Sentiment Classification of Public Perception on LHKPN Using SVM and Naive Bayes

Ahmad Rijal Hermawan[1], Isa Faqihuddin Hanif [2]*
Department of Industrial Technology and Informatics, Informatics Engineering[1]
Department of Industrial Technology and Informatics, Information Systems and Technology[2]
University of Muhammadiyah Prof. DR. HAMKA
Jakarta, Indonesia
rijalhermawanahmad@gmail.com[1], isa@uhamka.ac.id[2]

***Abstract*— The public's perception of the State Officials' Wealth Report (LHKPN) serves as a vital measure of confidence in the government's commitment to transparency and efforts to combat corruption.This research seeks to examine public sentiment as reflected on the social media platform X. A dataset comprising 1,200 tweets was gathered and processed through various text mining methods, such as case folding, data cleaning, tokenization, normalization, stemming, stopword elimination, and TF-IDF vectorization. The tweets were then manually annotated into two sentiment categories: positive and negative, with 77.3% of tweets labeled as positive and 22.7% as negative. Sentiment classification was conducted using two machine learning algorithms: Support Vector Machine (SVM) and Naive Bayes. The Naive Bayes algorithm recorded an accuracy of 86.66%, with a precision of 0.93, a recall score of 0.88, and an F1-score of 0.87. Conversely, the SVM model with a linear kernel demonstrated superior performance, achieving an accuracy rate of 93.33%, along with a precision of 0.93, recall of 0.98, and an F1-score of 0.95. To uncover frequently occurring topics, WordCloud visualizations were generated. These revealed that positive tweets often included words such as 'lapor' and 'transparan', while negative ones were more likely to contain terms like 'bohong' and 'korupsi'. These findings indicate that public sentiment toward the LHKPN initiative is largely favorable, despite persistent concerns surrounding integrity and trustworthiness in asset reporting. This study highlights the effectiveness of sentiment analysis in gauging public opinion and informing future policy improvements.**

Keywords*— *Sentiment Analysis; Naive Bayes; Support Vector Machine; LHKPN; Social Media

I. INTRODUCTION

Corruption continues to pose a significant challenge to governance, particularly in developing nations such as Indonesia. In response to this persistent problem, the Indonesian government formed the Corruption Eradication Commission (KPK), granting it extensive powers to conduct investigations and bring corruption cases to justice. [1]. One of the primary transparency measures implemented by the KPK is the State Officials' Wealth Report (Laporan Harta Kekayaan Penyelenggara Negara or LHKPN), which functions as a preventive mechanism to identify inconsistencies in the asset declarations of public officials [2]. By fostering openness in asset disclosure, the LHKPN system enhances accountability and contributes to reinforcing public confidence in governance.

In recent times, the emergence of digital platforms has significantly reshaped the way citizens interact with political matters and policy discussions. Platforms like X (previously known as Twitter) have evolved into arenas where individuals openly share their views regarding governmental transparency and ethical governance. These online dialogues provide a rich source of data that can be leveraged through sentiment analysis to better understand public opinion[3].

Sentiment analysis facilitates the automated categorization of public opinions based on textual data. This method has seen extensive use across multiple fields, including the analysis of political developments, policy assessments, and evaluations of public services. For example,[4]employed Support Vector Machine (SVM) to evaluate public sentiment on the issue of lobster seed exports, achieving an accuracy rate of 84.21%. Similarly, [5] utilized SVM to analyze public responses during the 2019 presidential election, and [3]examined sentiment surrounding the KPK Bill, which predominantly reflected negative opinions. These findings underscore the reliability of machine learning techniques—particularly SVM—in interpreting public sentiment on political matters.

Although sentiment analysis plays a significant role in understanding public discourse, specific investigations into public sentiment regarding the LHKPN remain scarce. This research seeks to fill that gap by conducting an analysis of tweets related to the LHKPN using two machine learning algorithms: Support Vector Machine (SVM) and Naive Bayes. Through performance evaluation and comparison of these models, the study offers empirical evidence on public perceptions of asset disclosure by state officials. The results aim to contribute to enhancing policy communication, promoting greater transparency, and utilizing social media as a strategic instrument to measure public trust.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2341, Copyright ©2025
Submitted : April 25, 2025, Revised : May 3, 2025, Accepted : May 6, 2025, Published : May 26, 2025

