Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.11 No.1, January 2026, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Sentiment Analysis and Emotional Reviews of Hospital Services Using Naïve Bayes and Support Vector Machine (SVM)

Muhammad Rio Lintang Cahya[1], Erwin Yudi Hidayat[2]

[1,2]*Informatics Engineering Department, Universitas Dian Nuswantoro, Semarang, Indonesia*

[1]riolintang03@gmail.com (*)

[2]erwin@dsn.dinus.ac.id

Received: 2025-12-01; Accepted: 2026-01-31; Published: 2026-01-03

Abstract— Hospitals are crucial healthcare institutions that play a vital role in providing medical services to the community. Patient perceptions and experiences regarding hospital services are often reflected in reviews left on digital platforms such as Google Maps. This study aims to analyze public sentiment and emotions toward hospital services in Semarang City using Google Maps user reviews. A total of 16,364 reviews from 21 hospitals were collected between 2023 and 2024. Sentiment labelling was performed using a lexicon-based approach to classify reviews into positive, negative, and neutral categories. To explore the emotions expressed in the reviews, the NRC Emotion Lexicon (EmoLex) was used to identify eight basic emotions. The analysis revealed that 'Trust' was the most dominant emotion (7,090 words), indicating high patient confidence, followed by 'Joy'. Furthermore, for predictive modelling using Naïve Bayes and Support Vector Machine (SVM) algorithms, SVM achieved 90% accuracy, whereas Naïve Bayes achieved only 78%. The results of this analysis are expected to serve as input for hospitals to improve service quality and as a reference for prospective patients in selecting a hospital.

Keywords— Sentiment Analysis; Emotion Review; SVM; Naïve Bayes; Hospital Service.

I. INTRODUCTION

A hospital is a health care institution that provides a range of health-related services to the community. These services include medical and nursing care, providing diagnosis, treatment, and surgery. Hospitals also provide various facilities needed by patients, such as inpatient rooms, operating rooms, pharmacies and Emergency Units [1]. Hospitals in Indonesia are divided into several classes based on the facilities and services they provide. Class A public hospitals have at least 250 beds; class B public hospitals have at least 200 beds; class C public hospitals have at least 100 beds; class D public hospitals have at least 50 beds [2]. The higher the hospital's class, the more comprehensive the facilities provided to patients. Therefore, patients are often referred to higher-level hospitals when treatment at the original hospital is unsatisfactory or when the hospital's facilities are insufficient to treat the patient's condition.

Patients who have been treated at the hospital can share experiences regarding the services they received. An important point of the hospital accreditation process is the public opinion of the services provided by the hospital [3]. Conventionally, the quality of hospital services is assessed through structured surveys. However, conventional surveys have limitations in capturing patients' spontaneous perceptions as expressed on digital platforms. Unlike surveys, which tend to be periodic and rigid, online reviews offer real-time, honest feedback that reflects patients' genuine emotions. One platform for providing feedback or reviews is Google Maps. Google Maps reviews capture a person's experience and express sentiments and emotions related to the satisfaction of other patients, making them a valuable source of information for choosing the right hospital. Prospective patients can read reviews to find out the advantages and disadvantages of the hospital based on patient experience. Patient satisfaction is defined as the level of feeling that arises from the performance of health services, as measured by comparing what the patient receives with what he expects [4].

However, not all reviews on Google Maps are objective. With the freedom to write both positive and negative reviews, these reviews can influence potential patients' perceptions of hospital services. Therefore, sentiment and emotion analysis are needed to identify and understand the meaning of these reviews [5]. Sentiment analysis is the analysis of natural language or word composition used to track public opinion about an event. Sentiment analysis focuses on the emotional perspective of text comments, aiming to automatically analysehe mood of the community, emotions, and the atmosphere of the day regarding a particular issue [6]. Sentiment and emotion analysis helps reveal patterns, trends, and public perceptions in greater depth, which may not be immediately apparent [7]. Sentiment analysis, or opinion mining, is a field of study that analyses people's opinions about entities such as products, services, organizations, individuals, issues, events, and topics [8]. Sentiment analysis is used to determine a person's opinion toward an event or issue, whether it is positive, negative, or neutral [9]. By understanding the positive and negative sentiments in reviews, as well as the emotions such as happiness, anger, and sadness they contain, hospitals can take more appropriate steps to improve the quality of their services [10]. In addition, the analysis results can serve as a reference for prospective patients in making more informed decisions about the hospital they choose.

Previous research [11] used SVM and Random Forest models and found that Random Forest was more stable and superior in accuracy for Traveloka features, achieving 71%.

DOI : https://doi.org/10.25139/inform.v11i1.11257

The research [12] used the Naïve Bayes Classifier with K-10, achieving an accuracy of 73%. Research [13] compared the NBC and M models, with the best results obtained with the SVM model, achieving 100% accuracy after applying SMOTE and Stemming. Research [14] using the KDD (Knowledge Discovery in Data) method produced an accuracy rate of 87%. Research [15] used SVM and Random Forest models and found that SVM was superior, achievin85.74% accuracy with SMOTE. Eesearch [16] uses an SVM with the BOW method, achieving an accuracy of 86%. Research [17] used an SVM model and achieved an accuracy 97%.

This research used Google Maps reviews of 21 hospitals in the city of Semarang, totalling 16,364 reviews, covering the period 2023–2024. Sentiment labelling used a lexicon-based dictionary to obtain positive, negative, and neutral sentiment scores for each review. However, the sentiment results in the form of positive, negative, and neutral did not fully describe the emotions in the reviews. Sentiment is not limited to positive and negative classes; without containing sentiment (neutral), sentiment can be described as a feeling [18]. Based on this review, this study proposed the Support Vector Machine (SVM) method. SVM was chosen as the main method because of its robust and efficient handling of high-dimensional sparse data, which is a common typology of Google Maps reviews. Although many studies have compared classification algorithms, there is still a gap: previous studies have focused only on sentiment polarity (positive/negative) and have not conducted an in-depth exploration of granular emotions (such as Trust and Fear). Therefore, this study aims to fill this gap by integrating the NRC Emotion Lexicon to identify emotions in reviews. EmoLex contains a list of words and their associations with eight basic emotions, namely anger, anticipation, fear, surprise, sadness, joy, trust, and disgust, as well as two sentiments, namely negative and positive [7].

