Indonesian Journal of Data and Science

Volume 7 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v7i2.445

Research Article

Weakly Supervised Sentiment Analysis of Gold Price Discussions Using Conventional Machine Learning and IndoBERT

M Rizki Hardika [1]; Brina Miftahurrohmah [2,*]
[1] Universitas Internasional Semen Indonesia, Gresik, Indonesia, m.hardika22@student.uisi.ac.id
[2] Universitas Internasional Semen Indonesia, Gresik, Indonesia, brina.miftahurrohmah@uisi.ac.id

Correspondence should be addressed to Brina Miftahurrohmah; brina.miftahurrohmah@uisi.ac.id

Received 10 March 2026; Accepted 30 July 2026; Published 31 July 2026

Abstract:
Introduction: Gold price movements attract substantial public and investor attention because gold serves as both a safe-haven asset and a hedging instrument. This study investigates Indonesian public sentiment toward gold price discussions on Platform X using a weakly supervised sentiment-analysis framework. **Method**: A total of 7,283 Indonesian-language tweets containing the keyword "harga emas" were collected during 2023–2025, with 4,429 tweets retained after preprocessing. Sentiment labels were generated using a domain-specific lexicon and validated through manual annotation. Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, and IndoBERT were evaluated using the same train–test partition. **Results and Discussion**: Manual validation achieved a Cohen's Kappa of 0.8718, indicating almost perfect inter-annotator agreement, while the lexicon-based labels achieved 70.62% accuracy against the manually annotated reference. IndoBERT achieved the highest performance on weakly supervised labels with 98.31% accuracy and a 98.16% macro F1-score, outperforming SVM, Naïve Bayes, and KNN. However, its accuracy decreased to 69.49% when evaluated against manually annotated data, demonstrating that downstream performance remains strongly influenced by weak-label quality. **Conclusion**: Weak supervision provides an efficient and scalable approach for large-scale Indonesian financial sentiment annotation, while contextual models such as IndoBERT offer superior classification performance; however, reliable manual validation remains essential to mitigate label noise and improve generalizability.

Keywords: Sentiment Analysis; Gold Price; Weak Supervision; IndoBERT; Platform X

1. Introduction

Gold has long been recognized as one of the most reliable investment instruments because of its intrinsic value, high liquidity, and ability to preserve wealth over time [1]. Unlike many financial assets whose values fluctuate considerably during periods of economic uncertainty, gold is widely regarded as a *safe-haven* asset that provides protection against market turbulence [2]. Gold also serves as an effective hedging instrument against inflation and financial crises, making it an important component of investment portfolio diversification [3], [4]. Consequently, fluctuations in gold prices continue to attract considerable attention from investors, financial institutions, and policymakers because they often reflect changes in economic conditions and market confidence [4], [5]

The period from 2023 to 2025 represents an important phase for gold investment because global gold prices experienced substantial fluctuations driven by persistent inflation, monetary tightening by major central banks, and increasing geopolitical tensions [1]. These economic conditions encouraged investors to seek safer investment instruments while simultaneously stimulating extensive public discussion regarding gold price movements across digital platforms [2]. As public opinion increasingly influences financial decision-making, understanding investor sentiment has become complementary to conventional market indicators in explaining financial market dynamics [5], [6]. Previous studies have shown that investor sentiment extracted from online discussions and social media can influence market behavior, asset valuation, and investment decisions [7], [8]

The rapid development of digital communication has fundamentally transformed the way financial information is created, disseminated, and consumed [5], [9] .Besides relying on macroeconomic indicators and fundamental analysis, investors increasingly monitor public opinion shared through social media to understand market perception and identify early reactions to financial events [5]. Social media platforms provide large volumes of user-generated content that reflect public responses almost instantly, making them valuable alternative sources of market intelligence [10]. Among various social media platforms, Platform X (formerly Twitter) is particularly suitable for financial sentiment analysis because its real-time, text-oriented, and open communication characteristics facilitate the rapid dissemination of investment-related information and opinions [11]. Furthermore, Indonesia has experienced substantial growth in digital technology adoption and social media usage, making Indonesian-language discussions on Platform X an important source for understanding public perceptions of gold investment in the domestic market [12]

Despite its potential, extracting meaningful information from social media remains challenging because textual data are inherently unstructured, noisy, and highly contextual [13]. Financial discussions frequently contain abbreviations, informal language, domain-specific terminology, and ambiguous expressions whose meanings depend on contextual interpretation rather than individual words [14]. For example, a decline in gold prices may represent negative sentiment for current investors while simultaneously indicating positive sentiment for prospective buyers expecting lower purchase prices. Such contextual ambiguity makes manual analysis impractical and motivates the application of Natural Language Processing (NLP), particularly sentiment analysis, to automatically classify opinions into positive, negative, and neutral categories [15], [13].

Various machine learning approaches have been proposed for sentiment analysis, ranging from conventional classifiers such as Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) to more recent transformer-based language models [16], [17]. Classical machine learning methods combined with TF-IDF feature representation remain attractive because they are computationally efficient, interpretable, and effective for sparse text classification [13]. Nevertheless, TF-IDF primarily captures lexical frequency and often cannot adequately represent contextual semantic relationships between words [18]. In contrast, transformer-based language models such as BERT and IndoBERT learn contextual representations by considering surrounding words within a sentence, enabling a deeper understanding of semantic meaning in Indonesian texts, particularly in financial discussions where sentiment is highly context-dependent [19]. Consequently, comparing conventional feature representation with contextual language representation provides a more comprehensive evaluation than comparing classification algorithms alone.

Several previous studies have compared the performance of Naïve Bayes and K-Nearest Neighbor in sentiment classification and reported inconsistent findings [20], [21], [22]. More recent studies have reported strong performance of Support Vector Machine for text classification tasks involving high-dimensional TF-IDF representations, making it one of the most widely used conventional machine learning algorithms for sentiment analysis [23], [24]. Furthermore, transformer-based language models have consistently demonstrated superior performance over traditional machine learning methods across various NLP tasks by capturing richer contextual information [18]. However, most previous studies focused primarily on comparing classification algorithms using general-purpose sentiment datasets, whereas studies specifically investigating Indonesian-language public sentiment toward gold price movements remain limited. Moreover, only a few studies have simultaneously compared classical machine learning models and contextual language models within the same financial-domain dataset.

