uploaded over the Internet, particularly in text form. Global Internet users now exceed 5,16 billion, constituting a penetration rate of 64.4 percent of the world's total population. While only a small fraction of individuals actively expresses their opinions online, sentiment analysis aims to categorize textual information into favorable, negative, or neutral states of mind. When dealing with unlabeled datasets, the Valence Aware Dictionary and sEntiment Reasoner (VADER) Lexicon proves to be an effective tool for extracting feature sentiment. This facilitates the direct application of machine learning techniques such as Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbor (KNN) to classify datasets. Fuzzy Matching (FM) serves as a dimensionality reduction technique. Experimental results utilizing three datasets from diverse sources reveal that the combination of FM and SVM yields the highest accuracy. Model validation through K-Fold cross-validation reveals notable accuracy rates across multiple datasets. For dataset A, the accuracy stands at 94.69% with manual labeling and improves slightly to 95.92 % with VADER labeling. Similarly, for dataset B, the accuracy shows a marginal increase from 96.94% manual labeling to 97.01% VADER labeling. Dataset C also displays an enhancement in accuracy, with manual labeling achieving 95.51% accuracy and VADER labeling demonstrating a higher accuracy of 96.73%. These results underscore the effectiveness of both manual and automated labeling techniques in enhancing model performance across diverse datasets.*

Keywords: *Lexicon-based, classification, customer, review, text analysis.*

Received April 01, 2024; accepted December 04, 2024
https://doi.org/10.34028/iajit/22/2/2

1. Introduction

One aspect of text mining data management is sentiment analysis, a field that examines people's perceptions, feelings, assessments, actions, and attitudes toward various entities, including individuals, organizations, goods, services, issues, themes, events, and characteristics [14]. Fundamentally, sentiment analysis involves a classification task; however, the actual classification process is more complex than a straightforward categorization, given the uncertainty in word usage, the absence of tone in writing, and the evolving nature of language itself [14]. Sentiment analysis is responsible for determining the sentiment polarity of textual content, which involves determining whether the emotions expressed in a particular text are positive, negative, or neutral [32].

In contemporary contexts, sentiment analysis is frequently applied in diverse scenarios, such as forecasting election results [19], providing information about brands to groups [22], summarizing product reviews, or even predicting stock market trends [27, 52]. Sentiment analysis classification is also employed in the healthcare field, as seen in attempts to predict therapeutic techniques for Covid-19 [41]. Automated sentiment analysis proves highly advantageous in cases where outcomes rely on human perspectives.

Naive Bayes (NB) achieved 96% accuracy in analyzing sentiment regarding childfree choices [30], while Support Vector Machine (SVM) achieved 94% accuracy in e-commerce sentiment analysis [21]. Recent studies, such as those conducted on Gojek, use Aspect-Based Sentiment Analysis (ABSA) to analyze user reviews, achieving an accuracy of up to 96.67% [38].

Sentiment analysis is also used to analyze the influence of social media, for example, to observe trending topics. The childfree movement and applications like GBWhatsApp have sparked extensive discussions on platforms like Twitter, showing a mix of positive and negative sentiments [36, 44]. While sentiment analysis continues to evolve, challenges remain in accurately capturing nuanced opinions, especially in the rapidly changing digital landscape. Future research may focus on improving model accuracy and addressing the complexity of human sentiment.

This study contributes several approaches and algorithms to conduct sentiment analysis. For the classification of product reviews, SVM, NB, and K-Nearest Neighbor (KNN) algorithms are employed.

SVM excels in nonlinear classification [12] benefiting from vector support and hyperplane; however, it entails a time-consuming classification process [19]. Across various dataset types and preprocessing setups, SVM consistently produces optimal results, suggesting its preference as the sentiment analysis algorithm [12]. Shaban *et al*.'s research [46] achieved a 98% accuracy rate using NB, while Romadhon and Kurniawan [41] applied KNN, resulting in a 75% accuracy rate. A comparative analysis of these three algorithms is conducted.

Nevertheless, SVM, NB, and KNN face challenges in handling language variations [15]. Operating as binary classification algorithms, they are constrained by exact word matching, potentially leading to misclassification in texts with language variations [28]. Furthermore, they struggle to understand the context of words in sentences, causing misclassification in cases of word ambiguities. Additionally, they are unable to handle spelling errors Mahilraj *et al*. [28] a common occurrence in informal texts [7, 22].

To address these limitations, a Fuzzy Matching (FM) algorithm is introduced to match words with similar meanings, consider the context of words in sentences, and tolerate spelling errors. FM proves effective in simplifying features on big data [20], identifying similar but not identical text elements [35]. Using fuzzy logic, which is to search the level of truth where it searches the same string and strings that are close to the other strings collected although not exactly the same order of the characters [40]. FM offer a distinct advantage by enhancing accuracy even in scenarios with limited sample sizes [51]. Moreover, FM excels in feature extraction, mitigating overfitting concern [50], thereby reinforcing its utility in various analytical context.

Large datasets with accurate annotations are crucial for the success of supervised machine learning [9]. The labeling process itself poses challenges, susceptible to errors that introduce bias and hinder the generalization process of predictive models. Labeling errors may arise from automatic labeling, input-output ambiguities, or human errors (lack of expertise), leading to a decline in prediction performance. Consequently, label cleaning becomes essential for improving model training and evaluation.

Manual labeling in sentiment analysis presents several disadvantages that can significantly impact the quality and efficiency of the data used for training models. Manual annotation is a labor-intensive process, often requiring extensive time and financial resources. For instance, a study indicated that a dataset that could be annotated in approximately 173 seconds using an automated technique would take about 575 hours if done manually [37]. The high cost and time commitment can limit the volume of data that can be realistically labeled, leading to insufficient datasets for effective model training [47].

Annotator demographics can introduce biases, affecting the consistency and reliability of sentiment labels. Research shows that demographic differences among annotators can lead to significant variations in sentiment ratings, impacting model accuracy by over 4.5 points [13]. This inconsistency can skew the model's learning process, resulting in less generalizable and potentially biased outcomes.

While manual labeling is essential for creating high-quality datasets, its inherent limitations highlight the need for more efficient and unbiased methods, such as semi-automated or fully automated annotation techniques, to enhance sentiment analysis capabilities.

