Discovering User Sentiment Patterns in Libraries with a Hybrid Machine Learning and Lexicon-Based Approach

Dini Nurmalasari[1]*, Dini Hidayatul Qudsi Authors [2], Nessa Chairani [3], Heri R Yuliantoro [4]
Department of Information Technology, Computer Engineering Technology, Caltex Riau Polytechnic [1], [2], [3]
Department of Tax Accounting, Caltex Riau Polytechnic [4]
dini@pcr.ac.id [1], dinihq@pcr.ac.id[2], nessa@pcr.ac.id [3], heriry@pcr.ac.id [4]

***Abstract*— The need to enhance library services is the focus of this study, which relies on user feedback for data-driven decision-making. Text data from library user surveys conducted at Politeknik Caltex Riau (PCR) is analyzed to categorize sentiment and identify areas for improvement. The biannual student and lecturer feedback collected from 2018 to 2023 through the institution's official survey system (survey.pcr.ac.id) is utilized, providing a comprehensive and robust picture of user needs across five years. Sentiment analysis is employed using the VADER method to classify user comments into positive or negative categories. Text preprocessing techniques, such as stemming, tokenizing, and filtering, are performed to ensure robust classification. Machine learning algorithms – Naïve Bayes, Support Vector Machine (SVM), and Random Forest – are then utilized to evaluate sentiment classification accuracy. The study offers significant findings. Both SVM and Random Forest achieve an outstanding accuracy of 99%, indicating highly reliable sentiment categorization. Notably, these algorithms also achieve 100% precision, recall, and F1-score, demonstrating their effectiveness in accurately identifying positive and negative user sentiment. While Naïve Bayes shows slightly lower accuracy at 98%, it maintains a high recall rate (100%), ensuring all negative feedback is captured. This research presents a novel approach combining user sentiment analysis with a comprehensive five-year dataset. This enables a deeper understanding of evolving user needs and priorities. The high accuracy and effectiveness of the employed algorithms highlight the potential of this methodology for libraries. Libraries can leverage user feedback for evidence-based service improvement and increased user satisfaction.**

Keywords*— *Sentiment Analysis; Vader Lexicon; Random Forest, Naïve Bayes, Library Opinion

I. INTRODUCTION

One method used to examine user comments, feedback, answers, and recommendations is sentiment analysis. Because of unstructured data and many linguistic variances, this process might be complex. The results of sentiment analysis can identify particular sentiments and emotions, such as happiness, sadness, or rage, in addition to identifying whether a comment is good, negative, or neutral. As a result, sentiment analysis is widely used to improve performance in a variety of industries, especially business.

To further improve services, the Perpustakaan Politeknik Caltex Riau (PCR) is making one such performance optimization initiative. PCR keeps improving its facilities and services as part of these initiatives, which include growing its library of books, course materials, and journals. In order to provide cross-collection services, partnerships with libraries from other universities are also sought. Furthermore, the French area of PCR's library includes basic information about France as well as insights into French culture. PCR added the Bank Indonesia Corner as a new facility in 2019. Nevertheless, PCR's statistical data on library patrons and book-borrowing members indicates a negligible rise in spite of these initiatives. According to infographic data from 2020, 2021, and 2022 (as seen in Figure 1), the number of visitors decreased in 2023, while the small number of visitors in 2021 was ascribed to distant learning practices.

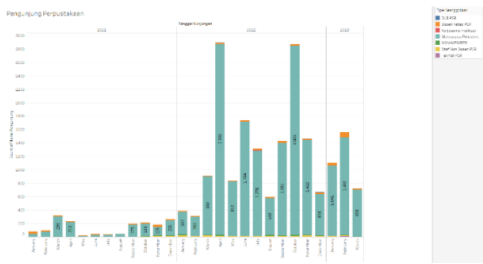


Figure 1 Number of Visits to PCR Library

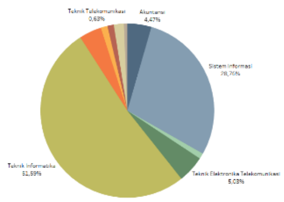


Figure 2 Number of Visits to PCR Library by Prodi

Currently, library patrons send suggestions—like adding book collections to the library—directly to the manager for prompt consideration, such as through email. The manager can immediately address these suggestions by giving library patrons a Google form to complete about which books should be budgeted for next year's purchase. However, this approach is

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2217, Copyright ©2024
Submitted : July 16, 2024 , Revised : August 30, 2024, Accepted : September 11, 2024, Published : November 19, 2024

limited to some library patrons, particularly students. They feel more at ease using the feedback mechanism to provide ideas and comments.