II. LITERATURE REVIEW

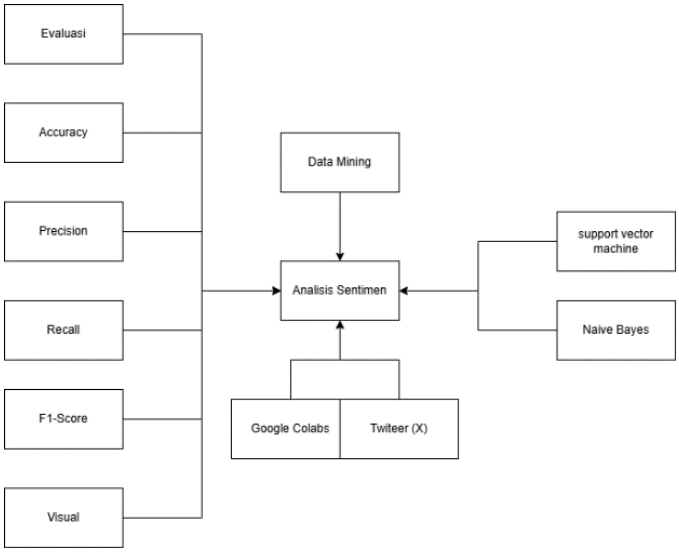

Figure. 1 Conceptual Framework

The conceptual framework outlines the process of data acquisition, preprocessing, sentiment labeling, classification, and evaluation used in this study.

A. Sentiment Analysis

Sentiment analysis is a method within Natural Language Processing (NLP) that focuses on detecting and categorizing the sentiments or emotional tones present in textual data. It is commonly employed to gain insights into public attitudes toward various entities, policies, or occurrences. When applied to social media, sentiment analysis enables the transformation of vast amounts of unstructured data into meaningful and actionable information [6].

B. Naive Bayes Classifier

Naive Naive Bayes is a classification technique grounded in Bayes' Theorem, operating under the strong assumption that each feature—such as individual words in a document—is independent of one another. Despite this assumption, the method has demonstrated high effectiveness in text analysis tasks, valued for its simplicity and strong performance in classification accuracy.[7].

C. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning technique known for its strong performance in handling high-dimensional datasets, such as textual data. It operates by identifying the optimal hyperplane that effectively separates data points into distinct classes. The advantage of SVM is its ability to handle irregular and complex data efficiently[8]

D. State Official Wealth Report (LHKPN)

The LHKPN is a wealth disclosure system administered by Indonesia's Corruption Eradication Commission (KPK), designed as a mechanism for public oversight to deter corruption. It enables citizens to evaluate the integrity of public officials by examining the transparency of their declared assets.[2]

E. Related Research

Various previous studies have effectively applied sentiment analysis to assess public opinion on governmental and political matters in Indonesia. These studies frequently employed machine learning algorithms such as Support Vector Machine (SVM) and Naive Bayes, recognized for their strong performance in text classification tasks. [4] Several studies have utilized Support Vector Machine (SVM) to classify sentiment on various political and governmental issues in Indonesia. For example, one study applied SVM to analyze public sentiment related to the controversy over lobster seed exports, achieving an accuracy rate of 84.21%, thereby demonstrating SVM's effectiveness in addressing issue-specific sentiment analysis on Indonesian Twitter data. Likewise, [5] implemented SVM to examine sentiments during the 2019 presidential election, attaining an accuracy of 91.5%, further affirming SVM's reliability in mining political opinions.[3]

also used SVM to explore public reactions to the revision of the KPK Law, identifying a predominant negative sentiment (60.9%). Their findings highlighted Twitter's potential as a real-time platform for gauging public responses to controversial legislative changes. In another study, [9] compared the performance of SVM and Naive Bayes in analyzing sentiment regarding the Jakarta gubernatorial election, with results showing that SVM surpassed Naive Bayes in both accuracy and precision.[10]

examined the performance of Naive Bayes Classifier (NBC), K-Nearest Neighbors (KNN), and SVM in evaluating public sentiment toward government performance. Among these algorithms, SVM delivered the highest performance, further validating its robustness in text-based sentiment analysis.While these studies have provided meaningful insights into sentiment analysis using machine learning in the context of elections, public policies, and services, none have specifically investigated public sentiment regarding state wealth transparency through the LHKPN system. This study addresses that gap by applying SVM and Naive Bayes to assess public opinion on LHKPN, thereby contributing a novel perspective. It expands the application of sentiment analysis into the domain of institutional transparency and preventive anti-corruption efforts—an area that remains underrepresented in current research.

III. RESEARCH METHODOLOGY

This study adopts a quantitative methodology by leveraging machine learning algorithms to categorize public sentiment toward the State Officials' Wealth Report (LHKPN) as expressed on the social media platform X. The method ological stages include data collection, preprocessing, annotation, feature extraction, classification, and evaluation.

1) Data Collection

A total of 1,200 tweets related to LHKPN were collected using the X API v2 (formerly Twitter API). Keywords such as "LHKPN", "lapor harta", and "transparansi pejabat" were used to retrieve relevant tweets. Data was gathered using the Tweepy library in Python, executed within the Google Colaboratory environment. In accordance with ethical research standards, all personally identifiable information was either removed or

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2341, Copyright ©2025
Submitted : April 25, 2025, Revised : May 3, 2025, Accepted : May 6, 2025, Published : May 26, 2025

anonymized during the preprocessing stage.

