International Journal of Advances in Data and Information Systems
Vol. 7, No. 1, April 2026, pp. 292~306
ISSN: 2721-3056, DOI: 10.59395/ijadis.v7i1.1517

Hybrid Relevance and Sentiment Classification of Indonesian Gold Tweets Using Machine Learning for Market Risk Signal Extraction

Antika Zahrotul Kamalia[1], Indra[2], Arief Wibowo[3] ,Jan Everhard Riwurohi[4] , Shiza Hassan[5]
[1]Department of Informatics Engineering, Universitas Pelita Bangsa, Indonesia
[2.3,4]Department of Computer Science, Universitas Budi Luhur, Indonesia
[5]Department of Computer Science, UIT University, Pakistan

Article Info
Article history:
Received Jan 17, 2026
Revised Feb 05 2026
Accepted Feb 16, 2026

Keywords:
Gold
Classification
Sentiment
Machine Learning
Risk

ABSTRACT
This study proposes a hybrid relevance–sentiment classification framework to analyze public opinion on physical Antam gold from Indonesian Twitter data and to support exploratory market-risk signal extraction. Tweets were collected during February–November 2025, after preprocessing and text-normalized deduplication, 1,271 unique tweets were retained. The approach combines weak supervision (rule-/lexicon-based silver labels) with TF-IDF-based machine learning in two stages: (1) relevance classification to separate tweets genuinely discussing physical Antam gold from non-relevant contexts (e.g., ANTM stock/capital-market discussions), and (2) two-class sentiment classification (positive vs negative) applied to relevance-filtered tweets. Random Forest achieved the strongest relevance performance (Accuracy = 0.984; macro-F1 = 0.943; 5-fold CV macro-F1 = 0.928 ± 0.033). For sentiment classification, performance was moderate and close across models; the most stable model under cross-validation (Logistic Regression/Naive Bayes) was used for downstream aggregation. Sentiment outputs were aggregated into a monthly sentiment index for descriptive comparison with gold prices; the observed association was weak, indicating that the index is better interpreted as a risk-perception proxy rather than a direct price predictor.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Antika Zahrotul Kamalia,
Department of Informatics Engineering, Universitas Pelita Bangsa, Indonesia
Jl. Inspeksi Kalimalang No.9, Cibatu, Cikarang Sel., Kabupaten Bekasi, Jawa Barat 17530, Indonesia
Email: antika.kamalia@pelitabangsa.ac.id

1. INTRODUCTION

Gold has long been regarded as a relatively stable investment asset and a safe haven when financial markets come under stress[1], [2]. Several studies show that gold tends to preserve its value or even appreciate during periods of economic turmoil, making it an important instrument for modern portfolio diversification [3]. In Indonesia, bullion is produced by PT Aneka Tambang Tbk. (Antam gold) It is a preferred choice among retail investors due to its high authenticity, official certification, and strong liquidity in domestic buy–sell transactions [4].

In financial risk management, quantifying uncertainty in gold prices is increasingly important given volatility driven by global macroeconomic conditions, exchange rates, and shifts in investor demand [5], [6]. Value-at-Risk (VaR) is a widely used measure for estimating potential losses at a given confidence level [7]. However, VaR models that rely solely on historical price information may be limited, as gold-price volatility can be influenced by non-linear market dynamics and investor reactions that are not always captured by traditional variables [8].

With the rapid growth of social media, public sentiment has emerged as a non-traditional indicator used to complement financial analysis of asset prices and volatility [9]. Social-media

sentiment may influence or even precede market movements by reflecting investors' perceptions and emotional responses in near real time [10], [11]. For gold, public discussions about prices, queues at bullion outlets, product availability, and purchase satisfaction may provide additional signals relevant to market dynamics.

Nevertheless, leveraging Twitter data in Indonesia poses a unique challenge due to entity/topic ambiguity: "Antam" may refer to physical gold products, while "ANTM" refers to the stock ticker of PT Aneka Tambang Tbk. Many tweets contain the term "antam" but actually discuss stock trading, technical analysis, or capital-market recommendations. Without explicit disambiguation, sentiment analysis may become biased and generate misleading signals for downstream financial interpretation [12], [13].

This study proposes a weakly supervised hybrid relevance–sentiment pipeline for Indonesian Twitter that separates relevance filtering (physical Antam gold vs. ANTM stock/irrelevant topics) from sentiment modelling using rule-/lexicon-based silver labels and task-specific TF-IDF features. Tweet-level outputs are aggregated into a relevance-filtered sentiment index, positioned as an exogenous, risk-related indicator for exploratory downstream analysis; full VaR/tail-risk integration is left for future work due to the limited observation horizon. Contributions: (1) a hybrid relevance filter tailored to the Antam–ANTM ambiguity; (2) weak supervision via silver labels for relevance and sentiment; and (3) task-specific TF-IDF modelling and construction of a relevance-filtered sentiment index.