This project will do sentiment analysis based on user comments about the library's resources and services in order to help improve library services. The goal of this sentiment study is to determine how users feel about the facilities and services offered by PCR's library so that improvements can be made. The VADER approach will be used to analyze the sentiment of textual user comments about PCR's library. By using evaluations from the dictionary to simulate human judgments, VADER (Valence Aware Dictionary for Sentiment Reasoning) will categorize data according to the compound value acquired.

This project will do sentiment analysis based on user comments about the library's resources and services in order to help improve library services. The goal of this sentiment study is to determine how users feel about the facilities and services offered by PCR's library so that improvements can be made. The VADER approach will be used to analyze the sentiment of textual user comments about PCR's library. By using evaluations from the dictionary to simulate human judgments, VADER (Valence Aware Dictionary for Sentiment Reasoning) will categorize data according to the compound value acquired.

A study on Sentiment Analysis of Ciletuh Geopark Tourist Destinations using VADER was carried out in 2022 by [4] and associates, and the classification of more than 75% favorable sentiment was the outcome. Using a lexicon-based approach on Twitter, Putri Amira et al. conducted another study in 2021 on the sentiment analysis of COVID-19 vaccinations in Indonesia. They found that over 20% of respondents had favorable sentiments and 44.36% had neutral sentiments. As a result, it was concluded that the public's attitude about COVID-19 vaccinations was neutral. Furthermore, Dicky Abimanyu and colleagues' 2022 study, Sentiment Analysis of Twitter Accounts in Apex Legends using VADER, produced a classification accuracy of 65.2% across 500 test samples, with 18% of the feelings being positive, 4.8% being negative, and 77.2% being neutral. Text classification, another name for sentiment analysis, is a broad field. It entails examining textual input to determine whether it contains neutral, negative, or positive attitudes. [5] [6]. Sentiment analysis of Twitter data, product reviews, and student evaluations of lecturers' performance have all been studied in relation to Indonesian text.

According to [7] book, the text preprocessing stage aims to Choose which data to clean. This phase includes filtering, normalization, tokenization, and transformation. Case folding, lowercase letter conversion, stopword removal (removing unnecessary words), and emoticon removal (removing emoticons from sentences) are other common preprocessing techniques. One technique used as a model for sentiment analysis is the Valence Aware Dictionary and Sentiment Reasoner (VADER) book [8]. It can use the existing lexical data to determine data diversity based on intensity. E. Hutto claims that the VADER technique is categorized using a human-centric approach that combines empirical validation with quantitative analysis, utilizing human judgment and wisdom. By assigning "negative," "positive," and "neutral" labels to phrases or sentences, VADER's lexical dictionary evaluates their mood. The entire statement is given a score by the lexical dictionary, which assesses words according to their sentiment category.

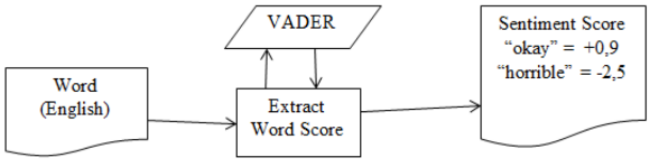


Figure 3 Polarity Score Process

To determine the polarity class of feelings, VADER will use lexical data from the NLTK server. It will then aggregate the 'compound' attribute from every available word to calculate the polarity of the sentence. "Positive" results are those with a compound score greater than 0.5; "neutral" results are those between -0.5 and 0.5; and "negative" results are those with a compound score less than -0.5.[9][10]

A confusion matrix is an evaluation tool for determining the efficiency or accuracy of the categorization process. Among the metrics used are F-measure, accuracy, recall, and precision [11]. Recall: Indicates the proportion of expectedly positive feedback data that were actually classified as such. Recall is equal to TPTP+FN. Precision: Indicates the percentage of feedback data correctly recognized as positive out of all data thus labeled. Precision is TPTP+FN. Accuracy: Shows how many feedback data points were accurately categorized (identified as positive or negative). TP+TNTP+FN+FP+FN = Accuracy. F1-Score: Shows the precision and recall harmonic mean. F1 Score = FN+FN+2TP2TP. [12]

Following data labeling, data modeling is carried out using classification techniques such as Random Forest, SVM, and Naïve Bayes. Because of its probabilistic methodology, the Naïve Bayes Algorithm is a popular option for sentiment analysis classification. Assuming independence between every pair of class variables, this supervised learning algorithm uses the Naïve Bayes approach [13]. applying the Bayes theorem. The fundamental algorithm used by Naïve Bayes is shown in the following:

$$P(A|B)=P(B)P(B|A)\times P(A) \quad (1)$$

where this probability is represented by P(A|B). Whereas (P(A) and (P(B)) represent the probabilities of A and B happening separately, (P(B|A) represents the likelihood of B happening with evidence of A.