2) Data Preprocessing

3) The tweets underwent multiple preprocessing steps to standardize the data:

1. *Case Folding*
   - Converting all characters in the text to lowercase to standardize the input data.
2. *Cleansing*
   - Removing unnecessary elements such as punctuation marks, special characters, numbers, and hyperlinks to reduce noise in the dataset.
3. *Tokenizing*
   - The text was segmented into individual words or tokens to enable more detailed analysis at the lexical level.
4. *Normalization*
   - Converting informal or slang words into their formal equivalents, for example using a dictionary or predefined list.
5. *Stemming*
   - Words were reduced to their root or base forms using an algorithm such as the Nazief-Adriani stemmer, which is specifically designed for the Indonesian language.
6. *Stopword Removal*
   - Common words that do not carry significant sentiment, such as "dan", "yang", and "itu", were removed because they are irrelevant for sentiment classification.

F. Data Annotation

The dataset was manually labeled into two sentiment categories: positive and negative. Annotation was performed by three independent human annotators with background in data science and communication studies. Each tweet was assessed in terms of its tone, context, and underlying meaning. Any disagreements among annotators were addressed through discussion, and inter-annotator agreement was tracked to maintain consistency in the labeling process. The final distribution was **77.3%** positive (925 tweets) and **22.7%** negative (275 tweets).

Annotation Examples:

- *Positive: "Bagus sekali pejabat ini, sudah melaporkan kekayaannya secara transparan."*
- *Negative: "Pejabat kok tidak lapor harta, pasti ada yang disembunyikan."*

G. Feature Extraction

Text data was transformed into numerical features using the Term Frequency–Inverse Document Frequency (TF-IDF) method. This method quantifies a word's relevance within a document in relation to the entire corpus, making it well-suited for high-dimensional text classification tasks. The resulting TF-IDF matrix served as input for the machine learning algorithms.

H. Data Splitting and Balancing

The dataset was split into 80% for training and 20% for testing using a stratified method to preserve the distribution of sentiment classes. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed on the training data. This technique enhances the representation of the minority class by generating synthetic samples through interpolation between existing instances, thereby improving the model's capacity to learn from the limited negative class data.

I. Classification Algorithms

Two classification models were implemented and compared:

- *Naive Bayes (Multinomial NB): A probabilistic model assuming feature independence, commonly used in text classification due to its simplicity and speed.*
- *Support Vector Machine (SVM): Implemented with a linear kernel and regularization parameter C = 1.0, SVM was selected for its ability to handle high-dimensional and sparse data typical of textual inputs.*

J. Evaluation Metrics

To evaluate the effectiveness of the models, several performance metrics were utilized:

- *Accuracy: Measures the proportion of total predictions that the model got right, reflecting its overall reliability.*
- *Precision: Represents the ratio of correctly predicted positive cases to all instances that were labeled as positive, highlighting the model's ability to avoid false positives.*
- *Recall: Indicates the proportion of actual positive instances that were successfully detected by the model, emphasizing its sensitivity to relevant data.*
- *F1-Score: Combines both precision and recall into a single score using their harmonic mean, making it particularly valuable when dealing with uneven class distributions.*
- *Confusion Matrix: A comprehensive table that categorizes prediction outcomes into true positives, true negatives, false positives, and false negatives. It provides an in-depth view of how the model performs across different classification outcomes, aiding in the identification of specific areas for improvement.*

IV. Results And Discussion

The first step in this research involved collecting tweets related to the State Officials' Wealth Report (LHKPN) using the X API v2. This was conducted using Python and the Tweepy library within Google Colaboratory. A total of 1,200 tweets were gathered based on keywords such as "LHKPN", "lapor harta", and "transparansi pejabat".

```
# Crawl Data

filename = 'Data Skripsi LHKPN Rijal 2.csv'
search_keyword = 'LHKPN KPK'
limit = 1000

!npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token {twitter_auth_token}
```

Figure 2 Crawling Data

Figure 2 shows the Python script implementation used to perform tweet crawling, including authentication process

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2341, Copyright ©2025
Submitted : April 25, 2025, Revised : May 3, 2025, Accepted : May 6, 2025, Published : May 26, 2025

and query formulation.

the raw tweets were reviewed. These included various expressions of public opinion ranging from support for transparency to criticism of corruption.