Another important issue concerns the reliability of sentiment labels. Constructing large manually annotated datasets for Indonesian financial texts is time-consuming and resource-intensive, leading many studies to adopt lexicon-based labeling as a form of weak supervision [25]. Although this approach enables efficient annotation of large-scale datasets, lexicon-generated labels may contain noise because financial sentiment often depends on contextual meaning that cannot always be captured through dictionary-based polarity alone [6]. Recent studies have shown that noisy labels generated through weak supervision can substantially affect downstream classification performance, making label quality assessment an essential component of machine learning pipelines [26], [27]. Therefore, lexicon-based labels should not be considered absolute ground truth but rather require validation through manual annotation before being used in supervised classification.

Based on these research gaps, this study investigates public sentiment toward gold price movements using Indonesian-language tweets collected from Platform X during the 2023–2025 period. Lexicon-based labeling is employed as an initial weak-supervision strategy and subsequently validated through manual annotation to evaluate label reliability before model development. This study compares the performance of Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine using TF-IDF feature representation and further evaluates IndoBERT as a contextual language model. The proposed framework is expected to provide a more comprehensive understanding of Indonesian public sentiment toward gold price movements while simultaneously evaluating the effectiveness of classical machine learning and transformer-based approaches for financial sentiment analysis.

2. Method

The research flow chart explains the steps to be taken in this study. The research flowchart is illustrated in **Figure 1**.

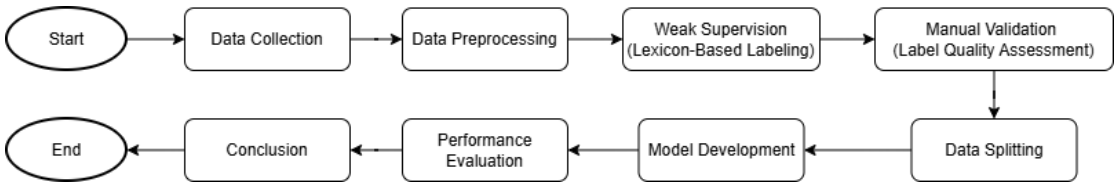

Figure 1. Research Flow

Data Collection

This study began by collecting public opinion data from Platform X using a data crawling technique. The data collection process was conducted through the X Application Programming Interface (API) using the Tweet Harvest tool executed in the Google Colab environment. This approach enables the retrieval of publicly available tweets based on predefined search keywords, ensuring that the collected data are relevant to the research objective. The keyword used in this study was "harga emas", as the research focuses on analyzing public sentiment expressed in Indonesian-language tweets regarding gold price movements. Using an Indonesian keyword allows the collected data to better represent discussions from Indonesian users related to gold investment and market conditions.

Data were collected from January 2023 to December 2025. This period was selected because the global gold market experienced considerable price fluctuations driven by persistent inflation, changes in monetary policy by major central banks, economic uncertainty, and increasing geopolitical tensions. These conditions intensified public interest in gold as a safe-haven investment and generated active discussions on social media, making this period suitable for investigating public sentiment toward gold price movements. To maintain the stability of the crawling process and comply with Platform X's access limitations, data collection was performed periodically, with each crawling session retrieving approximately 100 tweets per week. The collected tweets were stored separately in CSV format, with each file containing tweet metadata and textual content.

After the crawling process was completed, all CSV files were merged into a single dataset to facilitate subsequent processing and analysis. Preliminary inspection was then conducted to verify data consistency, identify duplicate records, and ensure the completeness of the collected data before entering the preprocessing stage. A total of 7,283 tweets were successfully collected during the crawling process.

Data Preprocessing

The raw tweets collected from Platform X were preprocessed to improve data quality and prepare the text for sentiment analysis. Initially, the dataset consisted of 7.283 tweets. A series of preprocessing techniques was then applied to remove noise, standardize text representation, and reduce data inconsistency. After completing the preprocessing stage, the number of usable tweets decreased to 4.429, which were subsequently used in the sentiment labeling process. The overall preprocessing workflow is illustrated in **Figure 2**.

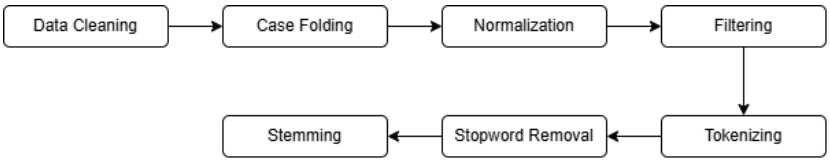

Figure 2. Preprocessing workflow

1) Data Cleaning: Data cleaning was conducted to remove irrelevant elements that do not contribute to sentiment classification, including URLs, user mentions(@username), hashtags(#), punctuation marks, emojis, numbers, and other special characters [28], [29].
2) Case Folding: Case folding converts all characters into lowercase to eliminate differences caused solely by letter capitalization [28].
3) Normalization: Normalization transforms non-standard words, abbreviations, internet slang, and informal expressions into their standard Indonesian forms [30].
4) Filtering: Filtering removes tweets or textual components that do not satisfy the predefined preprocessing criteria. [29].
5) Tokenization: Tokenization divides each tweet into individual word tokens, allowing every word to be processed independently during feature extraction and sentiment classification [14].
6) Stopword Removal: Stopword removal eliminates commonly occurring words that carry little semantic information, such as conjunctions, articles, and prepositions [28].
7) Stemming: Stemming converts inflected and derived words into their root forms, thereby reducing lexical variation and ensuring that words with similar meanings are represented consistently [31].

Weak Supervision (Lexicon-Based Labeling)

Weak supervision is an automatic labeling approach that generates approximate sentiment labels without requiring manually annotated datasets. In this study, sentiment labels were generated using a lexicon-based approach by matching each preprocessed tweet with an Indonesian sentiment lexicon containing positive and negative polarity words. The overall sentiment of each tweet was determined based on the accumulated polarity score and classified into positive, negative, or neutral classes. This approach was selected because it enables efficient large-scale sentiment labeling while maintaining transparent and interpretable labeling rules, making it widely adopted in financial sentiment analysis [25], [5], [6]. Weak supervision has also been widely studied as an efficient alternative to manual annotation in large-scale machine learning applications, particularly when obtaining gold-standard labels is expensive and time-consuming [26], [17].