The random forest algorithm is used for regression and classification. It works by creating a large number of decision trees during training and using the individual trees to determine the class mode (for classification) or average prediction (for regression) [14]. The durability, versatility, and ability to handle large datasets while reducing the risk of overfitting make

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2217, Copyright ©2024
Submitted : July 16, 2024 , Revised : August 30, 2024, Accepted : September 11, 2024, Published : November 19, 2024

this ensemble technique highly regarded [15] [16][17]. Support Vector Machines, or SVMs, are a reliable supervised learning method used for regression and classification tasks [18]. Its basic idea is to maximize the margin between data points in order to determine the optimal hyperplane for classifying them into discrete groups or for predicting continuous outcomes in regression. High-dimensional data can be handled effectively with SVM [19]. This template gives authors the formatting guidelines they need to create electronic copies of their papers. It was altered in Microsoft Word 2007 and saved as a "Word 97-2003 Document" for the PC. For three reasons, all standard paper components have been specified: (1) they are simple to use when formatting individual papers; (2) they automatically comply with electronic requirements, which makes it easier to produce electronic products later on or concurrently; and (3) they maintain style consistency throughout conference proceedings. There are built-in margins, column widths, line spacing, and type styles. Type style samples are given throughout the document and are indicated by italics in parentheses after the example. While many table text styles are offered, other elements—like multi-leveled equations, images, and tables—are not required. These components must be created by the formatter while taking into account the relevant criteria listed below.

Although user sentiment analysis has been the subject of numerous studies in various fields, additional work is required to adapt lexicon-based and hybrid machine learning techniques to the unique setting of libraries. Previous studies have focused on lexicon-based approaches or conventional machine-learning techniques [20], ignoring the potential for synergy [21]. Additionally, rather than thoroughly examining the range of data sources accessible inside the library setting, studies investigating user sentiment trends in libraries have frequently been limited to particular features, such as social media interactions or feedback forms [22]. This research gap offers a chance to contribute significantly to the field by creating a thorough framework that integrates the advantages of hybrid approaches to better understand user sentiments and obtain deeper insights into the dynamics of user sentiment within libraries.

II. Methodology

The following steps will be taken when utilizing VADER to do sentiment analysis on library user comments:

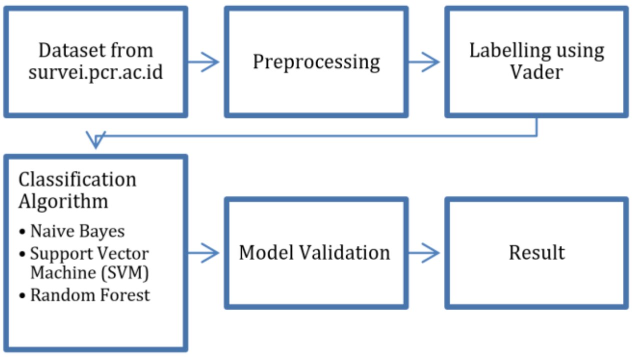


Figure 4 Research Methods

To begin this investigation, datasets from the survey were gathered using the research methodology shown in Figure 4.data from the pcr.ac.id website covering the years 2018–2023, totaling around 8,500 data points. Time-related data, program study specifics, and text-formatted feedback data were among the information gathered. To improve data purity and guarantee higher accuracy, a preprocessing step was then carried out. Using Tableau Prep Builder, the preparation included tasks like cleaning up the text feedback by deleting excessive spaces, deleting punctuation, changing the text to lowercase, fixing inconsistent wording, etc.

The Python programming language was then used to carry out additional preprocessing. Transformations, tokenization, normalization, stop word elimination, and sentence structure enhancements were all part of this phase. Sentiment analysis using rule-based VADER will be applied to cleaned data. Sorting sentiment into positive and negative categories is the aim. In order to rate texts on a scale from -1 (extremely negative) to 1 (positive), with 0 denoting neutrality, VADER will be executed using Python programming to determine sentiment polarity class. A compound score that ranges from -1 (most negative) to 1 (most optimistic) is the outcome of the final polarity computation. The information and procedures in sentiment analysis and text processing relate to building a backend for the intended application.[9].