	Text
921	50 Menteri dan Wakil Menteri Kabinet Merah Putih Prabowo-Gibran Belum Sampaikan LHKPN ke KPK https://t.co/4kFI3SZ2EU
321	@M45Broo_ Gak ada yg ribut terkait LHKPN andika coba kalo dr kim plus koar2 KPK suruh usut
101	@gea_asa @FarhanAtjeh @prastow Sgt bs utk @KPK_RI menjalankan pembuktian terbalik atas LHKPN pejabat2. Hny sj kembali kpd good will KPK itu sndr dan landasan UU/Aturan Pembuktian Terbalik- nya itu ada/blm ada. Kl blm ada aturannya mrk pejabat2 itu bs berk
920	KPK meminta Menteri dan Wakil Menteri Kabinet Merah Putih untuk segera melaporkan harta kekayaan dengan mengisi LHKPN. Batas waktu yang diminta KPK yakni tiga bulan setelah dilantik atau pada Januari 2025. https://t.co/NHtuO4G4o1
58	@intinyadeh LHKPN ini kayaknya lucu deh coba kita cek setiap pejabat berapa aja. Pasti jadi ladang meme banyak banget Lagian KPK ini percaya ga sama beginian? Mayoritas rakyat aja ga percaya eh
790	Tambah berat Ya Min @Pupufafa1 @Gerindra pak @prabowo .. Kau lapor kan itu LHKPN ke KPK Taim Miftah
948	@TOMSETOM_DR berani ga KPK cek LHKPN ne AL ?
969	@dhemit_is_back @KPK_RI @DitjenPajakRI Ini kelemahan lhkpn masih sebatas self asesmen harus ada pembuktian terbalik
410	Polisi harus masuk selidiki dugaan penggunaan jam kw itu. KPK juga harus masuk selidiki jam itu yg konon tak masuk LHKPN
1079	Eng ing eng.. KPK Turun Tangan Dalami Asal Usul Jam Tangan Mewah Milik Dirdik Jampidsus Kejagung Saya lihat dulu ya (LHKPN Qohar) kata Deputi Bidang Pencegahan dan Monitoring @KPK_RI Pahala Nainggolan https://t.co/euJwTp4hmN

Figure 3 Random data crawling results

Figure 3 displays a screenshot of the raw dataset structure, which includes tweet ID, timestamp, tweet text, and user information. All personal identifiers were anonymized for ethical compliance.

The tweets underwent multiple preprocessing steps to clean and standardize the data for sentiment classification.

No	Text Original	Text Processed
1	LHKPN itu cuma formalitas., ngisinya asal asalan dan KPK nya malas ngecek..	lhkpn formalitas ngisinya asal kpk nya malas ngecek
2	Buntut Kasus Aniaya Dokter Koas KPK Sebut Proses Analisis LHKPN Dedy Mandarsyah Berlangsung 1 Pekan https://t.co/5zCVXyHlct	buntut aniaya dokter koas kpk proses analisis lhkpn dedy mandarsyah pekan
3	Mari kita terus viralkan agar @KPK_RI tak perlu menunggu dua minggu untuk melakukan penyelidikan LHKPN dan penyidik jika telah memiliki dua bukti permulaan.	mari viralkan kpkri tunggu minggu lidi lhkpn sidik milik bukti mula
4	@KPK_RI apakah sudah terima LHKPN bapaknya lady ?	kpkri terima lhkpn bapak lady
5	KPK Ungkap Ada Pejabat Negara Tidak Jujur Isi LHKPN Sindir Jaksa Agung? https://t.co/B7yK4CVLYq	kpk jabat negara jujur isi lhkpn sindir jaksa agung

Figure 4 Data Processing Results

Figure 4 demonstrates a sample of tweets before and after preprocessing, showing the transformation from noisy text to a clean token list ready for vectorization.

Insert a screenshot of the code or interface used in Google Colab that demonstrates the data crawling process using the X API v2. This can be a snippet of Python code with function calls such as tweepy.Client() or search_recent_tweets() along with a visible terminal output showing total tweets collected., conducted via Google Colaboratory to collect tweet data containing keywords related to the LHKPN. This initial step resulted in 1,200 tweets for analysis.

Results tweets were manually labeled into positive and negative sentiments. This involved human annotators assessing the tone and implication of each tweet based on contextual interpretation.

	Text	label
2	lhkpn formalitas ngisinya asal kpk nya malas ngecek	negatif
3	buntut aniaya dokter koas kpk proses analisis lhkpn dedy mandarsyah pekan	positif
4	mari viralkan kpkri tunggu minggu lidi lhkpn sidik milik bukti mula	positif
5	kpkri terima lhkpn bapak lady	positif
6	kpk jabat negara jujur isi lhkpn sindir jaksa agung	negatif
7	kalo fair lapor lhkpn ga suruh nilai aset input harta lokasi kpk kerjasama surveyor finance nilai aset	positif
8	opposisi m udah aneh sampe puluh m kpk lhkpn ya gak tindak lhkpn rakyat yg ngecek ricek	negatif
9	kpkri ada analisis investigasi thd lhkpn an jokowi gibrantweet nih biar kpkri sehat kuat jumat sepeda kk	positif
10	dhemitisback nih yg d mksd kpk jabat negara yg asaln isi lhkpn lg isitulis	negatif

Figure 5 Manual data labeling results

Figure 5 shows labeled data samples, where tweets are presented with corresponding sentiment labels. This figure is intended to demonstrate the outcome of manual sentiment annotation conducted during the labeling stage. Each tweet is assessed based on its linguistic tone and context, and then assigned either a 'Positive' or 'Negative' sentiment.