Manual Validation (Label Quality Assessment)

To evaluate the quality of the weak labels, manual validation was performed on 354 tweets selected using stratified sampling to ensure proportional representation of each sentiment class. Each tweet was independently annotated by two annotators based on its contextual meaning. The agreement between the annotators was measured using Cohen's Kappa coefficient [32], which evaluates inter-annotator agreement while correcting for chance agreement. According to [33], Kappa values above 0.81 indicate almost perfect agreement. Disagreements were subsequently resolved through discussion to obtain consensus labels. The consensus labels were then compared with the lexicon-generated labels using a confusion matrix and evaluation metrics including accuracy, precision, recall, and F1-score. This validation was conducted to assess the reliability of the weak supervision process before the labeled dataset was used for feature representation and model training [15], [34], [35]

Data Splitting

After the feature representation process was completed, the labeled dataset was divided into training and testing sets using the holdout method with an 80:20 ratio. Approximately 80% of the data were allocated for training, while the remaining 20% were reserved for testing to evaluate the models on previously unseen data. This approach was adopted to provide an unbiased estimation of the classification models' generalization performance [34], [36].

The dataset was partitioned using the train_test_split function from the scikit-learn library with a fixed random_state to ensure the reproducibility of the experimental results. Furthermore, stratified sampling was applied based on the sentiment labels so that the proportions of the positive, negative, and neutral classes were preserved in both the training and testing sets. Maintaining the original class distribution helps reduce the bias caused by class imbalance and provides a more representative evaluation of classification performance [34], [35]. The same training and testing partitions were consistently used for all experiments involving Naïve Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and IndoBERT, thereby ensuring a fair comparison among the evaluated models.

Model Development

This stage focuses on constructing sentiment classification models using both conventional machine learning algorithms and a transformer-based language model. Two modeling strategies were implemented in this study. The first strategy employed TF-IDF feature representation combined with three conventional machine learning algorithms, namely Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The second strategy utilized IndoBERT, a transformer-based language model pre-trained on large-scale Indonesian corpora, which directly performs contextual representation learning and sentiment classification through supervised fine-tuning. This design enables a comprehensive comparison between conventional statistical text representation and contextual language representation for Indonesian financial sentiment analysis.

1) TF-IDF Representation

For the conventional machine learning approach, preprocessed tweets were first transformed into numerical vectors using the TF-IDF weighting scheme. Unigram and bigram features were adopted to capture both individual words and adjacent word combinations, while the *min_df* and *max_df* parameters were applied to remove extremely rare and overly frequent terms. The resulting sparse vectors were subsequently used as input features for the Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine classifiers [28], [13].

2) Naïve Bayes

Naïve Bayes (NB) is a probabilistic classification algorithm based on Bayes' theorem that estimates the posterior probability of each class from the observed features [16]. Because of its computational efficiency and its ability to handle high-dimensional sparse text data, the Multinomial Naïve Bayes classifier was employed to classify TF-IDF vectors into positive, negative, and neutral sentiment classes [13], [17].

$$P(C \mid X) = \frac{P(X|C)P(C)}{P(X)}$$

where $P(C \mid X)$ denotes the posterior probability of class Cgiven feature vector X, $P(X \mid C)$ represents the likelihood of observing feature vector Xin class C, $P(C)$ is the prior probability of class C, and $P(X)$ is the evidence.

3) K-Nearest Neighbor

K-Nearest Neighbor (KNN) classifies a document according to the majority class among its nearest neighbors in the feature space [17]. Since TF-IDF produces sparse and high-dimensional vectors, cosine similarity was employed to measure document similarity because it is more appropriate for text representation than Euclidean distance [37]. Based on preliminary experiments using the training dataset, the optimal parameter was determined as k = 7.

$$Cosine(x, y) = \frac{x \cdot y}{|| x || || y ||}$$

where x and y denote the TF-IDF vectors of two documents. A cosine similarity value closer to 1 indicates a higher degree of similarity between the documents.

4) Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm that constructs an optimal separating hyperplane by maximizing the margin between different classes [38]. Because TF-IDF vectors are sparse and high-dimensional, this study employed a linear kernel, which provides efficient computation while maintaining strong classification performance for text classification tasks [13]. The LinearSVC implementation from the scikit-learn library was adopted to classify tweets into positive, negative, and neutral sentiment categories.

5) IndoBERT Fine-tuning

Besides conventional machine learning models, this study also employed IndoBERT, a transformer-based language model pre-trained on large-scale Indonesian corpora using the Bidirectional Encoder Representations from Transformers (BERT) architecture [18]. Unlike TF-IDF-based approaches that represent text based on term frequency, IndoBERT generates contextual representations in which the meaning of each word is interpreted according to its surrounding context. The pre-trained IndoBERT model was initialized using publicly available weights and subsequently fine-tuned on the labeled tweet dataset to perform three-class sentiment classification. During the fine-tuning process, the model simultaneously learns contextual semantic representations and optimizes the classification layer for the financial sentiment analysis task.

Performance Evaluation

The performance of each sentiment classification model was evaluated using a confusion matrix, which summarizes the comparison between predicted labels and the corresponding ground-truth labels. Based on the confusion matrix, several evaluation metrics, including Accuracy, Precision, Recall, F1-score, and Macro F1-score, were calculated to assess the classification performance [34]. Since the sentiment classes in this study are imbalanced, Macro F1-score was emphasized because it evaluates all classes equally regardless of their distribution [35]).

1) Accuracy

Accuracy measures the proportion of correctly classified instances among all observations and provides an overall assessment of classification performance [34].

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

where TP denotes True Positive, TN True Negative, FP False Positive, and FN False Negative.

2) Precision

Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive. This metric reflects the classifier's ability to minimize false positive predictions [34].

$$Precision = \frac{TP}{TP + FP}$$

3) Recall

Recall measures the proportion of actual positive instances that are correctly identified by the classifier. A higher Recall indicates a better ability to detect positive samples [34].

$$Recall = \frac{TP}{TP + FN}$$

4) F1-score

F1-score is the harmonic mean of Precision and Recall and provides a balanced evaluation between these two metrics, particularly when the dataset is imbalanced [35].