2. RESEARCH METHOD

2.1 Research Framework

This study proposes a hybrid analytical pipeline to obtain a clean, gold-specific public sentiment signal for Antam Gold from Twitter and to evaluate the potential of the resulting sentiment index to support market risk signal extraction; robust risk forecasting/VaR-oriented evaluation is left for fu-ture work. The proposed framework comprises six main stages: (1) dataset input, (2) text prepro-cessing, (3) relevance filtering, (4) sentiment labelling, (5) monthly sentiment aggregation, and (6) mar-ket risk signal extraction. Each component is designed to address the noisy, ambiguous, and un-structured nature of Twitter data, particularly in the Indonesian context, where the term "Antam" may refer to either physical gold or ANTM stock.

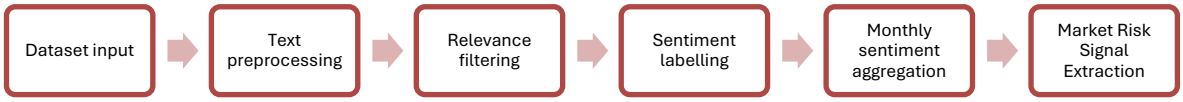


Figure 1. Research Method

This study uses a layered framework that begins with Twitter data collection via query, followed by text preprocessing to improve data quality. A hybrid approach is then applied to separate relevant from irrelevant tweets using rule-based filtering and machine learning classification. Relevant tweets are further processed for sentiment analysis using a two-stage procedure: lexi-con-based silver labels for initial polarity and a machine learning classifier for refinement. The resulting tweet-level sentiment is aggregated monthly to construct a Sentiment Index that captures shifts in public perception toward Antam Gold.

2.2 Dataset Collection and Description

Tweets were collected from X (Twitter) using tweet-harvest (Node.js) with API authentication over February–November 2025 to obtain a longer horizon for exploratory sentiment–price analysis. To reduce retrieval bias, we used a single neutral keyword set centered on Antam's physical gold (Logam Mulia), covering core product mentions and transaction/service contexts (e.g., "emas antam", "logam mulia/LM", "harga emas antam", "buyback", "butik logam mulia", "gram/batangan", "sertifikat", "cetak emas", "tabungan emas"), while explicitly excluding stock-market noise via negative keywords (e.g., "-saham", "-$ANTM", "-#ANTM", "-IHSG", "-IDX"). Importantly, affective terms were not used as retrieval triggers to avoid oversampling

explicitly emotional posts. Daily Antam gold prices were obtained from LogamMulia (PT Antam) and aligned by date with the sentiment index computed from relevance-filtered tweets. Given the prevalence of templated price updates and reposts, we removed exact duplicates based on normalized text (text_processed) before model training and evaluation; residual near-duplicate variants are treated as a limitation for future work.

2.4 Text Preprocessing

The text preprocessing stage transforms raw, unstructured tweets into standardized data ready for use in relevance and sentiment classification [14], [15]. Given Twitter's rich informal and noisy nature (URLs, mentions, emoticons, and case variations), this study implemented a series of preprocessing steps including case folding, tokenization, slang normalization (mapping informal words to formal forms using a slang-to-formal dictionary), stopword removal (an expanded NLTK Indonesian list with Twitter-specific stopwords), and stemming using Sastrawi to reduce morphological variation. After text cleaning and normalization, documents are represented numerically using TF-IDF so that machine learning models can leverage the relative weights of informative words within the corpus.
TF-IDF is calculated using the formula:

$$TF - IDF(t,d) = TF(t,d)\ x \log\left(\frac{N}{df_t}\right) \quad (1)$$

with:
TF(t,d) = frequency of occurrence of word t in document d
N = total number of documents
df_t = number of documents containing word t

In this study, we use task-specific TF–IDF representations by training two separate vectorizers for relevance and sentiment classification. The relevance TF–IDF is fitted on the full corpus (relevant + non-relevant) to capture discriminative vocabulary that separates discussions of physical Antam gold from ANTM stock/irrelevant contexts, where terms may overlap but appear in different combinations and usage patterns. The sentiment TF–IDF is fitted only on tweets predicted as relevant, so that the feature space emphasizes polarity-bearing expressions (e.g., mahal, puas, bagus, kecewa) within the physical-gold context. This separation avoids conflating two different objectives in a single representation and provides features better aligned with each task, leading to more stable and interpretable classifiers.

This comprehensive preprocessing pipeline is designed to address the unique characteristics of Indonesian Twitter data, including frequent slang use, complex word morphology, and contextual ambiguity. Applying these steps helps ensure that the textual representations fed into the machine learning models are of higher quality and, in turn, improves performance for both relevance and sentiment classification tasks [16][17].