Three machine learning algorithms will then be used to group the outcomes of the labeling process for both positive and negative feelings. Random Forest, SVM, and Naïve Bayes are the algorithms used. The Confusion Matrix will be used to validate the correctness of the modeling results produced by these three algorithms. To aid with interpretation, the analysis's findings will be graphically represented. Different kinds of graphs designed to meet particular requirements will be employed. The dashboard representation that is produced will be front-end web-based.

III. Result and Discussions

This section, which consists of multiple sections, will address the findings and analytical results based on the study methodology that was carried out.:

3. 1. Data Collection

Based on the library's feedback category, the survey.pcr.ac.id system provided the data. For data cleansing purposes, the data was then saved in Excel format. Faculty and student feedback processes are facilitated by the survey.pcr.ac.id system, which evaluates service performance in all departments on the Politeknik Caltex Riau campus, including the library department. Approximately 7693 items from four years—2018, 2021, 2022, and 2023—make up the data used in this study. Due to the COVID-19 epidemic, which caused classes and library visits to be suspended, data from 2019 and 2020 were not used..

3.2 Pre-processing

The results of each step in the pre-processing phase are described below. During the data cleaning process, a number of

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2217, Copyright ©2024
Submitted : July 16, 2024 , Revised : August 30, 2024, Accepted : September 11, 2024, Published : November 19, 2024

tasks were carried out, such as converting text into lowercase, tokenizing or simplifying words to their base forms, segmenting text into individual units, and eliminating unnecessary words.

```
def preprocessing(text):
    def strip_html(text):
        soup = BeautifulSoup(text, "html.parser")
        return soup.get_text()
    def remove_between_square_brackets(text):
        return re.sub('-\[[^]]*\]', '', text)
    def denoise_text(text):
        text = strip_html(text)
        text = remove_between_square_brackets(text)
        return text
    def remove_punctuation(text):
        return re.sub(r'[^\w\s]', '', text)
    def remove_non_ascii(text):
        return unicodedata.normalize('NFKD', text).encode('ascii', 'ignore').decode('utf-8', 'ignore')

    text = strip_html(text)
    text = remove_between_square_brackets(text)
    text = remove_punctuation(text)
    text = remove_non_ascii(text)

    return text.lower()
```

Figure 5 Cleansing Process

In order to clean text data by eliminating certain elements including HTML tags, punctuation, non-ASCII characters, and text enclosed in square brackets, this code uses a Python script for text preprocessing operations.

3.3 Punctuation removal

This step is used to remove punctuation in order to highlight specific words in a statement. In literary representation, punctuation has little bearing on meaning. Additionally, this step improves text representation consistency [23].

```
dataframe['Saran'] = dataframe['Saran'].str.replace('[^\w\s]','')
dataframe.head(10)
```

This approach involves substituting an empty character for each character other than letters, digits, or spaces. As a result, this stage's final output only includes letters, numbers, and spaces..

Lowercase

In order to maintain consistency during text analysis, lowercasing is the process of changing text into lowercase characters. In order to reduce the discrepancies in text representation, this step is essential [14]. The process involves utilizing a lambda function that breaks the text up into individual words, lowercases each word, and then uses the join command to bring these altered words back together. The Python syntax that demonstrates this procedure is shown below.

```
dataframe['Saran'] = dataframe['Saran'].apply(lambda x: " ".join(x.lower() for x in x.split()))
dataframe.head(10)
```

3.4 Stop Words Removal

Through this procedure, words like "the," "and," "is," and others that don't significantly add to a sentence's content are removed. It helps to improve data consistency, computational efficiency, and noise reduction [15]. Creating a variable to hold a list of English stop words and coming up with a method to distinguish between words that are on the list and those that are not are among the stages involved. Words that do not fit into the category of stop words are kept, while those that do are eliminated [16].

```
stop = stopwords.words('english')
dataframe['Saran'] = dataframe['Saran'].apply(lambda x: " ".join(x for x in x.split() if x not in stop))
dataframe['Saran'].head(10)
```

3.5 Spelling Correction

The process of locating and fixing spelling mistakes or inaccuracies that arise in a document is known as spelling correction.[24] This procedure could entail using dictionaries, statistical patterns, or language norms to replace misspelled words with their proper equivalents. Correcting typographical or spelling errors is the main goal of spelling correction, which also aims to increase the text's accuracy and readability.[25]

```
dataframe['Saran'].apply(lambda x: str(TextBlob(x).correct()))
dataframe['Saran'].head(10)
```