In the example shown in Figure 4, two tweets that express support or appreciation for transparency in public official reporting are labeled as Positive, while two tweets that express skepticism or criticism towards the LHKPN process are labeled as Negative. This stage is crucial for ensuring that the supervised learning model is trained with correct and contextually appropriate data.

Each tweet is assigned either a positive or negative sentiment, with a total of 925 positive and 275 negative entries. the data revealed an imbalance with significantly more positive tweets.

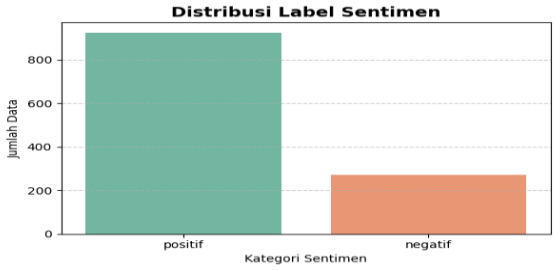

Figure 6 Distribution of Sentiment Data Labeling

Figure 6 presents a bar chart that visualizes the total count of positive and negative sentiments observed in the dataset. This chart serves to give a quick and intuitive overview of the sentiment distribution resulting from the manual annotation process.

As depicted, the positive sentiment category significantly outnumbers the negative one, with 925 tweets classified as positive and 275 as negative. The dominance of positive sentiment reflects the general tone of public discourse on Twitter regarding the LHKPN. Such visual representation supports further analysis and model training by confirming the class imbalance, which is subsequently addressed using techniques like SMOTE.

To prepare the data for machine learning, the tweet texts were transformed into numerical vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) method.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2341, Copyright ©2025
Submitted : April 25, 2025, Revised : May 3, 2025, Accepted : May 6, 2025, Published : May 26, 2025

	agung	analisis	aniaya	asal	bapak	bukti	buntut	\
0	0.000000	0.000000	0.000000	0.430786	0.000000	0.000000	0.000000	
1	0.000000	0.321926	0.339942	0.000000	0.000000	0.000000	0.333278	
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.294949	0.000000	
3	0.000000	0.000000	0.000000	0.000000	0.588212	0.000000	0.000000	
4	0.441058	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	

	dedy	dokter	formalitas	...	ngecek	ngisinya	nya	pekan
0	0.00000	0.000000	0.368098	...	0.403046	0.458526	0.251883	0.000000
1	0.30438	0.339942	0.000000	...	0.000000	0.000000	0.000000	0.379154
2	0.00000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000
3	0.00000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000
4	0.00000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000

	proses	sidik	sindir	terima	tunggu	viralkan
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.317003	0.000000	0.000000	0.000000	0.000000	0.000000
2	0.000000	0.322952	0.000000	0.000000	0.284954	0.383177
3	0.000000	0.000000	0.000000	0.460069	0.000000	0.000000
4	0.000000	0.000000	0.566548	0.000000	0.000000	0.000000

[5 rows x 37 columns]

Figure 7 Text Vaccination Results

Figure 7 illustrates a portion of the TF-IDF matrix, showing token importance across sample documents. This figure is designed to provide insight into how unstructured tweet text is quantitatively represented for machine learning processes.

To convert text data into a numerical format compatible with machine learning algorithms, this research implemented the Term Frequency–Inverse Document Frequency (TF-IDF) technique. This approach assigns a weight to each term based on how often it appears in a particular document (term frequency) and how rare it is across all documents in the dataset (inverse document frequency). Words that are frequently used in an individual tweet but rarely appear in the overall corpus are given greater importance, making them more influential in the classification process.

In contrast to Count Vectorization, the TF-IDF method demonstrates superior capability in identifying significant terms while effectively reducing the influence of frequently occurring, non-informative words. In this study, the preprocessing steps combined with the TF-IDF transformation generated 2,311 distinct features, which served as input for the classification algorithms.

This method was chosen due to its computational efficiency and strong performance in prior sentiment analysis studies. The resulting TF-IDF matrix provided a high-dimensional, sparse representation of tweet content, allowing SVM and Naive Bayes to learn sentiment-related patterns effectively.

was divided into training and testing sets in an 80:20 ratio to evaluate model performance effectively.

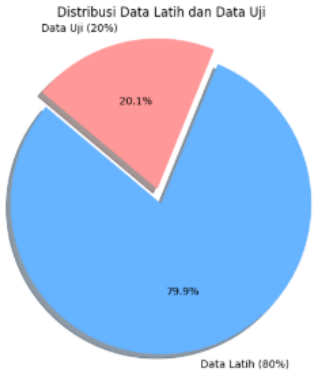

Figure 8 Data splitting results

Figure 8 displays the proportion and count of training and test data points per sentiment category. This figure aims to explain how the dataset was divided to ensure robust evaluation of model performance.