II. RESEARCH METHODOLOGY

This research aims to comprehensively reveal public perceptions and sentiments expressed in patient reviews across various aspects of hospital healthcare services in the city of Semarang. The research focuses not only on polarity (positive or negative sentiment) but also on granular emotion analysis to detect nuances in patients' feelings, such as whether the reviews reflect happiness with the service, sadness, fear, or other emotions.

The output of this analysis is projected to have dual value: first, as a data-based evaluation instrument for hospital management to identify weak points in their services; and second, as a transparent reference source for the community or prospective patients seeking the best healthcare options in Semarang. To achieve these objectives, this study began with the acquisition of review data, followed by a pre-processing stage for text standardization, sentiment and emotion labelling, and text transformation into numerical vectors using the TF-IDF algorithm, and finally a modelling and performance evaluation stage to ensure accurate predictions during model inference. The detailed workflow for this research is shown in Fig.1.

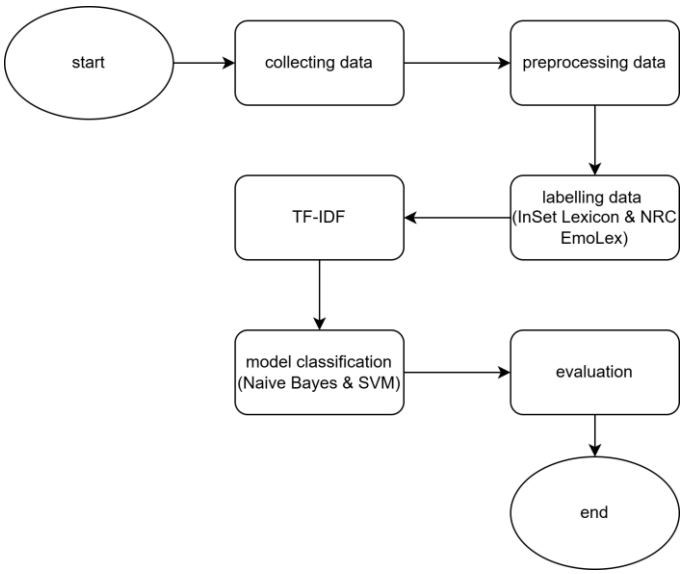

Fig.1. Research Flow

A. *Collecting Data*

Data collection in this research used scraping tools, namely Instant Data Scraper. The collected data consisted of user reviews on Google Maps for hospitals in the Semarang City area. The hospitals included were Amino Hospital, Banyumanik Hospital, Bhayangkara Hospital, Columbia Hospital, Elisabeth Hospital, Hermina Hospital, Karyadi Hospital, Diponegoro Hospital, Pantiwilasa Hospital, Permata Medika Hospital, Primaya Hospital, Roemani Hospital, Siloam Hospital, Sultan Agung Hospital, Tamtama Hospital, Telogorejo Hospital, Tugurejo Hospital, William Booth Hospital, and Wongsonegoro Hospital. After successful scraping, the data will be saved in XLSX format for further processing in the next stage. To protect privacy and research ethics, all user identities (account names) have been anonymized prior to the analysis stage. This dataset consists solely of public reviews that include positive, negative, and neutral opinions, with various informal language styles, about a hospital's services.

B. *Pre-processing Data*

Because Google Maps review data includes a variety of informal language styles, it is necessary to clean the review text first so that, when it enters the modelling stage, it will not introduce bias or noise that could affect the model. Therefore, the data cleaning stages that will be carried out include:

1) Drop Duplicates and Empty Data: This is done to prevent problems during the machine learning model-building process.

2) Data Cleaning: The aim is to eliminate noise elements or characters that are irrelevant to the analysis's context. This process includes removing numeric characters, punctuation marks, special symbols, and emoticons. In addition, this stage involves trimming or removing excess whitespace at the beginning and end, and between words, to ensure the integrity of the sentence structure before further processing [19].

DOI : https://doi.org/10.25139/inform.v11i1.11257

3) Casefolding: the process of converting or standardizing all characters in a text document into the same format, generally lowercase letters [20].

4) Tokenization: The process of decomposing or breaking down input sentences into smaller linguistic units. This process involves separating strings based on specific delimiters (such as spaces or special characters) to map the syntactic structure of a text document before further analysis [21].

5) Normalization: The procedure of standardizing non-standard word forms into standard forms in accordance with applicable linguistic rules (e.g., *KBBI* for Indonesian). The final goal is to reduce feature dimensions by unifying various spelling variations that refer to the same semantic entity.

6) Lemmatization: that aims to transform words into their basic or dictionary form (lemma) by considering the context in which the word is used in a sentence.

7) Stopword Removal: Filtration techniques to eliminate words that appear frequently but have low semantic value. Removing stopwords aims to reduce the dataset size and focus the algorithm's performance on keywords that contain substantive information from the document [21].

C. Labeling Data

In this research, for initial labelling, a lexicon dictionary containing words with values ranging from -5 (negative) to 5 (positive) was used. For the final result, if the score of a sentence is >= 1, it will be categorized as positive, if the score is <= -1, it will be categorized as negative, and anything else will be categorized as neutral [22]. The positive and negative dictionary list in Table I.