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5) Macro F1-score

Macro F1-score is computed as the arithmetic mean of the F1-score obtained for each sentiment class. Unlike Accuracy, Macro F1 gives equal importance to all classes and is therefore more appropriate for evaluating imbalanced multiclass classification problems [35].

$$Macro\ F1 = \frac{1}{N}\sum_{i=1}^{N} F1i$$

where Ndenotes the number of sentiment classes.

6) Cohen's Kappa

To assess the reliability of manual annotation, Cohen's Kappa coefficient was employed to measure the agreement between two annotators while correcting for agreement occurring by chance [32]. According to [33], Kappa values closer to 1 indicate stronger agreement, whereas values near 0 indicate agreement comparable to chance.

$$\kappa = \frac{Po - Pe}{1 - Pe}$$

where P_o represents the observed agreement between annotators and P_e denotes the expected agreement by chance. A higher Kappa value indicates stronger annotation consistency.

3. Results and Discussion

Data Collection

A total of 7,283 Indonesian-language tweets related to gold prices were successfully collected from Platform X using the keyword "harga emas" during the January 2023–December 2025 observation period. The collected dataset contained several metadata attributes, including *conversation_id_str*, *created_at*, *favorite_count*, *full_text*, *reply_count*, *retweet_count*, *username*, and *location*. However, only the full_text attribute was utilized in this study because it contains the textual opinions required for sentiment analysis.

Following the initial data collection, duplicate records, empty tweets, and invalid textual entries were removed to improve data quality. After this filtering process, 4,429 tweets remained and were subsequently used in the weak supervision labeling, manual validation, model development, and performance evaluation stages.

Data Preprocessing

Following the data collection stage, the dataset underwent several preprocessing steps to improve data quality before sentiment labeling and model development. The preprocessing pipeline consisted of data cleaning, case folding, normalization, tokenization, stopword removal, and stemming, as described in the proposed methodology. During this stage, duplicate records and empty tweets were removed, reducing the dataset from 7,283 collected tweets to 4,429 valid tweets. The resulting dataset was subsequently used for the weak supervision labeling process and sentiment classification experiments.

Table I. Data Cleaning

Before	After
@idextratime Beli sekarang eti harga emas akan naik terus. Apalagi dijaman perang ekonomi seperti sekarang. Kalau udah beli emas jgn lupa dicatat biar keliatan untung ruginya https://t.co/mrLl2bVaN2	*idextratime Beli sekarang eti harga emas akan naik terus Apalagi dijaman perang ekonomi seperti sekarang Kalau udah beli emas jgn lupa dicatat biar keliatan untung ruginya*

Table 1 presents the results of the data cleaning process. URLs, hashtags, mentions, punctuation marks, and other irrelevant characters were removed to reduce textual noise while preserving the semantic content of each tweet.

Table 2. Case Folding

Before	After
idextratime Beli sekarang eti harga emas akan naik terus Apalagi dijaman perang ekonomi seperti sekarang Kalau udah beli emas jgn lupa dicatat biar keliatan untung ruginya	*idextratime beli sekarang eti harga emas akan naik terus apalagi dijaman perang ekonomi seperti sekarang kalau udah beli emas jgn lupa dicatat biar keliatan untung ruginya*

Table 2 illustrates the case folding process, in which all characters were converted to lowercase to ensure consistent word representation throughout the dataset.

Table 3. Normalization

Before	After
idextratime beli sekarang eti harga emas akan naik terus apalagi dijaman perang ekonomi seperti sekarang kalau udah beli emas jgn lupa dicatat biar keliatan untung ruginya	*idextratime beli sekarang eti harga emas akan naik terus apalagi dijaman perang ekonomi seperti sekarang kalau sudah beli emas jangan lupa dicatat biar kelihatan untung ruginya*

Table 3 presents the normalization results. In the examples shown, no lexical changes were observed because the tweets already contained standardized vocabulary after the cleaning process.

Table 4. Tokenizing

Before	After
idextratime beli sekarang eti harga emas akan naik terus apalagi dijaman perang ekonomi seperti sekarang kalau sudah beli emas jangan lupa dicatat biar kelihatan untung ruginya	*['idextratime', 'beli', 'sekarang', 'eti', 'harga', 'emas', 'akan', 'naik', 'terus', 'apalagi', 'dijaman', 'perang', 'ekonomi', 'seperti', 'sekarang', 'kalau', 'sudah', 'beli', 'emas', 'jangan', 'lupa', 'dicatat', 'biar', 'kelihatan', 'untung', 'ruginya']*

Table 4 shows the tokenization process, where each tweet was segmented into individual word tokens to facilitate subsequent text processing and feature extraction.

Table 5. Stopword Removal

Before	After
['idextratime', 'beli', 'sekarang', 'eti', 'harga', 'emas', 'akan', 'naik', 'terus', 'apalagi', 'dijaman', 'perang', 'ekonomi', 'seperti', 'sekarang', 'kalau', 'sudah', 'beli', 'emas', 'jangan', 'lupa', 'dicatat', 'biar', 'kelihatan', 'untung', 'ruginya']	*['idextratime', 'beli', 'eti', 'harga', 'emas', 'dijaman', 'perang', 'ekonomi', 'beli', 'emas', 'lupa', 'dicatat', 'biar', 'untung', 'ruginya']*

Table 5 illustrates the stopword removal process, in which frequently occurring words with limited semantic contribution were removed while preserving sentiment-bearing terms.

Table 6. Stemming

Before	After
['idextratime', 'beli', 'eti', 'harga', 'emas', 'dijaman', 'perang', 'ekonomi', 'beli', 'emas', 'lupa', 'dicatat', 'biar', 'untung', 'ruginya']	*idextratime beli eti harga emas jam perang ekonomi beli emas lupa catat biar untung rugi*

Table 6 presents the stemming process, where inflected words were transformed into their root forms to reduce vocabulary variations and improve feature consistency for sentiment classification.