An 80:20 data split is a widely adopted approach in machine learning, enabling the model to train on the bulk of the dataset while assessing its ability to generalize using the remaining unseen portion. The accompanying visualization illustrates how the data is distributed across classes in both sets, providing clarity and transparency during the model preparation stage.

To address the issue of imbalanced data, this study employed the Synthetic Minority Over-sampling Technique (SMOTE) on the training dataset. By generating artificial samples for the underrepresented class—specifically, those labeled with negative sentiment—SMOTE helped balance the dataset. This enhancement enabled the classification models to better capture and recognize sentiment trends across both majority and minority classes, ultimately improving predictive performance.

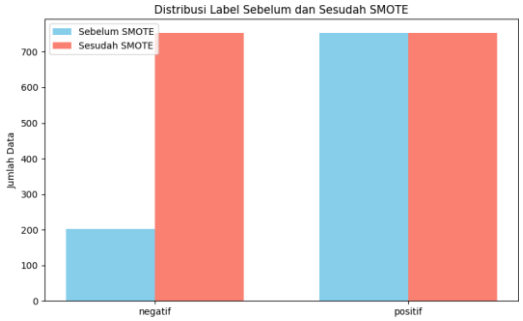

Figure 9 Data Smote Method

Figure 9 provides a visual explanation of the SMOTE process and how synthetic samples are generated. This technique is applied to balance the dataset by oversampling the minority class.

The illustration demonstrates how SMOTE operates conceptually by generating synthetic samples along the lines connecting minority class samples with their nearest neighbors. By introducing these artificial instances, the dataset becomes more balanced, which in turn improves the model's classification accuracy—especially in correctly identifying samples belonging to the minority class, such as those expressing negative sentiment.

results showed that 77.3% of the data were classified as positive sentiment and 22.7% as negative.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2341, Copyright ©2025
Submitted : April 25, 2025, Revised : May 3, 2025, Accepted : May 6, 2025, Published : May 26, 2025

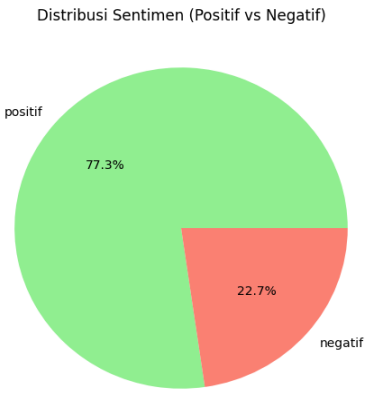

Figure 10 Sentiment Analysis Results

Figure 10 presents the overall sentiment analysis result. This pie chart is a summarization of the sentiment classification, showing the proportion of tweets categorized as positive and negative after the model predictions.

The dominance of positive sentiment in the chart reflects the overall tone of public perception on Twitter toward the LHKPN. The visual also validates the earlier manual annotation distribution and supports the interpretation that the public tends to support transparency initiatives.

TABLE I. MODEL PERFORMANCE EVALUATION REPORT

	Precision	Recall	F1-Score	Support
Negative	0.73	0.84	0.78	69
Positive	0.93	0.88	0.90	171
Accuracy			0.87	240
Macro Avg	0.83	0.86	0.84	240
Weighted Avg	0.87	0.87	0.87	240

TABLE I displays the classification report for the Naive Bayes algorithm, highlighting key performance indicators such as accuracy, precision, recall, and F1-score. This report provides a detailed overview of how well the model performs across various sentiment categories. The presented metrics reflect the model's capability to distinguish between positive and negative sentiments. Accuracy indicates the proportion of correct predictions overall, while precision and recall delve into how effectively the model identifies each sentiment type. The F1-score, which combines precision and recall into a single measure, proves especially useful in scenarios with imbalanced sentiment distributions.

TABLE II. SVM MODEL PERFORMANCE EVALUATION REPORT

	Precision	Recall	F1-Score	Support
Negative	0.95	0.81	0.88	69
Positive	0.93	0.98	0.95	171
Accuracy			0.93	240
Macro Avg	0.94	0.90	0.91	240
Weighted Avg	0.93	0.93	0.93	240

TABLE II Displays the classification report for the SVM model, which demonstrated superior performance compared to Naive Bayes across all evaluation metrics. This figure provides a comprehensive summary of the SVM model's effectiveness in performing the sentiment classification task.

The higher values in precision, recall, and F1-score compared to the Naive Bayes model suggest that SVM is more robust, especially in identifying minority class instances. This supports the selection of SVM as the preferred model for this particular sentiment analysis case.