TABLE I
POSITIVE DICTIONARY DICTIONARY LIST

Word	**Weight**	**Word**	**Weight**
Positive Dictionary:		Negative Dictionary:	
Hai	3	*Lara*	-5
Detail	2	*Mencelakai*	-5
Antusias	2	*Maaf*	-3
Senyum	2	*Tolong*	-2
Segera	3	*Satir*	-2
. . .	. .	. .	. .
Sholat	5	*Sibuk*	-3

TABLE II
NRC EMOLEX DIARY

Word	
Benci	
EmoLex Score	
Anger	1
Anticipation	0
Disgust	1
Fear	1
Joy	0
Sadness	0
Surpirse	0
Trust	0

Then, for emotion labelling, NRC EmoLex was used, which contains a set of words with scores formotions such as joy, happiness, fear, and so on [23]. This is shown in Table II. It is important to recognise that this lexicon-based tagging serves as a silver standard rather than a gold-standard, fully annotated by humans. Although this approach allows efficient processing of large datasets, automatic tagging may lack the contextual nuances of human tagging.

D. TF-IDF

The TF-IDF method converts text data into numerical or vector form for NLP processing [24]. This method is used to calculate a word's weight and assess its importance within a collection of documents (a corpus). TF-IDF consists of two components, namely*:*

1) Term Frequency (TF): Measures how often a word appears in a document. The more often the word appears, the greater its value. As shown in Equation (1), the value of $TF(t,d)$ represents the frequency of occurrence of a word (term) relative to the length of the document. Where $f_{t,d}$ is the number of occurrences of term t in document d, while N_d is the total number of words in document d. Division by N_d is performed as a normalization step to prevent bias in documents with more words.

2) Inverse Document Frequency (IDF): Measuring how unique a word is in the entire corpus. Words that appear frequently across many documents will have a low value because they are considered common, as shown in Equation (2). Where N represents the total number of documents in the data corpus, and df_t (document frequency) is the number of documents containing the word t. The logarithm function is used to mitigate the impact of very large values of N, so that the weight scale remains proportional. The final weight Equations (3) for each feature are obtained by multiplying the results of Equations (1) and (2).

$$IDF(t) = \log(\frac{N}{df_t}) \quad (2)$$

$$W_{t,d} = TF(t,d)\ x\ IDF(t) \quad (3)$$

E. IndoBERT Embedding

IndoBERT is a language model based on the Transformer (BERT) architecture that has been pre-trained on a large-scale Indonesian-language corpus, covering more than 4 billion words from sources such as news, social media, and Wikipedia [25]. In this mechanism, each word is represented as a high-dimensional dense vector (generally 768 dimensions). The main advantage of this embedding is its ability to capture semantic meaning from the surrounding sentence context in a bidirectional manner. This capability makes IndoBERT very effective at handling the ambiguity and non-standard language structures often found in user reviews.

F. Modelling

This research uses the Naïve Bayes and SVM baseline models. Previous researchers have widely used both models for classification problems and have demonstrated that they achieve optimal accuracy [6].

DOI : https://doi.org/10.25139/inform.v11i1.11257

1) Naïve Bayes: The Naïve Bayes Classifier is a classification method rooted in Bayes' theorem. As shown in Equation (4), this method uses probability and statistics to predict future probabilities based on experience [9]. Where $P(A|B)$ is the posterior probability of class A (e.g., positive/negative) given feature B (word). The variable $P(A)$ is the prior probability of class A, while $P(B|A)$ is the probability of feature B occurring in class A. The value of $P(B)$ is the probability of the feature occurring in general and is constant across classes.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (4)$$

2) Support Vector Machine (*SVM*): SVM is a learning algorithm based on Structural Risk Minimization that learns to separate data into different classes by constructing a hyperplane [11]. To determine the sentiment class of each document, the SVM model uses a linear decision function shown in Equation (5). In this Equation, $w \,.\, x$ is the dot product between the model weight w and the document feature x, which is then added to the bias b [27]. Classification is based on the sign of the calculation result: if $w.x + b > 0$, the document is classified into the positive class, and conversely, if the result is negative, it is classified into the negative class [28].

$$w \,.\, x + b = 0 \quad (5)$$

G. Evaluation

The model results will be evaluated using a confusion matrix. Several matrices will be generated, namely accuracy, precision, recall, and F1-score for each class. The results of the predictions can then be sorted into four groups, namely TP (positive data that is predicted correctly), TN (negative data that is predicted correctly), FP (negative data that is predicted as positive data), and FN (positive data that is predicted as negative data)[29].

Accuracy describes how accurately the model can classify correctly, as shown in Equation (5). Precision describes the level of agreement between the requested data and the model's predictions, as shown in Equation (6). Recall describes the model's success in rediscovering information, as shown in Equation (7), and the F1-Score is a harmonic combination of precision and recall, as shown in Equation (8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = \frac{2 * (Precision + Recall)}{Precision + Recall} \quad (8)$$

III. Result and Discussion

This section compares classification results from SVM and Naïve Bayes modelling. To achieve a more accurate comparison, this modelling will be conducted in two sessions: the first and second scenarios, each with different model parameters.

A. Collecting Data

The total number of reviews obtained during scraping was 16,364, spanning 2023 to 2024. Ratings were not used in this study because, in real-life cases, users may give low ratings but write positive reviews. After successful scraping, the data will be saved in XLSX format for continued processing in the next stage. The scraping results are shown in Table III.

TABLE III
SAMPLE DATASET

Hospitals	Reviews
RS Dr. Amino Gondohutomo	*Dokter untu og ngamukan*
RS Hermina Pandanaran	*Istimewa pelayanan kesehatan, gedung memadai mewah bersih bagus. Suka banget periksa ke sini. RS favorit keluarga ku …*
....	....
RSUP Dr. Kariadi	*Suasana nyaman untuk pasien dan keluarga yg menunggu. Semoga tambah bagus pelayanannya..*

B. Pre-processing Data

The next step is the data pre-processing process for the dataset obtained. This step consists of dropping duplicates, casefolding, tokenization, normalization, lemmatization, and stopword removal, which aim to ensure the dataset is clean during the modelling stage and to achieve more accurate results [30]. The pre-processing results are shown in Table IV.