3.6 *Lemmitization*

Lemmatization is a method for converting text into its root or base forms so that word variants are consistent. By decreasing the dimensionality of data or feature space and increasing the accuracy of sentiment outputs, this approach is essential to sentiment analysis [17]. Furthermore, lemmatization is helpful when dealing with words that are non-standard or exist as abbreviations [18].

```
nltk.download('wordnet')

from textblob import Word
dataframe['Saran'] = dataframe['Saran'].apply(lambda x: " ".join([Word(word).lemmatize() for word in x.split()]))
dataframe['Saran'].head(10)
```

The technique of lemmatization, which uses WordNet resources from NLTK to break words and convert them into their root forms, is explained in the text..

Labeling

In this work, researchers automatically labeled text data using VADER, a Python sentiment analysis program. The data had to be translated into English first, though, because VADER can only be used with English sentences. Both positive and negative data classes are taken into account in the investigation. As seen in Figure 6, 7839 data were classified as positive and 123 as negative based on the labeling results.

```
for i in range(0,len(text)):
    textB = TextBlob(text[i])
    sentiment = textB.sentiment.polarity
    new_df.at[i,'sentiment']= sentiment
    if sentiment < 0.00 :
        sentimentclass = 'Negative'
        new_df.at[i, 'sentimentclass']= sentimentclass
    else :
        sentimentclass = 'Positive'
        new_df.at[i, 'sentimentclass']= sentimentclass
```

Figure 6. Labelling Process

Two categories are produced by the labeling process: positive and negative. Ninety-eight percent fall into the good category, and two percent fall into the negative. The negative sentiment is expressed through suggestions for improving the library, like expanding the collection of books, improving the

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2217, Copyright ©2024
Submitted : July 16, 2024 , Revised : August 30, 2024, Accepted : September 11, 2024, Published : November 19, 2024

functionality of the website, and resolving fine management.

	sentimentclass	cnt	percent
0	Negative	123	0.015448
1	Positive	7839	0.984552

Figure 7. Sentiment Result

Modelling with Classification Techniques

For all phases of data processing in this study—from preprocessing to data sharing, modeling using classification techniques, testing, and accuracy measurement—we used the Python programming language. When the data is ready, we separate it into training and testing data, identify the data column and target column, and create data vectors for classification in order to get ready for modeling.

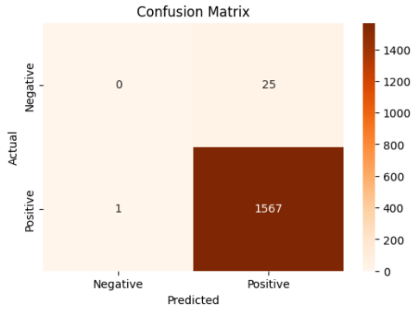
A total of 8253 data texts were gathered for the study and validated using cross-validation. During the investigation, the Confusion Matrix was used with the Random Forest, SVM, and Naïve Bayes algorithms to quantify accuracy. 7962 data points were left for testing using the Naïve Bayes, SVM, and LSTM algorithms after the data was cleaned using Case Folding, Stemming, Tokenizing, and Stopword Removal. Python was the language utilized for all aspects of data processing, including modeling, accuracy measurement, and pre-processing. Below is an explanation of the three algorithms' outcomes.

[13] [26] Naïve Bayes Algorithm Figure 6 illustrates how to apply the Multinomial Naïve Bayes method, which is part of the Naïve Bayes algorithm from the sklearn library, to data modeling in the Python language.

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report, accuracy_score

X = new_df['Saran']
y = new_df['sentimentclass']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

vectorizer = TfidfVectorizer()
X_train_vect = vectorizer.fit_transform(X_train)
X_test_vect = vectorizer.transform(X_test)

nb_classifier = MultinomialNB()
nb_classifier.fit(X_train_vect, y_train)

▾ MultinomialNB
MultinomialNB()
```

Figure 8. Naïve Bayes Algorithm

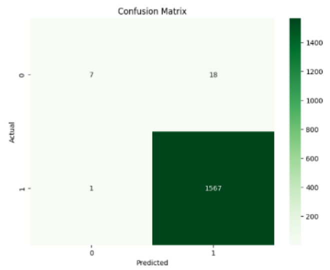
Our accuracy rate with the Naive Bayes Model was 98%. Here are the specifics: There were 1 false negative, 25 real negative, and 1567 true positives.