- *Naive Bayes*: Accuracy 86.66%, precision 0.93, recall 0.88, f1-score 0.87
- *SVM*: Accuracy 93.33%, precision 0.93, recall 0.98, f1-score 0.95

TABLE III. CONFUSION MATRIX NAIVE BAYES

	Actual: Positive	Actual: Negative
Predicted: Positive	TP: 150	FP: 11
Predicted: Negative	FN: 21	TN: 58

Table III The Naive Bayes algorithm attained an accuracy rate of 86.66%, but it exhibited a noticeable tendency to incorrectly classify several positive tweets. According to the confusion matrix, it successfully identified 150 tweets as true positives and 58 as true negatives. Nevertheless, 21 positive tweets were mistakenly categorized as negative (false negatives), and 11 negative tweets were classified as positive (false positives). Despite achieving high precision in detecting positive sentiment, the substantial number of false negatives led to a lower recall, highlighting the model's limitations in accurately capturing negative sentiment.

TABLE IV. CONFUSION MATRIX SVM

	Actual: Positive	Actual: Negative
Predicted: Positive	TP: 168	FP: 13
Predicted: Negative	FN: 3	TN: 56

On the other hand, the Support Vector Machine (SVM) model delivered better overall results, reaching an accuracy of 93.33%. As reflected in the confusion matrix, the model accurately recognized 168 positive tweets (true positives) and 56 negative tweets (true negatives), with only 3 positive tweets misclassified as negative (false negatives) and 13 negative tweets incorrectly labeled as positive (false positives). The minimal false negative rate suggests that the model possesses a high recall, indicating its effectiveness in identifying genuine positive sentiments. While a few misclassifications occurred in the negative class, the results overall affirm that SVM is more consistent and dependable than Naive Bayes for sentiment classification within this dataset.

Figure 11 Positive Sentiment WordCloud

Figure 15 illustrates the WordCloud for positive sentiment

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2341, Copyright ©2025
Submitted : April 25, 2025, Revised : May 3, 2025, Accepted : May 6, 2025, Published : May 26, 2025

tweets. This visualization highlights the most frequently occurring keywords in tweets labeled as positive.

The larger the word appears in the WordCloud, the more frequently it occurs in the dataset. Words such as "lapor," "transparan," and "harta" are indicative of public approval, often linked with praise for transparent reporting practices or admiration for public figures who disclose their wealth properly. The WordCloud serves as a visual summary of public expressions of trust and appreciation toward the LHKPN process. Prominent keywords consist of terms like “lapor,” “transparan,” and “harta,” along with mentions of notable individuals such as ministers or celebrities known for advocating transparency.

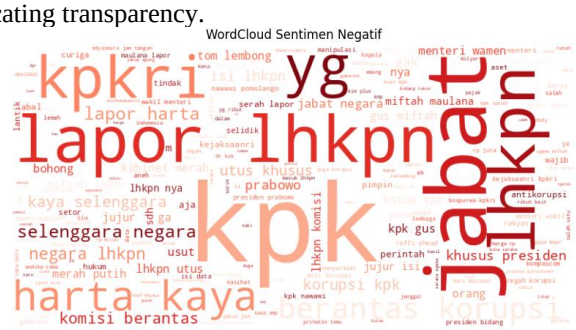

Figure 12 Negatif Sentiment WordCloud

Figure 16 shows the WordCloud for negative sentiment tweets. This figure emphasizes the dominant words used in tweets expressing dissatisfaction, skepticism, or criticism toward the LHKPN.

Prominent terms like "bohong," "korupsi," and "tidaklapor" suggest public concerns about dishonesty and lack of compliance by some officials. The negative WordCloud helps identify key issues in public discourse and reflects areas where transparency efforts may still be perceived as insufficient. Keywords like “bohong,” “korupsi,” and “tidaklapor” reflect public skepticism regarding the integrity of certain officials in disclosing their assets.

Dominant keywords:

- *Positive*: "lapor" (report), "transparan" (transparent)
- *Negative*: "bohong" (lie), "korupsi" (corruption)

SVM shown better performance due to its capability to handle high-dimensional data with complex distributions. These findings are consistent with previous studies.

V. CONCLUSION

This research effectively categorized public sentiment regarding the State Officials’ Wealth Report (LHKPN) shared on the social media platform X by utilizing Support Vector Machine (SVM) and Naive Bayes algorithms. A total of 1,200 tweets were gathered and manually annotated as either positive or negative, with 77.3% identified as positive and 22.7% as negative. The sentiment analysis process involved thorough text preprocessing and TF-IDF feature extraction, followed by an 80:20 split between training and testing datasets.

The evaluation revealed that the SVM model surpassed Naive Bayes across all perforsmance indicators, achieving 93.33% accuracy, 0.93 precision, 0.98 recall, and an F1-score of 0.95. Although Naive Bayes proved to be a fast and straightforward approach, its accuracy reached only 86.66%, and it struggled with identifying negative sentiment effectively. The analysis was further reinforced by WordCloud visualizations, which highlighted commonly used terms in each sentiment category—such as lapor (report), transparan (transparent), and jujur (honest) in positive tweets, and bohong (lie), korupsi (corruption), and tidaklapor (not reporting) in negative ones.