An alternative to manual labeling is Valence Aware Dictionary and sEntiment Reasoner (VADER), which significantly shortens labeling times through automatic processes. VADER Lexicon proves to be efficient for labeling [17]. VADER is a semantic and rule-based Lexicon utilized to calculate polarity scores and classify sentiments, overcoming the shortcomings of manual labeling [29].

In a study comparing VADER with the InSet Lexicon, VADER achieved an average accuracy of 82.65% in sentiment labeling, slightly lower than InSet's 85.8% [18]. Another research on ChatGPT reviews demonstrated that VADER, when combined with SVM, can yield high accuracy rates, reaching up to 92.72% [16]. VADER has been utilized to analyze public sentiment regarding tax policies, revealing both positive support and negative criticism [2]. In the context of COVID-19 vaccine discussions, VADER identified a predominantly positive public sentiment, although machine learning models outperformed it in classification accuracy [5].

The resulting score determines positive and negative polarization. When applied to consumer reviews, VADER achieved an accuracy of 70%, influenced by factors such as class imbalance, preprocessing techniques, labeling models, and classification models [6]. Similarly, it exhibited strong performance in predicting customer responses, with an average F1-score of 83.4% [10].

To enhance the functionality of sentiment analysis, the combination of algorithm classification and FM, along with robust annotation or labeling techniques, is crucial. VADER Lexicon emerges as an efficient approach to labeling product review datasets [17], addressing the drawbacks associated with manual labeling by employing a semantic and rule-based approach to generate polarity scores and identify moods [29].

2. Related Studies

A sentiment analysis model is a machine learning model used to determine the sentiment of a text. A high-quality sentiment analysis dataset is essential for producing an accurate sentimental analysis model. Research for analyzing user opinions is often obtained from online

media and social media, with Amazon product reviews [4], Twitter [19, 42, 43] or other social media platforms [3, 45] being common sources. The format of sentiment analysis datasets can vary depending on the data source. Generally, a sentiment analysis dataset requires at least two attributes: the review text and the label. The review text is a column containing the text to be analyzed, while the label column contains the sentiment label of the text, which can be positive, negative, or neutral. Additionally, a sentiment analysis dataset can be in JSON, XML, or other formats.

Previous research has indicated that sentiment analysis classification is widely used to address real-world problems. This study proposes the use of random sampling of minorities because it is a small-scale data set, but majority sampling can also be used, albeit requiring a large-scale data set that can reduce time complexity. The study also addresses the issue of limited data for multi-class classification problems based on class label analysis, which yields better predictions than the model while providing the same overall functionality for sentiment analysis related to COVID-19 [24]. Another study proposed an approach to conducting drug safety review analysis using Lexicon-based and in-depth learning techniques [26]. In addition, another paper focuses on analyzing various sentiment techniques in a dataset of tweet behaviors for various fields such as healthcare services, behavior estimation, etc. Furthermore, the results in this work explore and validate statistical machine learning classifiers that provide the percentage accuracy achieved in terms of positive, negative, and neutral tweets [11].

The diversity of dataset structures undoubtedly impacts the performance outcomes of sentiment analysis, necessitating efforts to select and enhance the quality of features within the dataset. In sentiment analysis, feature selection is a crucial technique that plays a role in improving the accuracy and efficiency of sentiment prediction models [50]. Term Frequency-Inverse Document Frequency TF-IDF possesses the capability to address common words that frequently appear throughout a document. It ensures that words which are more specific, unique, and carry greater weight associated with a particular sentiment exert a more significant influence [3, 4, 19]. However, TF-IDF has a weakness in handling spelling errors, and it can be susceptible to the influence of synonyms, treating similar words differently despite having the same meaning. When word order becomes a critical factor in analysis, the use of n-grams becomes more appropriate for feature selection [3]. While the n-gram technique is considered easy to implement and can be combined with other techniques, processing large-dimensional data remains a weakness [39]. To address this weakness, the current research employs FM to reduce feature dimensions, as FM can effectively operate on large-scale data [20] and handle spelling errors [50].

In machine learning, classification algorithms are employed to categorize data into classes. To effectively handle unstructured data, Decision Tree (DT) and KNN can serve as viable solutions [3, 4, 19]. NB is recognized for its speed and efficiency, offering ease of implementation [3, 4, 43, 45]. Two-way neural networks do not offer any advantages over standard neural networks because the standard artificial neural network provides slightly better results than two-way neural networks. Experimental results validate that the model offers very good results with a validation accuracy of 92.5% [24]. However, when dealing with complex, non-normally distributed data without intricate relationships, SVM proves to be the most suitable option [10, 11, 42, 43]. Table 1 displays previous research, all of which utilized manual data labeling. In contrast, this study employs the VADER Lexicon for data labeling, aiming to enhance sentiment analysis performance values.

Table 1. Comparison of sentiment analysis techniques for different data sources on customer reviews using manual labeling.

Author	Source of data	Handling feature	Classification algorithm	Accuracy
Arief and Deris [4]	Amazon product review	TF-IDF	DT, NB, SVM	88.13%
Firmansyah *et al.* [19]	Review from Twitter	TD-IDF	SVM, KNN	72.4%
Ruz *et al.* [43]	Critical event review	TF, TO, TF-IDF	NB, SVM, RF	80%
Araslanov *et al.* [3]	Short social network messages	TF-IDF, Chi-square, n-gram	LR, NB	80%
Setiabudi *et al.* [45]	Review from social media	n-gram	NB	78.3%

3. Method

The most commonly used method for sentiment classification is manual labeling, wherein humans read the text and assign it a sentiment label. However, manual labeling is time-consuming and requires significant effort, making it unsuitable for handling large amounts of data. In an effort to address this challenge, the study proposes an automatic sentiment classification method using the VADER Lexicon. This approach has demonstrated high accuracy, surpassing the precision achieved through manual labeling. Several comparisons are made between the classification performance of manually labeled data and data labeled automatically with VADER, as depicted in Figure 1.