```
Accuracy: 0.9843063402385436
              precision    recall  f1-score   support

    Negative       0.00      0.00      0.00        25
    Positive       0.98      1.00      0.99      1568

    accuracy                           0.98      1593
   macro avg       0.49      0.50      0.50      1593
weighted avg       0.97      0.98      0.98      1593
```

Figure 9. Accuracy Result using Naïve Bayes Model

[27] SVM Algorithm Linear SVC is the method used for modeling using the SVM technique. The SVM algorithm in the sklearn library includes the Linear SVC technique. Figures 10 and 11 show the outcome of modeling using the support vector machine approach.



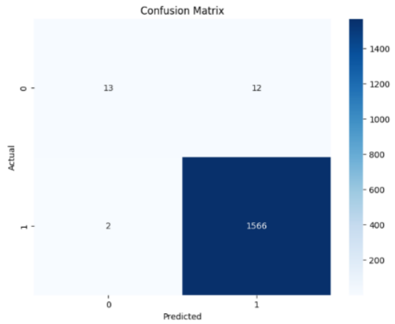

```
              precision    recall  f1-score   support

    Negative       0.88      0.28      0.42        25
    Positive       0.99      1.00      0.99      1568

    accuracy                           0.99      1593
   macro avg       0.93      0.64      0.71      1593
weighted avg       0.99      0.99      0.99      1593
```

Figure 10. Confusion Matrix using SVM Model

Using the SVM model, we were able to achieve 99% accuracy. The following are the specific findings: There were seven false positives, one false negative, 18 genuine negatives, and 1567 real positives. Algorithm for Random Forest: [16] A machine learning method called the Random Forest algorithm creates several decision trees while being trained. The mean prediction for regression of the individual trees or the mode of the classes for classification are the next outputs. Figure 11

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2217, Copyright ©2024
Submitted : July 16, 2024 , Revised : August 30, 2024, Accepted : September 11, 2024, Published : November 19, 2024

provides an illustration of this procedure.




```
              precision    recall  f1-score   support

    Negative       0.87      0.52      0.65        25
    Positive       0.99      1.00      1.00      1568