The findings indicate that public perception of the LHKPN initiative is largely positive, reflecting strong support for transparent disclosure of state officials’ assets. These insights can provide valuable guidance for policymakers, particularly the Corruption Eradication Commission (KPK), in evaluating public confidence in asset reporting systems and enhancing their outreach and communication efforts. Moreover, the existence of critical viewpoints underscores the public’s call for stronger verification mechanisms and stricter enforcement of asset declaration policies.

Looking ahead, this research is limited to binary sentiment classification using traditional machine learning algorithms. Future research is recommended to investigate more advanced approaches, such as deep learning techniques like Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT), which are capable of capturing deeper semantic meaning and contextual nuances in sentiment analysis. Additionally, incorporating multi-class sentiment classification (e.g., positive, negative, neutral, sarcastic), topic modeling, and temporal trend tracking could offer richer insights into public opinion and deliver more adaptive, real-time feedback to support transparency initiatives in governance.

ACKNOWLEDGMENT

The author gratefully acknowledges the blessings and guidance of Allah SWT, which have been instrumental in the successful completion of this research. Heartfelt appreciation is extended to the academic advisor for their insightful guidance, valuable suggestions, and consistent encouragement throughout the research journey. The author also wishes to thank family and friends for their unwavering support and motivation. Gratitude is further extended to all individuals and organizations involved in the data gathering process and in facilitating this study. It is hoped that the outcomes of this research will make a meaningful contribution to the fields of sentiment analysis and studies on public transparency.

REFERENCES

[1] R. Y. Oly Viana Agustine, Erlina Maria Christin Sinaga, “Politik Hukum Penguatan Kewenangan Komisi Pemberantasan Korupsi dalam Sistem Ketatanegaraan Legal Politics of the Strengthening of Authority in the Constitutional System,” *Konstitusi*, vol. 16, no. 2, pp. 314–338, 2019.

[2] U. M. Sosiawan, “Peran Komisi Pemberantasan Korupsi (KPK) Dalam Pencegahan dan Pemberantasan Korupsi,” *J. Penelit. Huk. Jure*, vol. 19, no. 4, p. 517, 2019, doi: 10.30641/dejure.2019.v19.517-538.

[3] R. Nooraeni, H. D. Sariyanti, A. F. F. Iskandar, S. F. Munawwaroh, S. Pertiwi, and Y. Ronaldias, “Analisis Sentimen Data Twitter Mengenai Isu RUU KPK Dengan Metode Support Vector Machine (SVM),” *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 1, pp. 55–60, 2020, doi:

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2341, Copyright ©2025
Submitted : April 25, 2025, Revised : May 3, 2025, Accepted : May 6, 2025, Published : May 26, 2025

10.31294/p.v22i1.6869.

[4] B. Pamungkas, A. Syaifuddin, and M. Muslimin, "Analisis Sentimen Twitter MenggunakanMetode Support Vector Machine (SVM) padaKasus Benih Lobster 2020," *J. Informatics, Inf. Syst. Softw. Eng. Appl.*, vol. 3, no. 2, pp. 10–20, 2021.

[5] O. Zoellanda A.Tane, K. Muslim Lhaksmana, and F. Nhita, "Analisis Sentimen pada Twitter Tentang Calon Presiden 2019 Menggunakan Metode SVM (Support Vector Machine)," *eProceedings Eng.*, vol. 6, no. 2, pp. 9716–9725, 2019.

[6] B. W. Sari and F. F. Haranto, "Implementasi Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom Dan Biznet," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 171–176, 2019, doi: 10.33480/pilar.v15i2.699.

[7] A. Rosadi *et al.*, "Analisis Sentimen Berdasarkanan Opini Pengguna pada Media Twitter Terhadap BPJS Menggunakan Metode Lexicon Based dan Naïve Bayes Classi er Twitter Text Mining," vol. 20, pp. 39–52, 2021.

[8] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, "Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi," *J. Teknoinfo*, vol. 14, no. 2, p. 115, 2020, doi: 10.33365/jti.v14i2.679.

[9] E. Putra Nuansa, "Analisis Sentimen Pengguna Twitter Terhadap Pemilihan Gubernur Dki Jakarta Dengan Metode Naïve Bayesian Classification Dan Support Vector Machine," *Inst. Teknol. Sepuluh Nop. Surabaya*, pp. 1–101, 2017.

[10] P. Simposium, N. Multidisiplin, and U. M. Tangerang, "Analisis Sentimen Kinerja Pemerintahan Menggunakan Algoritma," vol. 4, pp. 114–121, 2022.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2341, Copyright ©2025
Submitted : April 25, 2025, Revised : May 3, 2025, Accepted : May 6, 2025, Published : May 26, 2025