Weak Supervision (Lexicon Labeling)

After preprocessing, sentiment labels were automatically generated using a lexicon-based weak supervision approach. A domain-specific sentiment lexicon was developed to represent the characteristics of Indonesian discussions related to gold prices. The lexicon consists of positive terms (e.g., *naik*, *untung*, *bullish*, and *profit*) and

negative terms (e.g., *turun*, *anjlok*, *koreksi*, and *crash*). The sentiment polarity of each tweet was determined by comparing the cumulative number of positive and negative terms. Tweets with a higher positive score were assigned a positive label, those with a higher negative score were assigned a negative label, while tweets with equal polarity scores were labeled as neutral. This weak supervision strategy enabled efficient annotation of the entire dataset without requiring manual labeling for all 4.429 tweets. Examples of the generated sentiment labels are presented in **Table 7**.

Table 7. Examples of Lexicon-Based Sentiment Labels

Text	Sentiment
you brik my heart harga emas naik terus nyesel ga beli makanya beli dari sekarang cuss order	Positif
sekejam itu ya harga emas kalau aku beli 3 tahun lalu dapet 2pcs masih sisa duitnya WKWKWK. per hari ini dah 4jt ajib	Positif
eh harga emas turun today	Negatif

Manual Validation (Label Quality Assessment

To assess the quality of the sentiment labels generated through weak supervision, a manual validation process was conducted using a stratified sample of 354 tweets. The sample was independently annotated by two annotators according to the predefined sentiment criteria for gold price discussions. The agreement between annotators was measured using Cohen's Kappa, resulting in a value of 0.8718, which indicates almost perfect agreement. This result demonstrates that the manual annotation process was highly consistent and provides a reliable reference for evaluating the quality of the automatically generated labels.

After resolving annotation disagreements through consensus, the final human labels were compared with the labels generated by the lexicon-based approach. **Table 8** summarizes the performance of the weak supervision labeling process.

Table 8. Performance of Lexicon-Based Weak Supervision

Metric	Value
Cohen's Kappa	0.8718
Accuracy	70.62%
Precision	70.94%
Recall	70.62%
F1-score	70.48%

As shown in **Table 8**, the lexicon-based labeling approach achieved an accuracy of 70.62%, with a precision of 70.94%, recall of 70.62%, and F1-score of 70.48% when compared with the final human annotations. These findings indicate that the proposed lexicon provides provides sufficiently reliable labels for weakly supervised learning rather than fully supervised learning. However, several disagreements remain due to the inability of lexicon-based methods to capture contextual meanings and ambiguous financial expressions. Therefore, the validated dataset was subsequently used to train and evaluate the Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, and IndoBERT models in the subsequent experiments.

Model Development

1) Naïve Bayes
   The Naïve Bayes model was implemented using the Multinomial Naïve Bayes classifier available in the scikit-learn library. The model was trained using TF-IDF unigram and bigram features extracted from the preprocessed tweets. Hyperparameter tuning was conducted using GridSearchCV with five-fold cross-validation, resulting in

the optimal smoothing parameter of **α = 0.1**. The optimized model was subsequently trained using the entire training dataset and evaluated on the testing dataset.

2) K-Nearest Neighbor
The K-Nearest Neighbor model was developed using TF-IDF feature vectors with cosine similarity as the distance metric to measure document similarity. Hyperparameter optimization was performed using GridSearchCV with five-fold cross-validation by evaluating several values of *k* and weighting strategies. The optimal configuration consisted of k = 7, distance weighting, and cosine similarity, which was then employed to classify the testing data.
3) Support Vector Machine
The Support Vector Machine classifier was implemented using the Linear Support Vector Classification algorithm. Hyperparameter tuning was carried out using GridSearchCV with five-fold cross-validation by evaluating different values of the regularization parameter C. The best performance was achieved using C = 1, which was subsequently adopted to train the final classification model on the TF-IDF feature vectors.
4) IndoBERT
The transformer-based model was implemented using the pre-trained IndoBERT Base P1 model provided by IndoBenchmark. Unlike the conventional classifiers, IndoBERT directly processed the preprocessed tweets through a tokenizer to generate contextual embeddings. The model was fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 16, and three training epochs on the weakly supervised sentiment labels. During training, the average training loss consistently decreased from 0.3332 in the first epoch to 0.0788 in the second epoch and 0.0330 in the third epoch, indicating stable convergence of the fine-tuning process before evaluation on the testing dataset.

Performance Evaluation

To evaluate the effectiveness of the proposed sentiment classification models, two complementary experiments were conducted. First, four classification models, namely Naïve Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and IndoBERT, were evaluated using the weakly supervised sentiment labels generated through the lexicon-based labeling approach. Second, to assess the reliability of the weak supervision strategy, the best-performing model was further evaluated using a manually annotated dataset as a gold-standard reference. The evaluation employed four performance metrics, namely accuracy, precision, recall, and F1-score, which were derived from the confusion matrix.

Table 9. Performance Comparison on Weakly Supervised Labels

Model	Accuracy	Precision	Recall	Macro F1	Weighted F1
Naïve Bayes	80.02	79.96	80.02	77.05	79.71
K-Nearest Neighbor	76.98	77.38	76.98	73.13	76.46
Support Vector Machine	92.33	92.32	92.33	91.64	92.32
IndoBERT	98.31	98.31	98.31	98.16	98.31

Table 9 shows that IndoBERT achieved the highest performance across all evaluation metrics, followed by Support Vector Machine, Naïve Bayes, and K-Nearest Neighbor. In particular, IndoBERT obtained a macro F1-score of 98.16%, substantially outperforming the conventional machine learning models. Since the sentiment classes are not perfectly balanced, macro F1-score provides a more representative evaluation than accuracy because it assigns equal importance to each class regardless of its frequency. The results demonstrate that contextual language representation learned 9 IndoBERT is considerably more effective than TF-IDF-based statistical representations for Indonesian gold price sentiment classification.

1) Naïve Bayes
The Naïve Bayes classifier achieved an overall accuracy of 80.02%, with a weighted F1-score of 79.71% and a macro F1-score of 77.05%.

Table 10. Classification Report of Naïve Bayes

Class	Precision	Recall	F1-Score	Support
Negative	0.7892	0.7228	0.7545	202
Neutral	0.8050	0.8824	0.8419	510
Positive	0.7958	0.6494	0.7152	174
Macro Avg	0.7967	0.7515	0.7705	886
Weighted Avg	0.7996	0.8002	0.7971	886

The model demonstrated the highest recall for the neutral class (88.24%), indicating that neutral tweets were identified more accurately than positive and negative tweets. However, the positive class achieved the lowest recall (64.94%), suggesting that a considerable number of positive tweets were misclassified, primarily as neutral. This indicates that although Naïve Bayes performs reasonably well on sparse TF-IDF representations, its conditional independence assumption limits its ability to distinguish sentiment classes with overlapping vocabularies.