TABLE IV
TEXT PRE-PROCESSING

Raw Text	
Setelah saya berkonsultasi dengan yg lebih profesional, saya sangat terbantu. Saya mendapat pemahaman yang lebih baik 👍 …	
Text Pre-processing	
Case folding	*setelah saya berkonsultasi dengan yg lebih profesional saya sangat terbantu saya mendapat pemahaman yang lebih baik*
Tokenization	*[setelah, saya, berkonsultasi, dengan, yg, lebih, profesional, saya, sangat, terbantu, saya, mendapat, pemahaman, yang, lebih, baik]*
Normalization	*[setelah, saya, berkonsultasi, dengan, yang, lebih, profesional, saya, sangat, terbantu, saya, mendapat, pemahaman, yang, lebih, baik]*
Lemmatization	*setelah saya konsultasi dengan yang lebih profesional saya sangat bantu saya dapat paham yang lebih baik*
Stop word removal	*konsultasi profesional bantu paham*

C. Labelling Data

To address the limitations of manually annotating the entire dataset and to test the reliability of automatic labelling, manual validation was performed using Cohen's Kappa. Human annotators manually label a random sample of 390 reviews. A comparison between the system labels (Lexicon) and the manual labels showed a 70% agreement rate. This level of agreement is considered acceptable for generating initial ground truth before entering the classification model's training stage.

DOI : https://doi.org/10.25139/inform.v11i1.11257

1) Labelling Sentiment: Based on automatic labelling of 16,364 reviews, the public generally has a positive perception of hospital services in Semarang. As shown in Fig.2, the sentiment distribution is dominated by the Positive class at 63.1% (10,311 reviews). This dominance indicates that key service aspects, such as the competence of medical personnel and hospital facilities, have met the expectations of the majority of patients. Furthermore, Neutral sentiment accounts for 19.8% (3,231 reviews), which generally contains objective questions or informative statements without strong emotional content. On the other hand, although a minority, Negative sentiment at 17.2% (2,804 reviews) is a crucial indicator that highlights areas for improvement, particularly related to operational efficiency and queue management. The imbalance in the amount of data between these classes (an imbalanced dataset) poses a technical challenge that will be tested in the next classification modelling stage.

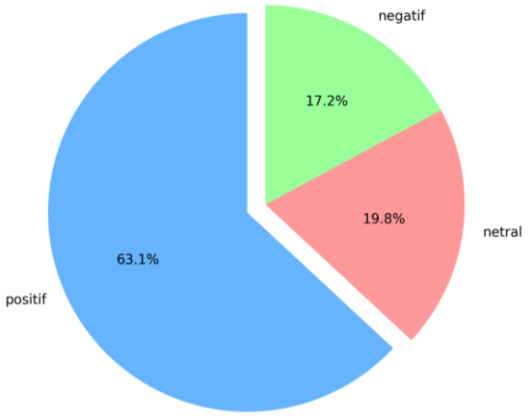

Fig.2. Percentage Sentiment

To provide a more concrete picture of the characteristics of each sentiment class, Table V presents a representation of the original reviews (raw text) resulting from the labelling. In the Negative sentiment category, users tend to express frustration with operational and procedural failures, as seen in reviews about "*antrian gak jelas*" and experiences of being "*dilewati*." The use of repetitive phrases such as "minus double minus triple minus" underscores patients' intense disappointment with service management. In contrast, Positive sentiment is dominated by direct appreciation of medical personnel's soft skills and treatment outcomes. Emotional keywords such as *"terimakasih"*, *"memuaskan"*, and *"dokter ramah"* are key indicators of patient satisfaction.

Meanwhile, the Neutral category often contains rhetorical questions or conditional feedback. For example, complaints about *"bau tidak sedap"* in the parking area are classified as Neutral by the lexicon system because the sentence structure is in the form of a question (*"apakah tidak ada..."*) and a suggestion, which, in terms of word weight, is not as aggressive as purely negative reviews. This demonstrates the characteristic of the lexicon method, which is based on the accumulation of word scores rather than the context of implicit sarcasm.

TABLE V
DATA LABELLING SENTIMENT

Raw Text	Sentiment
antrian gak jelas, antri duluan dilewati. giliran protes baru disuruh masuk. minus double minus triple minus	Negative
apakah tidak ada yang mencium bau semerbak setiap kali masuk / keluar dan berhenti di portal parkir? sudahlah koordinasi dgn pihak terkait? bau tidak sedap hal sepele tapi kalo diabaikan...jadi sumber ketidaknyamanan dan mana mau orang â€¦	Neutral
terimakasih bangsal ripd atas pelayananya yg sangat memuaskan.	Positive
saya pernah kunjungan dirumahsakit gondohutomo ini pelayanannya baik baik dokter yang ramah	Positive

2) Labelling Emotion: Unlike polar sentiment analysis (positive-negative), emotion analysis using NRC EmoLex provides a more granular picture of patients' feelings [31]. The emotion extraction results show that "Trust" is the most dominant emotion with a frequency of 7,090 words, followed by "Joy" with 3,190 words. The complete breakdown of emotion labels is shown in Table VI.

TABLE VI
EMOTION COLLECTION

Dominant Emotion	Total
Trust	7.090
Joy	3.190
Anticipation	2.418
Anger	250
Sadness	177
Disgust	177
Fear	164
Surprise	76
Disgust	177

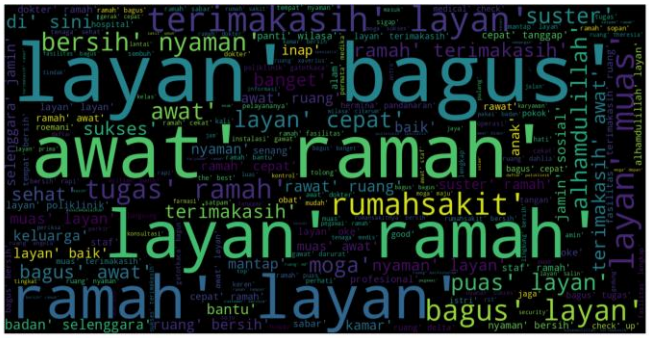

Fig.3. WordCloud Sentiment Positive

Based on Fig.3 (Positive Sentiment Word Cloud), the data visualisation shows the dominance of keywords such as *"ramah"*, *"nyaman"*, *"bersih"*, *"terimakasih"*, and *"puas"*. The high frequency of the words *"ramah"* and *"terimakasih"* confirms the finding that patient satisfaction is not driven solely by medical success but also depends heavily on the quality of interpersonal interactions between patients and medical personnel (doctors and nurses). In addition, the emergence of the words "clean" and "comfortable" indicates that the physical and environmental aspects of hospitals are vital to creating a positive psychological experience for patients and their families.