    accuracy                           0.99      1593
   macro avg       0.93      0.76      0.82      1593
weighted avg       0.99      0.99      0.99      1593
```

Figure 11. Confusion Matrix using Random Forest Model

We achieved 99% accuracy after using the Random Forest model. The following are the specific findings: There were 13 false positives, 2 false negatives, 12 true negatives, and 1566 real positives.

Comparison of Accuracy

The study assessed three approaches using criteria including F-1, recall, accuracy, and precision. It was discovered that the combination of SVM and Random Forest had the greatest accuracy at 99%, with 100.00% precision, recall, and F-1, after examining the output of the Naïve Bayes, Support Vector Machine (SVM), and Random Forest algorithms. With 98% accuracy, 98.00% precision, 100.00% recall, and 99.00% F-1, Naïve Bayes came in second. According to the findings, combining SVM with Random Forest is a great option for sentiment analysis of user reviews of libraries. Although this model is highly accurate at classifying sentiment, preprocessing methods and the quality of the training data affect how well it performs. Table 1 displays the comparison results, while Table 1 and Figure contain the graphs.

Table 1. Test Scenario Comparison Results

	Scenario 1	Scenario 2	Scenario 2
Algorithm	SVM	Random Forest	Naïve Bayes
Accuracy %	99%	99%	98%
Precision %	100%	100%	98%
Recall %	100%	100%	100%
AUC	100%	100%	99%

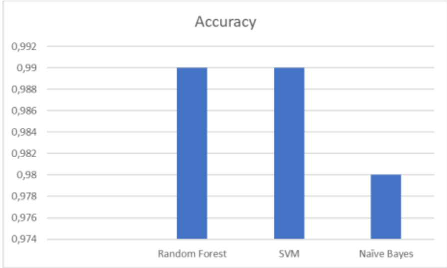


Figure 12. Comparison of Accuracy

Sentiment class word cloud

One tool for analyzing written content that makes use of graphics is the word cloud. In order to evaluate and enhance library resources and services, this study includes input from instructors and students who completed surveys on survei.pcr.ac.id. The word clouds generated from the feedback are displayed in Figures 15 and 16, emphasizing both positive and negative sentiments.

Figure 13 WordCloud Positive and Negative Sentiment

"Good service," "comfortable library," "better collection," and "improve will" are the most frequently found words in the word cloud display of positive sentiment text. The most commonly used words in the text that conveys negative mood are "improve wifi speed," "increase collection," "enhance library facilities," "assignment help needed," "additional references," and "expanded services."

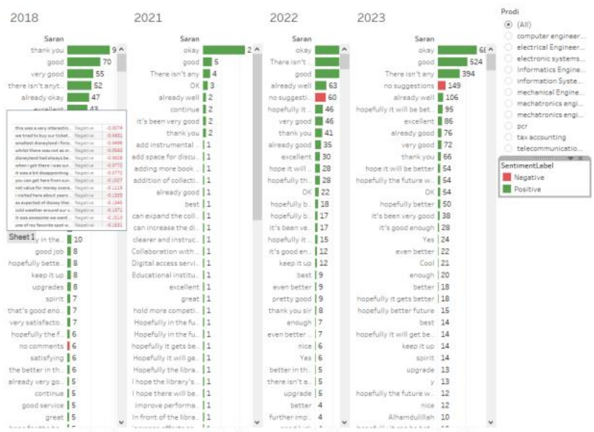


Figure 14 Visualization Tren Sentiment Result

Tableau software was used to visually portray the results

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2217, Copyright ©2024
Submitted : July 16, 2024 , Revised : August 30, 2024, Accepted : September 11, 2024, Published : November 19, 2024

of the sentiment analysis process, enabling the tracking of sentiment trends over time. Furthermore, it was feasible to observe the sentiment outcomes according to both time and study program because the data collection was structured by both year and program. The percentage of good and negative sentiments was constant year after year, according to the trend of this feedback sentiment data. This was due to the fact that the feedback text, which contained suggestions for library enhancements from both academics and students, had never before been thoroughly examined and its results thoroughly examined.

COnclusions

For sentiment analysis of library user input, a comparison of the Naïve Bayes, Support Vector Machine (SVM), and Random Forest algorithms reveals that the combination of SVM and Random Forest is very successful. It performs better in terms of precision, recall, and F1-score and reaches 99% accuracy. The complexity of the model and the properties of the data affect the algorithms' performance. SVM and Random Forest work very well with noisy and high-dimensional text data. Model performance is also influenced by appropriate preprocessing methods including vectorization, stemming/lemmatization, tokenization, and stop word removal. These findings offer a solid basis for creating improved sentiment analysis algorithms for libraries and create avenues for additional study in this field.

References

[1] S. Saifullah, Y. Fauziyah, and A. S. Aribowo, "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," *J. Inform.*, vol. 15, no. 1, p. 45, 2021, doi: 10.26555/jifo.v15i1.a20111.

[2] A. E. P. Nugraha, R. Riyanto, I. A. Sari, and D. P. Hadi, "Analisis Sentimen Konsumen Selama Pandemi Covid-19 di Kota Semarang Menggunakan Tableau," *J. Nusant. Apl. Manaj. Bisnis*, vol. 7, no. 2, pp. 185–193, 2022, doi: 10.29407/nusamba.v7i2.16068.

[3] D. Nurmalasari, H. Yuliantoro, and S. I. Yanti, "Dini Nurmalasari, Heri Yuliantoro, Saleha Indra Yanti," pp. 96–105.