2.5 Hybrid Relevance Classification

The relevance classification stage separates tweets that genuinely discuss Antam's physical gold (Logam Mulia) from tweets about ANTM stock/capital markets or other irrelevant topics. Because "Antam" is ambiguous in public discourse, we use a two-stage hybrid pipeline: rule-based filtering for initial screening and a machine-learning classifier for refinement. In the rule-based stage, physical-gold indicators (e.g., emas antam, logam mulia/LM, butik LM, buyback, gram, batangan, sertifikat, cetak emas) are used to tag tweets as relevant. At the same time, capital-market terms (e.g., saham ANTM, IHSG/IDX, lot, dividend, emiten, cut loss, TP/SL) indicate non-relevance. Ambiguous tokens (e.g., antam, logam, naik/turun) are not labeled directly and are deferred to the learning-based stage. The rule outputs are treated as silver labels for weak supervision, with manual annotation planned for future improvements.

Next, a TF-IDF n-gram classifier is trained to capture informal language variants and contextual cues beyond keyword rules. To reduce leakage from duplicated content, we remove exact duplicates prior to splitting and conduct a near-duplicate screening on the cleaned corpus. Since near-duplicate clusters were not substantial under our screening, we proceed with stratified

splitting; nonetheless, duplicate-aware grouping remains a limitation and is left for future improvement [18].

2.6 Hybrid Sentiment Classification

The sentiment classification stage was conducted to identify the polarity of public opinion toward Antam Gold in tweets confirmed as relevant in the previous stage. Given the rich informality, slang, and variety of expressions characteristic of Indonesians on Twitter, this study employed a two-stage hybrid approach: (1) lexicon-based sentiment labeling as an initial mapping, and (2) machine-learning sentiment classification to refine sentiment labels more accurately. This hybrid approach balances the interpretability of the lexicon and the generalizability of the machine learning model.

The lexicon-based approach is the initial stage for providing weak supervision of sentiment data [19], [20]. In the lexicon stage, each tweet is assigned a sentiment score by summing token scores: positive words receive +1, negative words −1, and out-of-lexicon words 0. Tweets are labeled positive (score > 0), negative (score < 0), or neutral/NA (score = 0). The lexicon is adapted from an Indonesian sentiment dictionary and manually augmented with common Twitter variants (e.g., slang and elongated forms such as mahalnyaaa, bagus bgt, parah sih, kecewa banget). Importantly, these lexicon-derived labels are treated as silver labels for weak supervision rather than ground truth; manual validation on a subset is planned in follow-up work. Each relevant tweet is analyzed by summing the sentiment scores of each token:

$$SentimentScore(d) = \sum_{t \in d} LexiconScore(t) \quad (2)$$

dengan:

- LexiconScore$(t) = +1$for positive words
- LexiconScore$(t) = -1$for negative words
- Words not in the lexicon are given a score of 0

Labels are assigned based on the following rules:

- Positive: SentimentScore(d) >0
- Negative: $SentimentScore(d) < 0$
- Neutral / NA: $SentimentScore(d) = 0$

Models are trained on TF-IDF n-gram features for binary polarity (positive vs. negative). Exact duplicates were removed using normalized text (text_processed) before an 80:20 stratified split (SEED = 42) and 5-fold StratifiedKFold CV (shuffle=True, random_state=42). Residual near-duplicates may still exist and are noted as a limitation for future duplicate-aware evaluation.

2.7 Sentiment Aggregation for Market Risk Signal Extraction

This stage converts tweet-level sentiment predictions into a time-aggregated sentiment indicator that can be used as a supporting, risk-related proxy signal for downstream market analysis. Because tweet sentiment is highly granular and noisy, temporal aggregation is applied to obtain a more stable indicator that is less sensitive to short-lived fluctuations[22]. Only tweets predicted as relevant by the best-performing relevance model are included, reducing contamination from ANTM stock and unrelated discussions. Each relevant tweet receives a sentiment label (positive, negative, or neutral/NA). The core index is computed using polarity tweets only (positive/negative), while neutral/NA tweets are retained for supplementary descriptive analysis. The sentiment index at time t is defined by the net-polarity formulation:

$$S_t = \frac{P_t - N_t}{T_t} \quad (3)$$

dengan:

S_t= Sentiment index on day t
P_t= Number of tweets labeled positive
N_t= Number of tweets labeled negative
$T_t = P_t + N_t$= Total tweets with polarity on that day

Where $S_t$denotes the sentiment index at time t, $P_t$is the number of positive tweets, $N_t$is the number of negative tweets, and $T_t$is the total number of polarity tweets. The index ranges from $-1$to $+1$, where positive values indicate a predominance of optimistic opinions and negative values indicate a predominance of pessimistic views. In our experiments, $S_t$is computed daily and then aggregated monthly for descriptive analysis. The resulting aggregates form the dataset for exploratory, risk-oriented analysis, such as descriptive comparison and association checks with gold price movements. In this study, the index is treated as a monitoring-oriented proxy rather than a causal or deterministic predictor; further validation and integration as an exogenous input into VaR or other tail-risk frameworks are left for future work [23], [24].