3.1. Data Collection

In this study, the dataset used comprises customer reviews gathered through website or internet-based surveys and processed online. The customer review dataset is a collection of texts presented in various formats. Three datasets, sourced from diverse platforms, are utilized: dataset A includes product reviews from Amazon, dataset B consists of movie reviews, and dataset C encompasses airline reviews from Twitter. The data were sourced from the Kaggle website [23] and

Amazon Dataset [8]. Search keywords included "product review," "customer review," and "polarity review." The data is categorized into two polarities: positive and negative. Subsequently, the dataset files are consolidated into a single file and processed using Microsoft Excel, a word processing program. The customer review data remains in a raw form, featuring various characters, symbols, numbers, URLs, mentions, etc., necessitating preprocessing to enhance consistency and reduce noise.

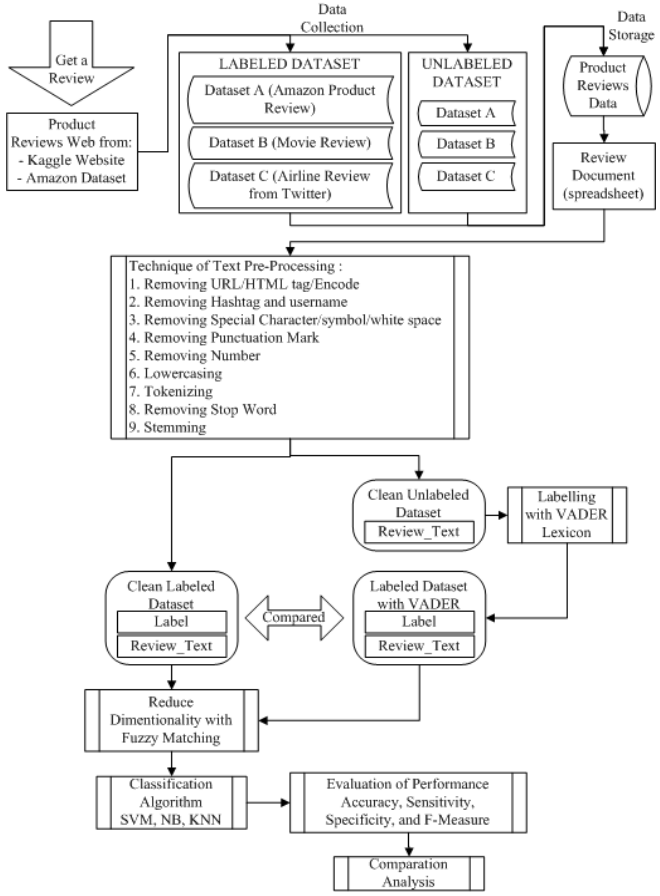

Figure 1. VADER Lexicon sentiment analysis research method.

3.2. Text Preprocessing Series

Preprocessing aims to clean the review data before processing and analysis to ensure that the data meets the requirements and is accurate [25]. Data in the form of Microsoft Excel files undergo nine preprocessing techniques, namely: removing URL/HTML tag/encode, removing hashtags and usernames, eliminating special characters/symbols/white spaces, removing punctuation marks, removing numbers, converting to lowercase, tokenizing, removing stop words, and stemming. These techniques are applied to make the data more structured and compatible.

3.3. Labeling with VADER Lexicon

At this stage, we label the dataset using the VADER Lexicon. Lexicon analysis is simpler than machine learning, which is more complex and demands more computing power [23]. In the Lexicon-based approach, a Lexicon captures words and their polarity according to the value of each word [30]. Previous research sources state that VADER is proficient in labeling sentiment data effectively [8]. When utilizing VADER, a lexical dictionary with 1773 weighted sentiment words is employed. The polarity value for each dataset is then determined, utilizing the weight or value of each word in a phrase to compute the overall polarity value. A positive sentiment is indicated by a polarity value greater than 0, while a negative sentiment is indicated by a polarity value less than 0 [31]. The method proposed in this work is to label datasets using the VADER Lexicon to expedite the labeling process and produce more precise results.

3.4. Dimentionality Reduction

FM algorithms help simplify features in text data. The FM algorithm utilizes a rule-based methodology to search for user-defined keywords and expressions. With a certain degree of matching error, this method can identify expressions in the text as they appear. Therefore, even with misspelled words and alternative suffixes or prefixes, matches can still be found in the text. The FM distance metric can be utilized to calculate the difference between two strings [33, 34].

3.5. Customer Review Classification

The customer review classification stage is executed using several algorithms, and for the distribution of training data, data testing is conducted through K-fold cross-validation with trials of various values of k. Cross-validation is a statistical method used to evaluate model performance by separating the data into two subsets: Training data and test data. K-fold cross validation is a specific case where the data is divided into k parts (folds) of equal size. One part is used for test data, while the remaining part (k-1) is used for training data. This process is repeated k times, ensuring each part serves as both training and test data [1].

As a learning algorithm based on optimization theory, SVM employs a hypothetical space in the form of a linear function in a high-dimensional feature space. By integrating learning bias generated from statistical learning, the algorithm is trained on parameters. This strategy operates by maximizing the distance between classes to find the optimal hyperplane. In a higher-dimensional class space, the hyperplane serves as a function that separates two classes for classification. SVM achieves linear separation by transforming data into a higher-dimensional space using kernel methods. Various kernel functions are frequently employed, such as the Radial Basis Function (RBF), polynomial, and linear functions [31]. To ensure comparability, SVM will be evaluated against similar classifiers, specifically NB and KNN [52].

3.6. Evaluation Performance

During the assessment phase, system testing is conducted to compute accuracy, sensitivity, specificity, and f-measure values to evaluate the performance of classification results. A metric commonly used to assess the accuracy of a strategy is performance evaluation. In this assessment, system development employs a confusion matrix [48, 49]. In real-world circumstances, the accuracy number calculates the percentage of correctly predicted outcomes. *Recall* measures the proportion of outcomes for which the correct value was determined, while precision reflects the correctness of the test results. The overall performance of the model is represented by the *F1-score*, calculated by summing the precision and recall values. The formulas used to determine the values are represented by Equations (1), (2), (3), and (4) below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision +\ Recall} \quad (4)$$

- **Information**
  - True Positive (TP)=shows that the outcome was accurately detected.
  - True Negative (TN)=indicates a negative result that was accurately recognized.
  - False Positive (FP)=displays negative findings found positive.
  - False Negative (FN)=displays good outcomes found negative.

3.7. Comparative Analysis

At this stage, compare the performance of sentiment analysis using a labeled dataset against a dataset automatically labeled with the VADER Lexicon.