DOI : https://doi.org/10.25139/inform.v11i1.11257

Then, in Fig.4 (Negative Sentiment Word Cloud), the narrative of the reviews shifted drastically toward technical and operational aspects. The most prominent words *include "gawat darurat", "instalasi", "daftar", "antre", "tunggu",* and *"jam"*. The size of the words *'gawat darurat'* and "*instalasi*" strongly signals that the Emergency Room (ER) is the service point most likely to trigger dissatisfaction. Furthermore, the dominance of the words *"antre", "tunggu", and "daftar"* confirms that bureaucratic inefficiency and long waiting times are the main drivers of negative sentiment, underscoring the need to improve patient flow management.

Fig.4. WordCloud Sentiment Negative

Then, as seen in Fig.5 (Neutral Sentiment Word Cloud), there is a mixture of ambiguous words. Positive words such as *"layan", "bagus",* and *"ramah"* still appear, but in descriptive, informative, or conditional contexts (e.g., *"apakah dokternya ramah?"* or *"tempatnya bagus tapi jauh"*). This demonstrates the characteristics of the lexicon method, which works based on word matching; when emotionally charged words appear in question or objective statement structures without sufficient intensity, the algorithm tends to group them into the neutral class.

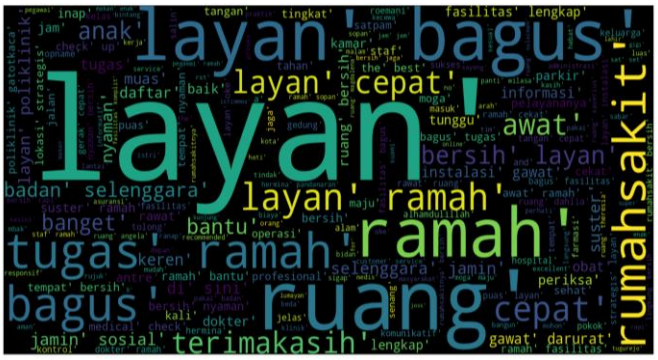

Fig.5. WordCloud Sentiment Neutral

D. Weighting TF-IDF

The next step is to split the data into training and test sets at an 80:20 ratio. Before modelling, the TF-IDF step is required to extract features from each word by converting text data into numerical vectors. This process weights each term by its frequency of occurrence in the document (TF) and its scarcity across the entire corpus (IDF). Table VII shows the 20 features with the highest total TF-IDF weights. The results show that the feature "*layan"* has the highest value. This indicates that the most frequently discussed aspect is respondents' experience with the service they received during their hospital visit.

TABLE VII
TOP FEATURE TF-IDF

Feature	Total Weight TF-IDF
Layan	1175.32
Ramah	709.41
Bagus	678.78
Layan bagus	404.44
Awat	401.08
Ruang	381.60
Terimakasih	361.67
Bersih	333.78
Muas	333.62
Layan ramah	307.70
Cepat	290.16
Nyaman	282.71
Tugas	220.91
Awat ramah	202.74
Rumahsakit	199.28
Layan muas	188.54
Fasilitas	166.52
Tugas ramah	152.76
Baik	146.68
Layan cepat	146.22

E. Modelling

At this stage, the classification model was developed by comparing the performance of two machine learning algorithms, Naïve Bayes and Support Vector Machine (SVM). To comprehensively assess the model's effectiveness, the experiment was divided into two main testing scenarios. The first scenario was a baseline test, in which both algorithms were run using standard parameters (default parameters) without any modifications. The goal is to assess the model's basic classification ability before optimization. The second scenario focuses on improving model performance through hyperparameter tuning. In this stage, the GridSearchCV technique is applied to automatically find the best parameter combination using a cross-validation scheme. With GridSearchCV, the model is evaluated against various candidate parameter values to produce the optimal model. Several parameters were tested using GridSearchCV for both algorithms, as shown in Table VIII.

TABLE VIII
PARAMETER GRID SEARCHCV

Models	Parameter	Param Grid	Best parameter
NB	Alpha	[0.1, 0.5, 1.0, 5.0, 10.0]	5.0
	Fit_prior	[True, False]	False
SVM	C (Regulation)	[0.01, 0.1, 1, 10, 100]	10
	Kernel	[‘linear’, ‘rbf’]	‘linear’

F. Evaluation

Based on the model experiment results in Table IX, the SVM algorithm consistently outperforms Naïve Bayes in the baseline model. Then, in the comparative experiment results in Table X, the SVM algorithm with TF-IDF features proved to be the best model, achieving 90% accuracy and an F1-Score of 89%. This achievement is 12% higher than the Naïve Bayes tuning model, which achieved only 78% accuracy.

DOI : https://doi.org/10.25139/inform.v11i1.11257

TABLE IX
EVALUATION BASELINE MODELS

Models	Accuracy	Precicion	Recall	F1-Score
NB	75%	71%	75%	70%
SVM	88%	88%	88%	88%

TABLE X
EVALUATION HYPERPARAMETER TUNING MODELS

Models (Tunning parameter)	Feature Extraction	Accuracy	Precision	Recall	F1-Score
NB	TF – IDF	78%	77%	78%	76%
SVM	TF – IDF	90%	89%	90%	89%
	BERT Embedding	78%	77%	78%	78%

The application of SVM with BERT Embedding achieved an accuracy of only 78%, comparable to Naïve Bayes. Although BERT is a sophisticated Deep Learning architecture, its use as a static feature extractor without fine-tuning was not sufficient to capture the nuances of specific and informal hospital reviews.