[4] F. Aziz, A. R. Thaha, and N. A. Ma'ruf, "Analisis Sentimen Destinasi Wisata Geopark Ciletuh," *J. Ilm. Pariwisata*, vol. 27, no. 1, p. 60, 2022, doi: 10.30647/jip.v27i1.1469.

[5] N. GUPTA, P., TIWARI, R. & ROBERT, "Sentiment analysis and text summarization of online reviews: A survey," in *International Conference on Communication and Signal Processing, ICCSP*, 2016, pp. 241–245.

[6] F. V. S. and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *J. SIMETRIS*, vol. 10, no. 2, pp. 681–686, 2019.

[7] R. H. Bater Makhabel, Pradeepta Mishra, Nathan Danneman, *R: Mining spatial, text, web, and social media data*.

[8] S. Elbagir and J. Yang, "Language Toolkit and VADER Sentiment," in *MultiConference Eng. Computer. Sci*, 2019, pp. 12–16.

[9] C. J. and G. E. Hutto, "VADER: A Parsimonious Rule-based Model fo," in *Eighth Int. AAAI Conf. Weblogs Soc. Media*, 2014, p. 18.

[10] J. H. and E. Gilbert, "VADER: A parsimonious rulebased model for sentiment analysis of social media text," in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM*, 2014, pp. 216–225.

[11] Sebastian Raschka dan Vahid Mirjalili, *Python Machine Learning*. 2019.

[12] A. H. Al Kabir, S. Basuki, and G. W. Wicaksono, "Analisis sentimen kritik dan saran pelatihan aplikasi teknologi informasi (PATI) menggunakan algoritma support vector machine (SVM)," *J. Repos.*, vol. 1, no. 1, p. 39, 2019, doi: 10.22219/repositor.v1i1.11.

[13] D. N. MI Zul, F Yulia, "Social media sentiment analysis using K-means and naïve bayes algorithm," in *2nd International Conference on Electrical Engineering and Informatics (ICon EEI)*, 2019.

[14] & R. A. Imam Riadi, Rusydi Umar, "Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode Naïve Bayes," *Decode*, vol. 4, no. 1, pp. 191–203, 2024.

[15] A. K. Eko, E. R. D., Maharani, D., & Syahputra, "Pemanfaatan Metode Naive Bayes Untuk Klasifikasi Status Gizi Balita Pada Kelurahan Karang Anyer," *Decode*, vol. 4, no. 2, pp. 392–405, 2024.

[16] I. Afdhal, R. Kurniawan, I. Iskandar, R. Salambue, E. Budianita, and F. Syafria, "Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 1, pp. 122–130, 2022, [Online]. Available: http://ojs.serambimekkah.ac.id/jnkti/article/view/4004/pdf

[17] & G. T. Cesar, W., Riki Ramdani Saputra, "Perancangan Model Sistem Pendukung Keputusan Untuk Menentukan Formasi CASN Menggunakan Naïve Bayes dan Simple Additive Weighting," *Decoding*, vol. 4, no. 1, pp. 239–250, 2024.

[18] and D. A. K. Indrayuni, A. Nurhadi, "Implementasi Algoritma Naive Bayes, Support Vector Machine, dan K_Nearest Neighbors untuk Analisa Sentimen Aplikasi Halodoc," *Fakt. Exacta*, vol. 14, no. 2, p. 64, 2021.

[19] Y. Ping, Y. Zhou, C. Xue, and Y. Yang, "Efficient representation of text with multiple perspectives," *J. China Univ. Posts Telecommun.*, vol. 19, no. 1, pp. 101–111, Feb. 2012, doi: 10.1016/S1005-8885(11)60234-3.

[20] Z. Drus and H. Khalid, "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review," *Procedia Comput. Sci.*, vol. 161, pp. 707–714, 2019, doi: https://doi.org/10.1016/j.procs.2019.11.174.

[21] S. J and K. U, "Sentiment analysis of amazon user reviews using a hybrid approach," *Meas. Sensors*, vol. 27, p. 100790, 2023, doi: https://doi.org/10.1016/j.measen.2023.100790.

[22] V. Kumar, "Exploring the Use of Sentiment Analysis in Library User Studies: Approaches and Challenges," 2023, pp. 446–457.

[23] S. M. Fani, R. Santoso, and S. Suparti, "Penerapan Text Mining Untuk Melakukan Clustering Data Tweet Akun Blibli Pada Media Sosial Twitter Menggunakan K-Means Clustering," *J. Gaussian*, vol. 10, no. 4, pp. 583–593, 2021, doi: 10.14710/j.gauss.v10i4.30409.

[24] Mutammimah, H. Sujaini, and R. D. Nyoto, "Analisis Perbandingan Metode Spelling Corrector Peter Norvig dan Spelling Checker BK-Trees pada Kata Berbahasa Indonesia," *J. Sist. dan Teknol. Inf.*, vol. 5, no. 1, pp. 12–16, 2017.

[25] E. Shafiera, "Pengaruh penerapan spelling correction menggunakan metode symspell pada incident categorization," pp. 1–98, 2022, [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/65196

[26] D. Nurmalasari and H. Ribut Yuliantoro, "Implementasi Ekstraksi Fitur untuk Pengelompokan Dokumen Proposal Menggunakan Algoritma NaÃ¯ve Bayes," *J. Komput. Terap.*, vol. 8, no. 1, pp. 194–203, 2022, doi: 10.35143/jkt.v8i1.5351.

[27] M. A. N. Y. Uan, Y. U. A. N. X. I. N. O. Uyang, and Z. H. X. Iong, "A Text Categorization Method using Extended Vector Space Model by Frequent Term Sets *," vol. 114, no. 60972145, pp. 99–114, 2013.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2217, Copyright ©2024
Submitted : July 16, 2024 , Revised : August 30, 2024, Accepted : September 11, 2024, Published : November 19, 2024