4. Result

The collected dataset was meticulously analyzed to reveal its composition, as detailed in Table 2. The dataset comprises a rich collection of product reviews, each reflecting the sentiments of users who have purchased or utilized specific products or services.

Table 2. Distribution of positive and negative sentiment in different datasets.

Dataset	Data training of product review			
	#Pos	#Neg	#Total	#Words
Dataset A (Amazon product review)	254	247	501	41,262
Dataset B (movie review)	239	261	500	118,729
Dataset C (airline review from Twitter)	310	201	511	8,671

An analysis of Table 2 reveals that although the three datasets show a relatively similar number of data rows, their word counts vary significantly. Dataset B, which consists of in-depth movie reviews, stands out with the highest average word count. Although Dataset A also includes customer reviews, the word count is significantly lower, likely due to shorter and more concise product reviews. Dataset C, sourced from Twitter, shows the shortest word count, reflecting the character limit imposed by the social media platform. The number of positive and negative words is relatively balanced in datasets A and B, except in dataset C where there are more positive words.

Table 3. Text preprocessing steps and sample results.

Preprocessing stage	Sample result
Initial text	Caution!: These tracks are not the “original” versions but are re-recorded versions. So, whether the tracks are “remastered” or not is irrelevant. 2 stars of 5 for this track but bad $40. The link track at http://www.movienow.com just for @Robert #newcomer
Removing URL/HTML tag/encode	Caution!: These tracks are not the “original” versions but are re-recorded versions. So, whether the tracks are “remastered” or not is irrelevant. 2 stars of 5 for this track but bad $40. The link track at just for @Robert#newcomer
Removing hashtag and username	Caution!: These tracks are not the “original” versions but are re-recorded versions. So, whether the tracks are “remastered” or not is irrelevant. 2 stars of 5 for this track but bad $40. The link track at just for
Removing special character/symbol/ white space	Caution!: These tracks are not the “original” versions but are re-recorded versions. So, whether the tracks are “remastered” or not is irrelevant. 2 stars of 5 for this track but bad 40. The link track at just for
Removing punctuation mark	Caution These tracks are not the original versions but are re-recorded versions So whether the tracks are remastered or not is irrelevant 2 stars of 5 for this track but bad 40 The link track at just for
Removing number	Caution These tracks are not the original versions but are re-recorded versions So whether the tracks are remastered or not is irrelevant stars of for this track but bad The link track at just for
Lowercasing	caution these tracks are not the original versions but are re-recorded versions so whether the tracks are remastered or not is irrelevant stars of for this track but bad the link track at just for
Tokenizing	caution|these|tracks|are|not|the|original|versions|but|are|re-recorded|versions|so| whether|the|tracks|are|remastered|or|not| is|irrelevant|stars|of|for|this|track|but|bad| the|link|track|at|just|for
Removing stop word	caution|tracks|not|original|versions|re-recorded| versions|whether|tracks| remastered|not|irrelevant|stars|track|bad| link|track|just
Stemming	caution|track|not|original|version|re-record|version|whether|track|remaster| not|irrelev|star|track|bad|link|track|just

The initial stage of the process involves preprocessing the data, which encompasses several essential methods. As depicted in Table 3, the raw data consists of product reviews submitted by customers. The first step within preprocessing is data cleaning, where all non-letter characters and noise are meticulously removed. To ensure consistency, all letter cases are standardized to lowercase. Subsequently, stop

word removal is implemented, utilizing a pre-established stop word list to eliminate superfluous terms from the manuscript. The following stage entails stemming, which involves identifying the root form of each word and replacing it with the corresponding English grammatical structure. Finally, the documents are segmented into token portions by utilizing space characters as delimiters, thereby completing the tokenization process.

Before going through the preparation stage, the dataset will be automatically annotated using VADER as opposed to being annotated manually. With validated valence scores reflecting sensory polarity (positive/negative) and sensation strength on a scale of -4 (negative) to +4 (positive), with 0 denoting neutrality, the VADER Lexicon includes approximately 7500 lexical elements. The terms "okay," "for," "bad," and "sick," for example, had scores of 0.9, 3.1, -2.5, and -1.5, respectively. Each lexical feature in a text is evaluated by VADER, which then assigns a positive, negative, or neutral score to it. The compound score, a matrix that normalizes all scores from -1 to +1, is then created by adding these scores together. A composite score can be classified into one of three groups: neutral (between -0.05 and 0.05), negative (less than -0.05), or positive (greater than 0.05) [4]. The compound score of every word in the phrase determines its polarity. An example of how the VADER Lexicon is assessed is shown in Table 4.

Table 4. Scoring sentiment with VADER Lexicon.

Dataset	Review	Score	Annotation
Dataset A	Amazing!: This soundtrack is my favorite music…	6.8717	Positive
	The Worst!: A complete waste of time…	-3.7692	Negative
Dataset B	I have always been a huge fan of "Homicide:…	3.3589	Positive
	I never really understood the controversy…	-5.7692	Negative
Dataset C	@VirginAmerica it was amazing,…	1.2056	Positive
	@VirginAmerica and it's a really big bad thing…	-0.6410	Negative

The effectiveness of this search idea relies on the ability to determine whether a string being searched is similar to a string present in the dictionary, even when the character arrangement differs. The Levenshtein distance similarity is a function used to establish "similarity." Each word is compared with every other word to identify comparable words. To prevent the use of duplicate terms, only distinct, comparable words are retained.

Labeling the dataset using the VADER Lexicon produces a composition of positive and negative annotations, as seen in Figure 2. Customer reviews with the highest number of positive sentiments are in dataset A, whereas dataset B has the least number of positive reviews. This dataset is then classified using three algorithms, namely SVM, NB, and KNN.

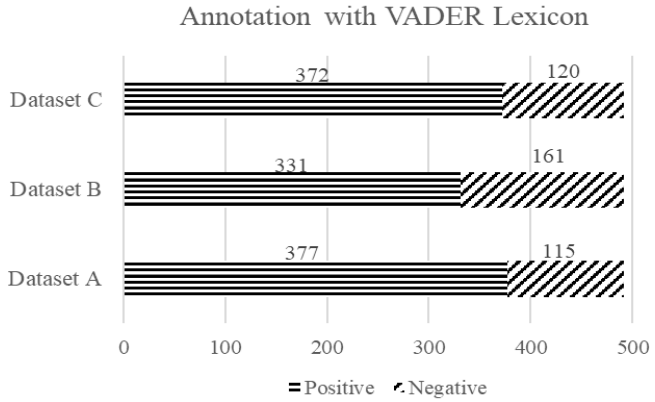

Figure 2. Visualization of sentiment annotation assigned by VADER Lexicon.