These results indicate that, for medium-sized datasets with non-standard language characteristics (noisy text), the traditional TF-IDF weighting method is more effective at highlighting distinctive keywords (such as *'antre', 'ramah'*) than the general semantic vector representations of pre-trained models. Therefore, the combination of SVM + TF-IDF is proposed as the method in this study due to its balance between computational efficiency and superior prediction accuracy.

Based on the Confusion Matrix in Fig. 6, the SVM model with Hyperparameter Tuning shows very robust performance, especially in recognizing the majority class. The model correctly predicted 1,975 Positive reviews (True Positives), with an error rate of only 21 Positive reviews incorrectly predicted as Negative. This indicates that positive keyword features (such as *'ramah', 'bersih', 'bagus'*) have very distinctive vector patterns that are easily separated by the SVM hyperplane. However, the 28 Negative reviews that were incorrectly predicted as Positive may have been due to linguistic complexities challenging to capture with word-frequency-based methods (TF-IDF), such as sarcasm, in which users often use positive words to express disappointment. For example, *"hebat ya, antre obat saja sampai 3 jam"*. The SVM algorithm assigns a high weight to the word "Hebat" as a positive feature, thereby failing to detect its implied negative meaning. Then there are ambiguous reviews, reviews that contain mixed sentiment, such as "*Dokternya sangat ramah, tapi sistem pendaftarannya berantakan*." If the weight of the word '*ramah*' in the training corpus is more dominant than "*berantakan*," the model tends to pull these reviews into the positive area. In addition, the highest error rate occurs in the Neutral class, where many Neutral reviews are misclassified as Positive or Negative. This is natural because the Neutral class often has high vocabulary overlap with the other two classes, usually in the form of descriptive questions or statements about the hospital's physical condition that do not express strong subjective emotions.

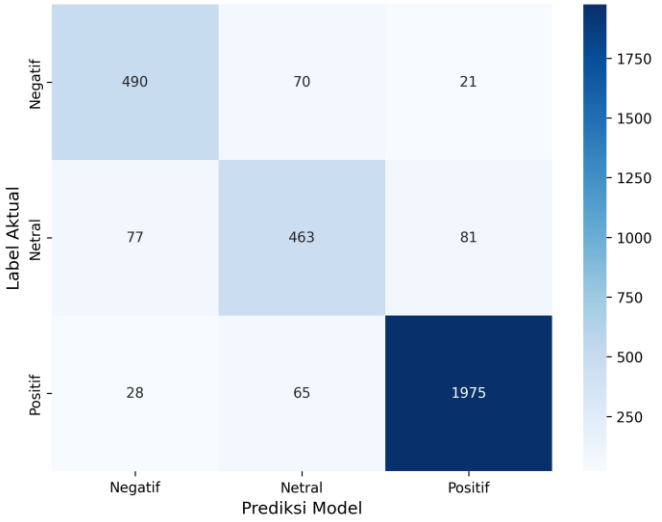

Fig.6. Confusion Matrix of SVM Tunning Parameter with TF – IDF

Based on Table XI, the advanced matrix evaluation of SVM results (tuning parameter) with TF-IDF showed that the Weighted Average F1-Score of 0.89, which is close to the total accuracy (90%), indicates that the model is very stable in handling the original data distribution. Additionally, the Macro Average F1-Score of 0.85 indicates that the model generalizes well across classes. Despite a significant imbalance in the data (2,068 positive vs. 581 negative), the model still maintains an F1-Score of 0.83 for the Negative class. The minor difference between the Weighted Average and Macro Average indicates that the proposed SVM model had no excessive bias towards the majority class. This advantage is often difficult to achieve with standard classification methods on imbalanced datasets.

TABLE XI
EVALUATION MATRIX OF SVM TUNING PARAMETER WITH TF - IDF

Metric	Precision	Recall	F1-Score	Data
Class :				
Negative	82%	84%	83%	581
Netral	77%	75%	76%	621
Positive	95%	96%	95%	2068
Average:				
Macro	85%	85%	85%	3270
Weighted	89%	90%	89%	3270
Total Accuracy:	90%			

In addition, even though the data is imbalanced, with the positive class (63.1%) dominating the Negative class (17.2%), the evaluation results show that the model can still achieve a high F1-Score for the minority class. This can be explained by the characteristics of the SVM algorithm and TF-IDF feature extraction. First, negative reviews of hospital services tend to use very specific vocabulary with strong emotional weight, such as '*antre*', '*lama*', '*kecewa*', and '*kotor*'. This differs significantly from the more general positive vocabulary. The TF-IDF weighting method assigns high values to these distinctive words, making the negative-class features highly separable in the high-dimensional vector space. Second, the SVM algorithm operates under the principle of Structural Risk Minimisation, aiming to find the optimal hyperplane with the maximum margin between classes, rather than merely

DOI : https://doi.org/10.25139/inform.v11i1.11257

estimating class probabilities as Naïve Bayes does. As long as there are sufficient support vectors to define the negative class boundaries, SVM can accurately separate the minority class without being overly biased by the majority class.

To demonstrate the effectiveness of the proposed model, an accuracy comparison was conducted against several relevant studies that used the SVM algorithm across various text domains. As presented in Table XII, the SVM method with TF-IDF optimization developed in this study achieved an accuracy of 90%. This achievement outperformed the research by [16] in the tourism domain (86%) and [32] in the entertainment domain (79.4%).