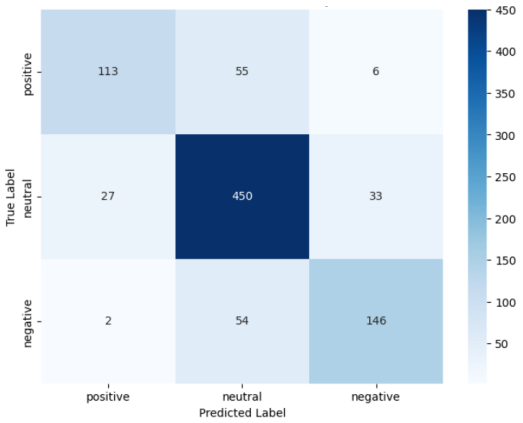

Figure 3. Confusion Matrix of Naïve Bayes

2) K-Nearest Neighbor

The K-Nearest Neighbor classifier produced an overall accuracy of 76.98%, a weighted F1-score of 76.46%, and a macro F1-score of 73.13%, representing the lowest performance among the evaluated models.

Table 11. Classification Report of K-Nearest Neighbor

Class	Precision	Recall	F1-Score	Support
Negative	0.7000	0.6584	0.6786	202
Neutral	0.7751	0.8784	0.8235	510
Positive	0.8559	0.5805	0.6918	174
Macro Avg	0.7770	0.7058	0.7313	886
Weighted Avg	0.7738	0.7698	0.7646	886

Similar to Naïve Bayes, KNN achieved its highest recall on the neutral class (**87.84%**). However, the recall for the positive class decreased to **58.05%**, indicating that many positive tweets were incorrectly assigned to the neutral category. This performance suggests that neighborhood-based classification using cosine similarity

is less effective when sentiment classes share similar TF-IDF representations, making class boundaries more difficult to distinguish.

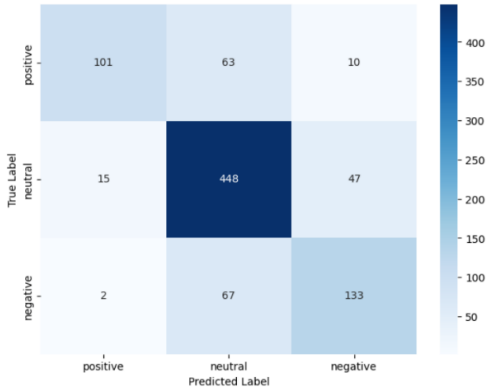

Figure 4. Confusion Matrix of K-Nearest Neighbor

3) Support Vector Machine

Support Vector Machine achieved an overall accuracy of 92.33%, with a weighted F1-score of 92.32% and a macro F1-score of 91.64%.

Table 12. Classification Report of Support Vector Machine

Class	Precision	Recall	F1-Score	Support
Negative	0.9296	0.9158	0.9227	202
Neutral	0.9283	0.9392	0.9337	510
Positive	0.9006	0.8851	0.8928	174
Macro Avg	0.9195	0.9134	0.9164	886
Weighted Avg	0.9232	0.9233	0.9232	886

Compared with Naïve Bayes and K-Nearest Neighbor, SVM demonstrated consistently high performance across all sentiment classes. The confusion matrix shows only a small number of misclassified instances, indicating that the linear decision boundary effectively separates high-dimensional TF-IDF feature vectors. These findings confirm that SVM is highly suitable for sentiment classification tasks involving sparse textual features.

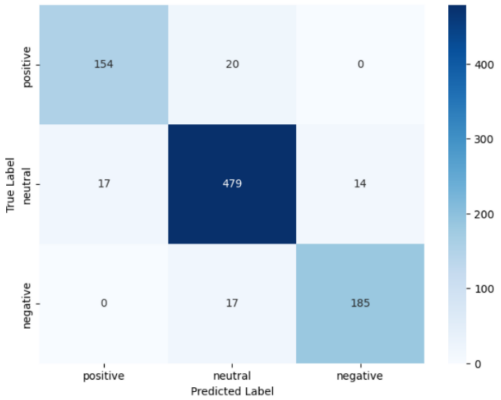

Figure 5. Confusion Matrix of Support Vector Machine

4) IndoBERT

IndoBERT achieved the best performance among all evaluated models, obtaining an accuracy, precision, recall, and weighted F1-score of 98.31%, with a macro F1-score of 98.16%.

Table 13. Classification Report of IndoBERT

Class	Precision	Recall	F1-Score	Support
Negative	1.0000	0.9901	0.9950	202
Neutral	0.9824	0.9863	0.9843	510
Positive	0.9655	0.9655	0.9655	174
Macro Avg	0.9826	0.9806	0.9816	886
Weighted Avg	0.9831	0.9831	0.9831	886

Compared with the conventional machine learning models, IndoBERT benefited from contextual language representations that capture semantic relationships beyond individual word frequencies. Unlike TF-IDF-based approaches, IndoBERT considers the surrounding context of each word, enabling the model to better interpret ambiguous financial expressions. This capability is particularly beneficial for weakly supervised datasets, where lexicon-generated labels may contain contextual inconsistencies that cannot be captured by simple keyword matching. Consequently, IndoBERT produced substantially fewer misclassifications across all sentiment classes and achieved the highest Macro F1-score among the evaluated models.

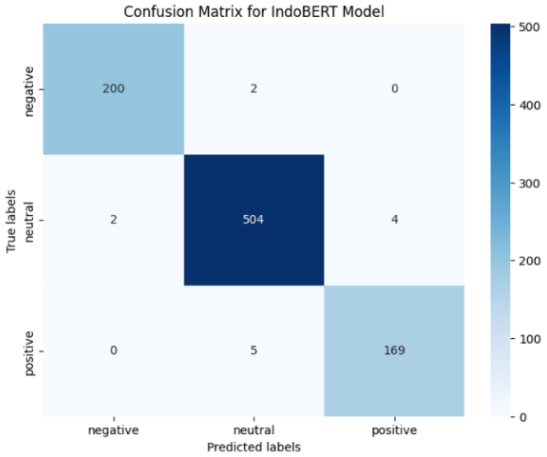

Figure 6. Confusion Matrix of IndoBERT

5) Error Analysis

Although the overall classification performance was relatively high, several misclassifications were observed across the evaluated models. These errors mainly occurred because financial sentiment expressed in social media often contains implicit meanings, contextual interpretations, and mixed sentiment signals that are difficult to capture using lexical and statistical features alone. Therefore, further analysis was conducted to identify the underlying causes of several representative classification errors.