Classification was conducted on three datasets using three algorithms, and their respective performance results were subsequently compared. Table 5 presents the comprehensive data on the evaluation results. The classification performance of three datasets, whether manually labeled or automatically labeled with VADER, was compared. As depicted in the table's data, the accuracy values for data labeled with VADER exhibit superior values. Figure 3 illustrates that, overall, SVM outperforms other algorithms, whereas KNN demonstrates the opposite result, displaying the lowest performance. The primary result involves comparing classification performance by assessing a dataset labeled manually with a dataset labeled automatically using the VADER Lexicon. In dataset A, the accuracy was 94.69% for manual labeling and 95.92% for VADER labeling; in Dataset B, it was 96.94% for manual labeling and 97.01% for VADER labeling; and in Dataset C, it was 95.51% for manual labeling and 96.73% for VADER labeling.

Table 5. Comparative performance of sentiment analysis models.

Classification algorithm		Dataset					
		Dataset A		Dataset B		Dataset C	
		M	V	M	V	M	V
SVM	Accuracy	94.69	95.92	96.94	97.01	95.51	96.73
	Precision	93.14	96.00	94.24	95.71	98.47	94.07
	Recall	97.36	98.63	99.57	89.33	93.45	92.50
	F-Score	95.20	97.30	96.83	92.41	95.90	93.28
NB	Accuracy	86.33	86.87	88.37	88.78	84.29	86.33
	Precision	86.94	90.38	89.14	80.65	87.79	75.24
	Recall	87.92	90.14	85.65	83.33	83.64	65.83
	F-Score	87.43	90.26	87.36	81.97	85.66	70.22
KNN	Accuracy	76.33	85.71	83.06	84.08	76.73	80.82
	Precision	78.33	88.92	84.83	77.69	76.22	64.77
	Recall	77.74	92.33	77.83	67.33	85.09	47.50
	F-Score	78.03	90.59	81.18	72.14	80.41	54.81

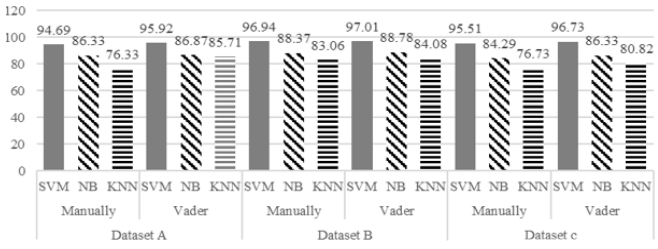

Figure 3. Classification accuracy of different models on various datasets.

While some improvements in accuracy were modest, as exemplified by the 0.07% increase for SVM on dataset B, others were considerably more substantial, such as the 9.38% increase for KNN on dataset A. The practical significance of these enhancements is contingent upon the specific application domain. In high-stakes contexts, even marginal improvements can have a substantial impact, particularly in fields like medical diagnosis or financial fraud detection, where accuracy is paramount. Furthermore, in large-scale applications processing millions of items, a seemingly minor 1% increase in accuracy can translate to a significant number of classifications being affected. Additionally, VADER demonstrated a more balanced precision and recall in certain instances, which could be advantageous depending on the specific requirements of the application.

The impact of VADER on our sentiment analysis task has significant implications for the precision, recall, and F1-scores observed across various datasets and classification algorithms. As a Lexicon and rule-based sentiment analysis tool, VADER exhibits varying effects on highly imbalanced datasets, which is evident in our study. The degree of imbalance in datasets A, B, and C influences how VADER's sentiment scores interact with our classification algorithms, including SVM, NB, and KNN.

Specifically, VADER's Lexicon-based approach appears to favor precision over recall, particularly for minority classes in imbalanced datasets. This is highlighted by the consistently high precision scores for SVM, which range from 93.14% to 98.47%. When examining algorithm-specific observations, SVM demonstrates high precision and recall across all datasets, suggesting its effective utilization of VADER's sentiment scores. Dataset B shows the most balanced performance, indicating an optimal interaction between VADER features and SVM for this particular data distribution.

In contrast, NB presents lower overall scores compared to SVM, which may imply that NB's probabilistic approach is less effective at leveraging VADER's deterministic sentiment scores. Notably, the significant drop in recall for Dataset C (65.83%) indicates potential challenges in addressing minority classes or extreme sentiments. KNN generally exhibits lower performance across metrics, revealing its unsuitability for VADER-processed features. The high variability in recall (ranging from 47.50% to 92.33%) suggests that KNN is particularly sensitive to dataset characteristics when working with VADER scores.

VADER's influence on metric variations is notable, as its rule-based nature contributes to relatively stable precision scores across datasets, particularly for SVM and NB. However, significant fluctuations in recall, especially for KNN, may be attributed to VADER's varying effectiveness in capturing nuanced sentiments in different dataset contexts. The F1-scores, which balance precision and recall, indicate that VADER's integration is most effective with SVM, providing a favorable compromise between identifying relevant instances and minimizing false positives.

Despite these insights, limitations remain. VADER's fixed Lexicon may fail to capture domain-specific sentiments, potentially hindering performance in specialized datasets. Additionally, exploring how adjustments to VADER's compound score thresholds affect classification performance could yield insights into optimizing its integration with machine learning models. Future work should consider combining VADER scores with other natural language processing features to enhance model performance, particularly for algorithms like KNN that demonstrated lower effectiveness. Furthermore, developing ensemble methods that incorporate VADER-based classifiers alongside other approaches may improve overall performance and robustness across diverse datasets.

By addressing these factors, we can deepen our understanding of how VADER influences our sentiment analysis task and identify strategies to leverage its strengths while mitigating its limitations in the context of imbalanced datasets.