TABLE XII
COMPARISON OF MODEL ACCURACY WITH PREVIOUS RESEARCH

Research	Method	Object	Accuracy
[13]	SVM + TF – IDF	*Peraturan Pemerintah*	99%
[16]	SVM + Doc2Vec	Rinca Island keyword (X apps)	86%
[32]	SVM +TF – IDF	XfactorID keyword (X apps)	79.44%
[17]	SVM + BERT Embedding	Biznet keyword (X apps)	97%
Proposed	SVM + TF – IDF	Hospital Services (Gmaps)	90%

Even though [13] ported an accuracy of 99%, this difference can be attributed to the dataset's characteristics. Study subjects were formal texts of 'Government Regulations' which have a standard language structure and consistent vocabulary, so that the classification patterns became more distinct (linearly separable). In contrast, the hospital review dataset presents far more complex linguistic challenges. Although intensive pre-processing and slang normalization were performed in this study, variation in informal language styles and the ambiguity of patient emotions still resulted in higher feature dimensions than in legal documents. Therefore, the 90% accuracy in this informal domain represents a very robust model performance.

Additionally, compared with [17], which achieved 97% using a Deep Learning approach (BERT Embedding), this study shows that the classic SVM method also delivers competitive performance (close to 90%) at much lower computational cost. These findings confirm that for the case study of healthcare service reviews in Indonesia, SVM with appropriate pre-processing is an effective and efficient solution.

IV. CONCLUSION

This research successfully mapped public perceptions and emotions regarding hospital services in Semarang City and developed an accurate automatic classification model. Based on an analysis of 16,364 Google Maps reviews, the public generally responded positively, with sentiment at 63.1%, dominated by trust and joy. These findings indicate that the quality of medical services and health facilities in Semarang has met the expectations of the majority of patients.

In terms of intelligent system development, algorithm comparison shows that Support Vector Machine (SVM) with Hyperparameter Tuning is the best method for classifying these reviews. SVM achieves 90% accuracy, significantly superior to Naïve Bayes, which reaches only 78%. The advantage of SVM lies in its ability to handle high-dimensional text data and separate the hyperplane between sentiment classes with a clear margin, as evidenced by the minimal fatal errors in the Confusion Matrix. However, this study still has limitations in the classification of the Neutral class, which has the lowest recall, and an imbalanced dataset that heavily favours the positive class. For further research, it is recommended to apply data imbalance handling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or data augmentation to improve the model's sensitivity to minority classes [33].

REFERENCES

[1] U. K. Nisak and Cholifah, "Buku Ajar Statistik Di Fasilitas Pelayanan Kesehatan," *UMSIDA Press*, pp. 1–107, 2020, Accessed: Jan. 03, 2026, doi:10.21070/2020/978-623-6833-94-0

[2] Kemenkes, "*Peraturan Menteri Kesehatan Republik Indonesia*," 2020.

[3] J. D. C. Aruan, B. Rahayudi, and A. Ridok, "Analisis Sentimen Opini Masyarakat terhadap Pelayanan Rumah Sakit Umum Daerah menggunakan Metode Support Vector Machine dan Term Frequency-Inverse Document Frequency," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, vol. 6, no. 5, pp. 2072–2078, 2022.

[4] Y. Soumokil, M. Syafar, and A. Yusuf, "Analisis Kepuasan Pasien Di Rumah Sakit Umum Daerah Piru," *Jurnal Ilmiah Kesehatan Sandi Husada*, vol. 10, no. 2, pp. 543–551, 2021, doi: 10.35816/jiskh.v10i2.645.

[5] M. Isnan, G. N. Elwirehardja, and B. Pardamean, "Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model," *Procedia Comput. Sci.*, vol. 227, pp. 168–175, 2023, doi: 10.1016/j.procs.2023.10.514.

[6] I. H. Kusuma and N. Cahyono, "Analisis Sentimen Masyarakat Terhadap Penggunaan E-Commerce Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, vol. 8, no. 3, 2023.

[7] A. S. Aribowo and S. Khomsah, "Implementation Of Text Mining For Emotion Detection Using The Lexicon Method (Case Study: Tweets About Covid-19) Implementasi Text Mining Untuk Deteksi Emosi Menggunakan Metode Leksikon (Studi Kasus: Twit Tentang Covid-19)," *Jurnal Informatika dan Teknologi Informasi*, vol. 18, no. 1, pp. 49–60, 2021, doi: 10.31515/telematika.v18i1.4341.

[8] J. Supriyanto, D. Alita, and A. R. Isnain, "Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 4, no. 1, pp. 74–80, Mar. 2023, doi: 10.33365/jatika.v4i1.2468.

[9] A. R. Kardian and D. Gustiana, "Analisis Sentimen Berdasarkan Opini Pengguna pada Media Twitter Terhadap BPJS Menggunakan Metode Lexicon Based dan Naïve Bayes Classifier," *Jurnal Ilmiah Komputasi*, vol. 20, no. 1, Mar. 2021, doi: 10.32409/jikstik.20.1.401.

[10] A. Nugraha, "Analisis Sentimen dalam Text Mining: Memahami Emosi Melalui Bahasa," *Teknologipintar.org*, vol. 3, no. 12, 2023.

[11] S. Rohimah, M. Afdal, M. Mustakim, and R. Novita, "Analisis Sentimen Traveloka Berdasarkan Ulasan Google Play Store Menggunakan Algoritma Support Vector Machine dan Random Forest," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 3, pp. 1709–1716, Dec. 2024, doi: 10.47065/bits.v6i3.6300.

[12] D. R. Andriyani, M. Afdal, and S. Monalisa, "Analisis Sentimen Masyarakat Terhadap Penghapusan Honorer Berdasarkan Opini Dari Twitter Menggunakan Naïve Bayes Classifier," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3541.

[13] M. A. Rani and N. Hendrastuty, "Perbandingan Algoritma NBC Dan SVM Untuk Melakukan Analisis Sentimen Terhadap PP NO.82 Tahun 2021," *Technology and Science (BITS)*, vol. 6, no. 4, 2025, doi: 10.47065/bits.v6i4.6496.