In the Naïve Bayes model, one example of misclassification occurred for the tweet “harga emas gadai diskon borong selagi murah”, which was manually labeled as positive but predicted as neutral. From an investment perspective, expressions such as *diskon*, *borong*, and *murah* indicate a favorable buying opportunity and therefore reflect positive sentiment. However, Naïve Bayes relies heavily on word occurrence probabilities and assumes feature independence. Since these terms may appear less frequently than more common neutral terms such as *harga* and *emas*, the model failed to recognize the positive intent and assigned the tweet to the neutral class. A similar limitation was observed in the K-Nearest Neighbor model for the tweet

"harga emas global rebound emas antam stabil februari", which was labeled as positive but classified as negative. Although the term *rebound* generally indicates positive market recovery, KNN determines the class based on neighboring documents in the TF-IDF feature space. As a result, the tweet was likely surrounded by negative instances with similar lexical patterns, causing the model to assign an incorrect label.

For Support Vector Machine, a representative error occurred in the tweet "tembus rekor aprilmei harga emas dunia anjlok", which was manually labeled as neutral but predicted as positive. This tweet contains contradictory sentiment indicators. The phrase *tembus rekor* is typically associated with positive market performance, whereas *anjlok* indicates a negative price movement. Because SVM constructs a decision boundary using weighted textual features, the positive influence of *tembus rekor* likely outweighed the negative contribution of *anjlok*, resulting in a positive prediction. These examples demonstrate that conventional machine learning models struggle to interpret implicit sentiment, contextual financial expressions, and conflicting sentiment signals.

These findings also suggest that a portion of the classification error may originate from the weak supervision labeling process itself. Financial sentiment often depends on investor perspective and contextual interpretation, making certain tweets difficult to label consistently even for human annotators. For example, a decline in gold prices may be perceived negatively by current investors but positively by prospective buyers seeking lower purchase prices. Consequently, label noise introduced during lexicon-based annotation may propagate into the training data and affect downstream classification performance, particularly for conventional machine learning models that rely heavily on surface lexical features.

In contrast, IndoBERT produced substantially fewer classification errors because its contextual language representations capture semantic relationships beyond word frequency and consider surrounding linguistic context. This capability enables the model to better interpret financial expressions whose sentiment polarity depends on context rather than individual keywords, resulting in more accurate sentiment classification for Indonesian gold price discussions.

4. Conclusion

This study proposed a weakly supervised sentiment classification framework for analyzing public sentiment toward gold prices on the X platform by combining lexicon-based labeling with conventional machine learning algorithms and a transformer-based language model. A total of 7.283 tweets were collected using the keyword "harga emas". After preprocessing, duplicate removal, and filtering, 4.429 tweets were retained for sentiment labeling and model development. Weak supervision was implemented using a domain-specific sentiment lexicon, enabling efficient automatic annotation without requiring manual labeling of the entire dataset.

The experimental results demonstrate that IndoBERT consistently outperformed the conventional machine learning models. IndoBERT achieved the highest performance with an accuracy of 98.31% and a macro F1-score of 98.16%, followed by Support Vector Machine, Naïve Bayes, and K-Nearest Neighbor. These findings indicate that contextual language representations learned by transformer-based models provide superior capability in capturing semantic information compared with TF-IDF-based representations used by conventional machine learning algorithms for Indonesian gold price sentiment classification.

To evaluate the reliability of the weak supervision strategy, the automatically generated sentiment labels were validated against manually annotated data. The manual annotation process achieved a Cohen's Kappa value of 0.8718, indicating almost perfect inter-annotator agreement. When compared with the final human annotations, the lexicon-based labeling achieved an accuracy of 70.62%, suggesting that although the proposed lexicon provides reasonably reliable labels, a certain degree of labeling noise remains unavoidable. Furthermore, when the best-performing model (IndoBERT) was evaluated using the manually annotated dataset as a gold-standard reference, the accuracy decreased to 69.49%. This result suggests that although IndoBERT effectively learned the automatically generated labels, its performance remained constrained by the quality of the weak supervision process. Therefore, the findings of this study

should be interpreted within the context of weakly supervised sentiment classification rather than a fully supervised learning framework.

Overall, this study demonstrates that weak supervision can serve as an efficient and scalable alternative for sentiment annotation in large-scale social media datasets, particularly when manually labeled data are limited. However, the quality of automatically generated labels remains a critical factor affecting downstream classification performance, especially in financial-domain texts where sentiment may depend on contextual interpretation and investor perspective. Future research should expand the manually annotated dataset to establish a stronger gold-standard benchmark, enrich the domain-specific financial sentiment lexicon, and investigate more advanced weak supervision techniques or recent Indonesian transformer models to reduce labeling noise and further improve sentiment classification performance. In addition, incorporating hybrid labeling strategies that combine lexicon-based methods with human annotation may improve label quality and enhance model generalization in financial sentiment analysis.

References:

[1] T. I. Tanin, A. Sarker, R. Brooks, and H. X. Do, "Does oil impact gold during COVID-19 and three other recent crises?," *Energy Economics*, vol. 108, p. 105938, Apr. 2022, https://doi.org/10.1016/j.eneco.2022.105938.

[2] A. A. Sikiru and A. A. Salisu, "Assessing the hedging potential of gold and other precious metals against uncertainty due to epidemics and pandemics," *Quality & Quantity*, vol. 56, no. 4, pp. 2199–2214, Aug. 2022, https://doi.org/10.1007/s11135-021-01214-7.

[3] I. N. Agustin, "Can Social Responsible Investment and Gold be a Good Diversifier for Indonesia Sharia Investors?," *Jurnal Keuangan dan Perbankan*, vol. 26, no. 1, pp. 146–160, Feb. 2022, https://doi.org/10.26905/jkdp.v26i1.6929.