5. Conclusions

This study investigated the effectiveness of a combined approach using FM and SVM for sentiment analysis in text data. Our findings demonstrate that utilizing the VADER Lexicon for feature extraction on unlabeled datasets and applying FM for dimensionality reduction, followed by SVM classification, yields high accuracy in sentiment polarity identification. Across three diverse datasets, the proposed model achieved accuracy exceeding 94% for both manually labeled and VADER-labeled data. These results highlight the efficacy of the combined FM-SVM approach in accurately assessing sentiment in text, even with unlabeled data.

By leveraging the advantages of VADER Lexicon, FM, and SVM, this study paves the way for efficient and accurate sentiment analysis in diverse text data, even with unlabeled information. As research in this area continues, further advancements in sentiment analysis can unlock valuable insights from the vast amount of textual data generated online, facilitating informed decision-making, and fostering enhanced communication across various domains. For future research, explore the integration of different language models to improve semantic understanding. Investigating the application of this model to multilingual sentiment analysis. or also examine the model's effectiveness on more complex and nuanced emotions. We encourage future research to explore VADER's performance with advanced techniques that combine complex transformer-based models, which can provide a more comprehensive understanding of its applicability across various methods.

Acknowledgment

This study was made possible by the generous funding provided by the international grant UIC241504. The authors are deeply indebted to Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), Al Maarif University College, and Universitas Dinamika for their invaluable contributions in terms of facilities, expertise, and collaboration.

References

[1] Agatha R. and Polina A., "Analisis Sentimen Terhadap Penggunaan Marketplace di Indonesia Menggunakan Metode Support Vector Machine dengan Seleksi Fitur Chi Square," *Seminar Nasional Riset and Inovasi Teknologi*, vol. 1, no. 1, pp. 314-323, 2022. https://e-proceeding.itp.ac.id/index.php/sinarint/article/view/63/37

[2] Anggraeni W., Roji F., and Alkautsar M., "Analisis Sentimen Publik Terhadap Kebijakan Insentif Perpajakan Dengan Pendekatan VADER (Valence Aware Dictionary and Sentiment Reasoner)," *Jurnal Proaksi*, vol. 10, no. 4, pp. 465-477, 2023. DOI: 10.32534/jpk.v10i4.4732

[3] Araslanov E., Komotskiy E., and Agbozo E., "Assessing the Impact of Text Preprocessing in Sentiment Analysis of Short Social Network Messages in the Russian Language," *in Proceedings of the International Conference on Data Analytics for Business and Industry: Way towards a Sustainable Economy*, Sakheer, pp. 1-4, 2020. DOI:10.1109/ICDABI51230.2020.9325654

[4] Arief M. and Deris M., "Text Preprocessing Impact for Sentiment Classification in Product Review," *in Proceedings of the 6th International Conference on Informatics and Computing*, Jakarta, pp. 1-7, 2021. DOI:10.1109/ICIC54025.2021.9632884

[5] Arya V., Mishra A., and Gonzalez-Briones A., "Sentiments Analysis of Covid-19 Vaccine Tweets Using Machine Learning and VADER Lexicon Method," *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 11, no. 4, pp. 507-518, 2023. DOI: 10.14201/adcaij.27349

[6] Asri Y. and Fajri M., "Sentiment Analysis of PLN Mobile Review Data Using Lexicon VADER and Naive Bayes Classification," *in Proceedings of the International Conference on Networking, Electrical Engineering, Computer Science, and Technology*, Bandar Lampung, pp. 132-137, 2023. DOI:10.1109/IConNECT56593.2023.10327064

[7] Barushka A. and Hajek P., "The Effect of Text Preprocessing Strategies on Detecting Fake Consumer Reviews," *in Proceedings of the 3rd ACM International Conference Proceeding Series, Association for Computing Machinery*, pp. 13-17, 2019. DOI:10.1145/3383902.3383908

[8] Bashar M., "A Hybrid Approach to Explore Public Sentiments on COVID-19," *SN Computer Science*, vol. 3, no. 3, pp. 1-19, 2022. DOI:10.1007/s42979-022-01112-1

[9] Bernhardt M., Castro D., Tanno R., Schwaighofer A., Tezcan K., Monteiro M., Bannur S., Lungren M., Nori A., Glocker B., Alvarez-Valle J., and Oktay O., "Active Label Cleaning for Improved Dataset Quality under Resource Constraints," *Nature Communications*, vol. 13, no. 1, pp. 1-11, 2022. DOI:10.1038/s41467-022-28818-3

[10] Borg A. and Boldt M., "Using VADER Sentiment and SVM for Predicting Customer Response Sentiment," *Expert Systems with Applications*, vol. 162, pp. 113746, 2020. DOI:10.1016/j.eswa.2020.113746

[11] Chouhan K., "Sentiment Analysis with Tweets Behaviour in Twitter Streaming API," *Computer Systems Science and Engineering*, vol. 45, no. 2, pp. 1113-1128, 2023. DOI:10.32604/csse.2023.030842

[12] De Oliveira D. and De Campos Merschmann L., "Joint Evaluation of Preprocessing Tasks with Classifiers for Sentiment Analysis in Brazilian Portuguese Language," *Multimedia Tools and Applications*, vol. 80, pp. 15391-15412, 2021. DOI: 10.1007/s11042-020-10323-8

[13] Ding Y., You J., Machulla T., Jacobs J., Sen P., and Hollerer T., "Impact of Annotator Demographics on Sentiment Dataset Labeling," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1-22, 2022. DOI:10.1145/3555632

[14] Drus Z. and Khalid H., "Sentiment Analysis in Social Media and its Application: Systematic Literature Review," *Procedia Computer Science*, vol. 161, pp. 707-714, 2019. https://doi.org/10.1016/j.procs.2019.11.174

[15] Duong H. and Nguyen-Thi T., "A Review: Preprocessing Techniques and Data Augmentation for Sentiment Analysis," *Computational Social Networks*, vol. 8, no. 1, pp. 1-16, 2021. DOI:10.1186/s40649-020-00080-x

[16] Ernawati S. and Wati R., "Evaluasi Performa Kernel SVM dalam Analisis Sentimen Review Aplikasi ChatGPT Menggunakan Hyperparameter dan VADER Lexicon," *Jurnal Buana Informatika*, vol. 15, no. 01, pp. 40-49, 2024. DOI:10.24002/jbi.v15i1.7925