[14] A. E. Perkasa and A. N. Putri, "Penerapan Naïve Bayes Untuk Analisis Sentimen Pada Ulasan Aplikasi Mobile Legends," *Technology and Science (BITS)*, vol. 6, no. 4, 2025, doi: 10.47065/bits.v6i4.6507.

DOI : https://doi.org/10.25139/inform.v11i1.11257

[15] E. Triningsih, M. Afdal, I. Permana, and N. Evrilyan Rozanda, "Analisis Sentimen Terhadap Program Makan Bergizi Gratis Menggunakan Algoritma Machine Learning Pada Sosial Media X," *Technology and Science (BITS)*, vol. 6, no. 4, pp. 2240–2250, 2025, doi: 10.47065/bits.v6i4.6534.

[16] T. H. J. Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 660–667. doi: 10.1016/j.procs.2021.12.187.

[17] S. Hidayatulloh, L. Muflikhah, and R. S. Perdana, "Implementasi Embedding IndoBERT dan Support Vector Machine (SVM) Untuk Analisis Sentimen Publik terhadap Layanan Biznet," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, vol. 9, no. 9, pp. 2548–964, Sep. 2025.

[18] D. C. Rahmadani, S. Khomsah, and M. Y. Fathoni, "Analisis Emosi Wisatawan Menggunakan Metode Lexicon Text Analysis," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 10, no. 1, May 2024, doi: 10.28932/jutisi.v10i1.6690.

[19] T. C. M. Yunanda, M. Hanafi, and W. M. P. Dhuhita, "Sentiment Analysis on TikTok Shop Reviews Using Long Short-Term Memory Method to Find Business Opportunity," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 1, pp. 1–7, Sep. 2023, doi: 10.25139/inform.v9i1.6524.

[20] C. M. Tri Yunanda, M. Hanafi, and W. M. Pradnya Dhuhita, "Sentiment Analysis on TikTok Shop Reviews Using Long Short-Term Memory Method to Find Business Opportunity," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 1, pp. 1–7, Sep. 2023, doi: 10.25139/inform.v9i1.6524.

[21] N. V. Pusean, N. Charibaldi, and B. Santosa, "Comparison of Scenario Pre-processing Performance on Support Vector Machine and Naïve Bayes Algorithms for Sentiment Analysis," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 8, no. 1, pp. 57–63, Jan. 2023, doi: 10.25139/inform.v8i1.5667.

[22] M. Gultom, J. Marikros, and W. Rusli, "Penerapan Vader Sentiment untuk Mendeteksi Sentimen Bahasa Inggris berbasis Website," *SEMINAR NASIONAL CORISINDO*, pp. 13–18, Aug. 2024.

[23] W. G. Irngamsyah, R. Helilintar, and L. Sinta Wahyuniar, "Deteksi Emosi Masyarakat Tentang Penyakit Gagal Ginjal Akut Dengan Metode Emolex Dan Logistic Regression," *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, vol. 9, pp. 2549–7952, 2025.

[24] L. Cahyaningrum, A. Luthfiarta, and M. Rahayu, "Sentiment Analysis on the Impact of MBKM on Student Organizations Using Supervised Learning with Smote to Handle Data Imbalance," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 1, pp. 58–66, Jan. 2024, doi: 10.25139/inform.v9i1.7484.

[25] R. L. Mulianingrum and E. Y. Hidayat, "Comparative Performance of SVM and BERT-Base Using Hybrid Pre-processing for Fast Fashion Sentiment Analysis," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 6, pp. 3464–3478, 2025.

[26] N. Z. B. Jannah and K. Kusnawi, "Comparison of Naïve Bayes and SVM in Sentiment Analysis of Product Reviews on Marketplaces," *Sinkron*, vol. 8, no. 2, pp. 727–733, Mar. 2024, doi: 10.33395/sinkron.v8i2.13559.

[27] A. G. Ghifari, G. Y. Ananada, and K. Purwandari, "A Comparative Sentiment Analysis of Public Opinion on Indonesia's National Football Coach Using CRNN and SVM," *Procedia Comput. Sci.*, vol. 269, pp. 1485–1493, 2025, doi: 10.1016/j.procs.2025.09.090.

[28] D. Wang and Y. Zhao, "Using News to Predict Investor Sentiment: Based on SVM Model," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 191–199. doi: 10.1016/j.procs.2020.06.074.

[29] M. R. Rahman, A. F. Diansyah, and Hanafi, "Sentiment Analysis on the Shopee Application on Playstore Using the Random Forest Classification Method," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 1, pp. 20–24, Nov. 2023, doi: 10.25139/inform.v9i1.5465.

[30] D. Al Akhdaan, Taufik Edy Sutanto, and Muhaza Liebenlito, "Confident Learning pada IndoBERT: Peningkatan Kinerja Klasifikasi Sentimen," *The Indonesian Journal of Computer Science*, vol. 13, no. 5, Oct. 2024, doi: 10.33022/ijcs.v13i5.4401.

[31] S. Khomsah, R. Dias Ramadhani, and S. Wijayanto, "Big Data Analytics to Analyze Sentiment, Emotions, and Perceptions of Travelers (Case Study: Tourism Destination in Purwokerto Indonesia)," *Jurnal E-Komtek (Elektro-Komputer-Teknik)*, vol. 5, no. 2, pp. 284–297, Dec. 2021, doi: 10.37339/e-komtek.v5i2.791.

[32] N. Charibaldi, A. Harfiani, and O. Samuel Simanjuntak, "Comparison of the Effect of Word Normalization on Naïve Bayes Classifier and K-Nearest Neighbor Methods for Sentiment Analysis," *Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 1, pp. 25–31, Dec. 2023, doi: 10.25139/inform.v9i1.7111.

[33] M. Dewi *et al.*, "Penerapan SMOTE-NCL untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Jantung Koroner," *JIP (Jurnal Informatika Polinema)*, vol. 10, no. 1, 2023.

DOI : https://doi.org/10.25139/inform.v11i1.11257