[4] D. G. Baur and T. K. McDermott, "Is gold a safe haven? International evidence," *Journal of Banking & Finance*, vol. 34, no. 8, pp. 1886–1898, Aug. 2010, https://doi.org/10.1016/j.jbankfin.2009.12.008.

[5] A. Algaba, D. Ardia, K. Bluteau, S. Borms, and K. Boudt, "Econometrics Meets Sentiment: An Overview Of Methodology And Applications," *Journal of Economic Surveys*, vol. 34, no. 3, pp. 512–547, Jul. 2020, https://doi.org/10.1111/joes.12370.

[6] L. Barbaglia, S. Consoli, S. Manzan, L. Tiozzo Pezzoli, and E. Tosetti, "Sentiment analysis of economic text: A lexicon-based approach," *Economic Inquiry*, vol. 63, no. 1, pp. 125–143, Jan. 2025, https://doi.org/10.1111/ecin.13264.

[7] M. Baker and J. Wurgler, "Investor Sentiment in the Stock Market," *Journal of Economic Perspectives*, vol. 21, no. 2, pp. 129–151, Apr. 2007, https://doi.org/10.1257/jep.21.2.129.

[8] X. Huang and H. Song, "Investor Sentiment Combined with Multisource Information to Predict Stock Prices: An Analysis of China's A-Share Market," *Scientific Programming*, vol. 2021, pp. 1–9, Dec. 2021, https://doi.org/10.1155/2021/9094032.

[9] E. Blankespoor, "Firm communication and investor response: A framework and discussion integrating social media," *Accounting, Organizations and Society*, vol. 68–69, pp. 80–87, Jul. 2018, https://doi.org/10.1016/j.aos.2018.03.009.

[10] A. Giachanou and F. Crestani, "Like It or Not," *ACM Computing Surveys*, vol. 49, no. 2, pp. 1–41, Jun. 2017, https://doi.org/10.1145/2938640.

[11] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 1003–1015, Aug. 2021, https://doi.org/10.1109/TCSS.2021.3051189.

[12] Simon Kemp, "Digital 2025: Indonesia," *DataReportal – Global Digital Insights*, Feb. 25, 2025. .

[13] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019, https://doi.org/10.3390/info10040150.

[14] Daniel Jurafsky and James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd ed. 2026.

[15] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, May 2012, https://doi.org/10.2200/S00416ED1V01Y201204HLT016.

[16] A. F. A. H. Alnuaimi and T. H. K. Albaldawi, "An overview of machine learning classification techniques," *BIO Web of Conferences*, vol. 97, p. 00133, Apr. 2024, https://doi.org/10.1051/bioconf/20249700133.

[17] Y. Zhang, Q. Li, and Y. Xin, "Research on eight machine learning algorithms applicability on different characteristics data sets in medical classification tasks," *Frontiers in Computational Neuroscience*, vol. 18, Jan. 2024, https://doi.org/10.3389/fncom.2024.1345575.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.

[19] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857, https://doi.org/10.18653/v1/2020.aacl-main.85.

[20] S. H. Ramadhani and M. I. Wahyudin, "Analisis Sentimen Terhadap Vaksinasi Astra Zeneca pada Twitter Menggunakan Metode Naïve Bayes dan K-NN," *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 6, no. 4, pp. 526–534, Feb. 2022, https://doi.org/10.35870/jtik.v6i4.530.

[21] N. Wulandari, Y. Cahyana, R. Rahmat, and H. Hikmayanti, "Sentiment Analysis on the Relocation of the National Capital (IKN) on Social Media X Using Naive Bayes and K-Nearest Neighbor (KNN) Methods," *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 724–731, Jun. 2025, https://doi.org/10.30871/jaic.v9i3.9552.

[22] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *Jurnal KomtekInfo*, pp. 1–7, Jan. 2023, https://doi.org/10.35134/komtekinfo.v10i1.330.

[23] E. Hokijuliandy, H. Napitupulu, and Firdaniza, "Application of SVM and Chi-Square Feature Selection for Sentiment Analysis of Indonesia's National Health Insurance Mobile Application," *Mathematics*, vol. 11, no. 17, p. 3765, Sep. 2023, https://doi.org/10.3390/math11173765.

[24] N. F. Rozy, N. Amini, N. Hakiem, and S. H. Afrizal, "Analysis of Multi-Class Sentiment on Indonesian Twitter Using Support Vector Machine Classification Algorithm with Particle Swarm Optimization," in *Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence*, Oct. 2023, pp. 62–67, https://doi.org/10.1145/3633598.3633609.

[25] A. M. van der Veen and E. Bleich, "The advantages of lexicon-based sentiment analysis in an age of machine learning," *PLOS ONE*, vol. 20, no. 1, p. e0313092, Jan. 2025, https://doi.org/10.1371/journal.pone.0313092.

[26] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel," *Proceedings of the VLDB Endowment*, vol. 11, no. 3, pp. 269–282, Nov. 2017, https://doi.org/10.14778/3157794.3157797.

[27] B. Frenay and M. Verleysen, "Classification in the Presence of Label Noise: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, May 2014, https://doi.org/10.1109/TNNLS.2013.2292894.

[28] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[29] R. Feldman and J. Sanger, *The Text Mining Handbook*. Cambridge University Press, 2006.

[30] Bo Han and Timothy Baldwin, *Lexical Normalisation of Short Text Messages: Makn Sens a #twitter*. Portland, Oregon, USA: Association for Computational Linguistics, 2011.

[31] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, Mar. 1980, https://doi.org/10.1108/eb046814.

[32] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, https://doi.org/10.1177/001316446002000104.

[33] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, https://doi.org/10.2307/2529310.

[34] D. Berrar, "Performance Measures for Binary Classification," in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 546–560.

[35] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, https://doi.org/10.1186/s12864-019-6413-7.

[36] M. A. Pienaar and K. Naidoo, "Classification and predictive models using supervised machine learning: A conceptual review," *Southern African Journal of Critical Care*, p. e2937, May 2025, https://doi.org/10.7196/SAJCC.2025.v411.2937.

[37] C. C. Aggarwal, *Machine Learning for Text*. Cham: Springer International Publishing, 2018.

[38] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, https://doi.org/10.1007/BF00994018.