[17] Es-Sabery F., Es-Sabery I., Hair A., Sainz-De-Abajo B., and Garcia-Zapirain B., "Emotion Processing by Applying a Fuzzy-Based VADER Lexicon and a Parallel Deep Belief Network Over Massive Data," *IEEE Access*, vol. 10, pp. 87870-87899, 2022. https://ieeexplore.ieee.org/document/9863839

[18] Fathoni M., Puspaningrum E., and Sihananto A.,

"Perbandingan Performa Labeling Lexicon InSet dan VADER pada Analisa Sentimen Rohingya di Aplikasi X dengan SVM," *Modem: Jurnal Informatika dan Sains Teknologi*, vol. 1, no. 3, pp. 62-76, 2024. https://doi.org/10.62951/modem.v1i3.112

[19] Firmansyah F., Zulfikar W., Maylawati D., Arianti N., Muliawaty L., Septiadi M., and Ramdhani M., "Comparing Sentiment Analysis of Indonesian Presidential Election 2019 with Support Vector Machine and K-Nearest Neighbor Algorithm," *in Proceedings of the 6th International Conference on Computing Engineering and Design*, Sukabumi, pp. 1-6, 2020. DOI:10.1109/ICCED51276.2020.9415767

[20] Gao Y., Zhang H., Li S., Shi C., and Gao H., "Short Circuit Fault Location Method of Distribution Network Based on Fuzzy Matching," *in Proceedings of the IEEE 6th Conference on Energy Internet and Energy System Integration (EI2)*, Chengdu, pp. 1499-1505, 2022. DOI:10.1109/EI256261.2022.10116989

[21] Hamka M. and Tukiran., "Analisis Sentimen Pengguna E-Commerce dan Marketplace Menggunakan Support Vector Machine," *Jurnal Rekayasa Sistem Informasi dan Teknologi*, vol. 1, no. 4, pp. 273-282, 2024. https://doi.org/10.59407/jrsit.v1i4.555

[22] Hong Y. and Shao X., "Emotional Analysis of Clothing Product Reviews Based on Machine Learning," *in Proceedings of the 3rd International Conference on Applied Machine Learning*, Changsha, pp. 398-401, 2021. DOI:10.1109/ICAML54311.2021.00090

[23] Hossen M. and Dev N., "An Improved Lexicon Based Model for Efficient Sentiment Analysis on Movie Review Data," *Wireless Personal Communications*, vol. 120, no. 1, pp. 535-544, 2021. DOI:10.1007/s11277-021-08474-4

[24] Humayun M., Javed D., Jhanjhi N., Almufareh M., and Almuayqil S., "Deep Learning Based Sentiment Analysis of COVID-19 Tweets via Resampling and Label Analysis," *Computer Systems Science and Engineering*, vol. 47, no. 1, pp. 575-591, 2023. DOI:10.32604/csse.2023.038765

[25] Khader M., Awajan A., and Al-Naymat G., "The Impact of Natural Language Preprocessing on Big Data Sentiment Analysis," *The International Arab Journal of Information Technology*, vol. 16, no. 3A, pp. 506-513, 2019. https://iajit.org/portal/PDF/Special%20Issue%20 2019,%20No.%203A/18596.pdf

[26] Lee E., Rustam F., Shahzad H., Washington P., Ishaq A., and Ashraf I., "Drug Usage Safety from Drug Reviews with Hybrid Machine Learning Approach," *Computer Systems Science and Engineering*, vol. 45, no. 3, pp. 3053-3077, 2023. DOI:10.32604/csse.2023.029059

[27] Li M. and Shi Y., "Sentiment Analysis and Prediction Model Based on Chinese Government Affairs Microblogs," *Heliyon*, vol. 9, no. 8, pp. 1-16, 2023. DOI:10.1016/j.heliyon.2023.e19091

[28] Mahilraj J., Tigistu G., and Tumsa S., "Text Preprocessing Method on Twitter Sentiment Analysis Using Machine Learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 11, pp. 233-240, 2020. DOI:10.35940/ijitee.K7771.0991120

[29] Mardjo A. and Choksuchat C., "HyVADRF: Hybrid VADER-Random Forest and GWO for Bitcoin Tweet Sentiment Analysis," *IEEE Access*, vol. 10, pp. 101889-101897, 2022. DOI:10.1109/ACCESS.2022.3209662

[30] Maree M., Eleyat M., and Mesqali E., "Optimizing Machine Learning-based Sentiment Analysis Accuracy in Bilingual Sentences via Preprocessing Techniques," *The International Arab Journal of Information Technology*, vol. 21, no. 2, pp. 257-270, 2024. https://doi.org/10.34028/iajit/21/2/8

[31] Muhammadi R., Laksana T., and Arifa A., "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. https://doi.org/10.23917/khif.v8i1.15213

[32] Nasser A. and Sever H., "A Concept-based Sentiment Analysis Approach for Arabic," *The International Arab Journal of Information Technology*, vol. 17, no. 5, pp. 778-788, 2020. https://doi.org/10.34028/iajit/17/5/11

[33] Nurcahyawati V. and Mustaffa Z., "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, 2023. DOI:10.11591/eei.v12i3.4830

[34] Oliveira M., Mourthe A., and Duque M., "Extracting Events from Daily Drilling Reports Using Fuzzy String Matching," *The APPEA Journal*, vol. 62, no. 2, pp. S158-S161, 2022. DOI:10.1071/aj21118

[35] Patil R., Peshave P., and Kamble M., "Application of Fuzzy Matching Algorithms for Doctors Handwriting Recognition," *in Proceedings of the IEEE Bombay Section Signature Conference*, Mumbai, pp. 1-5, 2022. DOI:10.1109/IBSSC56953.2022.10037486

[36] Prasetyo A., Ridwan T., and Voutama A., "Analisis Sentimen Terhadap Aplikasi GBWhatsapp Menggunakan Naive Bayes Classifier dan Random Forest Classifier," *Jurnal Sistem Informasi*, vol. 11, no. 1, pp. 1-9, 2024. https://doi.org/10.30656/jsii.v11i1.6936

[37] Qureshi M., Asif M., Hassan M., Mustafa G.,

Ehsan M., Ali A., and Sajid U., "A Novel Auto-Annotation Technique for Aspect Level Sentiment Analysis," *Computers, Materials and Continua*, vol. 70, no. 3, pp. 4987-5004, 2022. DOI:10.32604/cmc.2022.020544

[38] Rahman R., Pranatawijaya V., and Sari N., "Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Gojek," *Konvergensi Teknologi dan Sistem Informasi*, vol. 4, no. 1, pp. 70-82, 2024. DOI:10.24002/konstelasi.v4i1.8922

[39] Rajput G., Kundu S., and Kumar A., "The Impact of Feature Extraction on Multi-Source Sentiment Analysis," *in Proceedings of the 10th International Conference on System Modeling and Advancement in Research Trends*, Moradabad, pp. 510-515, 2021. DOI:10.1109/SMART52563.2021.9676201

[40] Rohman I., Aqharabah B., and Solekan R., "Chatbot Untuk Cek Persediaan Stok Barang Menggunakan Metode Fuzzy String Matching Berbasis Mobile," *Prosiding Seminar Nasional Teknologi dan Sains, Kediri: Universitas Nusantara PGRI Kediri*, vol. 2, pp. 281-286, 2023. https://doi.org/10.29407/stains.v2i1.2840

[41] Romadhon M. and Kurniawan F., "A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia," *in Proceedings of the 3rd East Indonesia Conference on Computer and Information Technology*, Surabaya, pp. 41-44, 2021. DOI:10.1109/EIConCIT50028.2021.9431845

[42] Rukhsar S., Awan M., Naseem U., Zebari D., Mohammed M., Albahar M., Thanoon M., and Mahmoud A., "Artificial Intelligence Based Sentence Level Sentiment Analysis of COVID-19," *Computer Systems Science and Engineering*, vol. 47, no. 1, pp. 791-807, 2023. DOI:10.32604/csse.2023.038384

[43] Ruz G., Henriquez P., and Mascareno A., "Sentiment Analysis of Twitter Data during Critical Events through Bayesian Networks Classifiers," *Future Generation Computer Systems*, vol. 106, pp. 92-104, 2020. DOI:10.1016/j.future.2020.01.005

[44] Safitri Y., Kurniawan R., and Suhardi, "Analisis Sentimen Mengenai Childfree Menggunakan Metode Naïve Bayes," *The Indonesian Journal of Computer Science*, vol. 13, no. 4, pp. 6320-6332, 2024. DOI:10.33022/ijcs.v13i4.4136

[45] Setiabudi R., Iswari N., and Rusli A., "Enhancing Text Classification Performance by Preprocessing Misspelled Words in Indonesian Language," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 19, no. 4, pp. 1234-1241, 2021. DOI:10.12928/TELKOMNIKA.v19i4.20369

[46] Shaban W., Rabie A., Saleh A., and Abo-Elsoud M., "Accurate Detection of COVID-19 Patients based on Distance Biased Naïve Bayes (DBNB) Classification Strategy," *Pattern Recognition*, vol. 119, pp. 1-15, 2021. DOI:10.1016/j.patcog.2021.108110

[47] Shirazi G., Azmi R., and Shakibian H., "A Semi-Automated Labeled Data Generation Approach Based on Deep Learning to Improve Sentiment Analysis in the Persian Language," *in Proceedings of the 9th International Conference on Web Research*, Tehran, pp. 242-246, 2023. DOI:10.1109/ICWR57742.2023.10138965

[48] Sutoyo E., Rifai A., Risnumawan A., and Saputra M., "A Comparison of Text Weighting Schemes on Sentiment Analysis of Government Policies: A Case Study of Replacement of National Examinations," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6413-6431, 2022. DOI:10.1007/s11042-022-11900-9

[49] Wankhade M., Rao A., and Kulkarni C., "A Survey on Sentiment Analysis Methods, Applications, and Challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731-5780, 2022. DOI:10.1007/s10462-022-10144-1

[50] Yang W., Xu J., Xiang J., Yan Z., Zhou H., Wen B., Kong H., Zhu R., and Li W., "Diagnosis of Cardiac Abnormalities Based on Phonocardiogram Using a Novel Fuzzy Matching Feature Extraction Method," *BMC Med Inform Decis Mak*, vol. 22, no. 1, pp. 1-13, 2022. DOI:10.1186/s12911-022-01976-6

[51] Yu H. and Kim J., "Indoor Positioning by Weighted Fuzzy Matching in Lifi Based Hospital Ward Environment," *in Proceedings of 4th International Conference on Control Engineering and Artificial Intelligence*, Singapore, pp. 1-6, 2020. DOI:10.1088/1742-6596/1487/1/012010

[52] Zhao W., Guan Z., Chen L., He X., Cai D., Wanget B., and Wang Q., "Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis," *IEEE Transactions on Knowledge and Data* Engineering, vol. 30, no. 1, pp. 185-197, 2018. DOI:10.1109/TKDE.2017.2756658

Vivine Nurcahyawati is a Ph.D. candidate at the Faculty of Computing, Universiti of Malaysia Pahang, her interest is in Data Mining and Natural Language Processing. She has 7 years of experience as a System Analyst and Database Administrator and 17 years in teaching at the Department of Information System, Universitas Dinamika, Surabaya, Indonesia. Her research areas include Data Mining, Natural Language Processing, and Software Engineering.

Zuriani Mustaffa is Senior Lecturer in Faculty of Computing, Universiti Malaysia Pahang, Malaysia. Her research interest includes Computational Intelligence (CI) algorithm, specifically in Swarm Intelligence (SI) and Machine Learning Techniques. Her research area focuses on Hybrid Algorithms which involves optimization and machine learning techniques with particular attention for time series predictive analysis. She has authored and co-authored various scientific articles in the field of interest.

Mohammed Khalaf is a Director of Quality Assurance and Accreditation, and Senior Lecture at Al-Maarif University college, and Iraq. Dr. Khalaf received his B.Sc. Degree in Information System from the College of Computer, University of Anbar, in 2008, Iraq. He gained his Master in Information Technology from Universiti Tenaga Nasional (UNITEN), in 2012, Malaysia. He obtained his Ph.D. in Computer Science from Liverpool John Moores University, in 2018, UK. His research interests include Data Science, Artificial Intelligence, Machine Learning and Advanced Algorithm Development, Network and Communication.