2.8 Duplicate Check and Leakage Prevention

To mitigate potential leakage from reposted or templated content, we first removed exact duplicates by matching normalized tweet text (text_processed) after basic normalization (e.g., lowercasing and URL/mention cleaning). We then conducted a near-duplicate screening on the cleaned corpus using the same normalized representation. The screening did not indicate substantial near-duplicate clusters; therefore, we did not impose additional duplicate-group constraints during splitting. Nevertheless, because exact-text matching cannot fully eliminate non-identical near-duplicates, we treat duplicate-aware (group-based) splitting as a potential improvement for future work.

2.9 Model Training and Evaluation Setup

To ensure objective evaluation, we used an 80:20 hold-out split combined with 5-fold cross-validation [25]. After preprocessing and exact-duplicate removal based on normalized text (text_processed), the final dataset contains 1,271 unique tweets, split stratified by class labels into 1,016 training and 255 test tweets using a fixed random seed (SEED = 42). Performance is reported on the held-out test set and summarized with 5-fold StratifiedKFold cross-validation performed on the training set only (shuffle = True, random_state = 42), reporting macro-F1 (mean ± std) as the primary metric, with accuracy/precision/recall as complementary measures. Confusion matrices and brief qualitative error checks are reported as an interpretability aid to inspect remaining false positives/negatives. For sentiment classification, lexicon-derived labels are treated as silver labels (weak supervision) for internal evaluation, with manual validation planned for future work. Because deduplication is performed by exact-text matching, near-duplicate variants may still exist and are treated as a limitation for future improvement.

1. Accuracy

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

- TP (True Positive): Positive data that is correctly predicted as positive.
- TN (True Negative): Negative data that is correctly predicted as negative.
- FP (False Positive): Negative data that is incorrectly predicted as positive.
- FN (False Negative): Positive data that is incorrectly predicted as negative.

2. Precision

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

3. Recall

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

4. F1-Macro

$$F1_{macro} = \frac{1}{2}(F1_{relevant} + F1_{not\ relevant}) \quad (7)$$

5. Cross-Validation (5-Fold CV)

$$CV - Mean = \frac{1}{5}\sum_{i=1}^{5} F1_i \quad (8)$$

6. Confusion Matrix

A confusion matrix is used to visualize the distribution of predictions per class, allowing model error patterns to be observed. The four-cell matrix (TP, TN, FP, FN) helps identify:

- relevant → irrelevant errors (false negatives)
- irrelevant → relevant errors (false positives)
- model bias toward the dominant class
- model weakness in ambiguous tweets

This analysis is important because errors in relevance filtering can directly degrade the quality of downstream sentiment signals.

7. Error Analysis

To understand the types of errors made by the model, error analysis is performed on a sample of incorrect predictions. These stages include:

- identifying tweets with the highest FP and FN
- analyzing language characteristics (slang, mixed topics of gold and stocks)
- detecting ambiguous words (e.g., "antam", "lm", "harga naik", "lot", "tp")
- evaluating whether errors originate from preprocessing, rule-based analysis, or the ML model

Error analysis is used to strengthen model interpretation and determine whether additional features, rule-based refinement, or data augmentation are needed [26].

2.10 Ethical Considerations

This study addresses the ethical aspects of social media data use, particularly privacy and minimizing the risk of individual identification. Data were collected from public Twitter content and analyzed at the text level and monthly aggregation, without the goal of tracking or identifying users. To maintain confidentiality, this research does not include account identification information (e.g., usernames, profile links, or user ID) in the reporting, and sample tweets are presented only as needed to interpret the results. This practice aligns with general recommendations for ethical research based on social media data, which emphasize caution regarding user privacy expectations and the potential impacts of data publication[27][28].

Furthermore, this research considers biases and limitations that can arise from a weak supervision approach. Rule-based relevance labels and lexicon-based sentiment labels are treated as silver labels rather than ground truth, potentially introducing biases toward certain linguistic varieties (e.g., slang, irony, figurative expressions) and leading to uneven errors. To maintain transparency and accountability, this study reports evaluation metrics, a confusion matrix, and conducts qualitative error analysis to identify error patterns.

3. RESULTS AND DISCUSSION

3.1. Dataset Characteristics and Descriptive Analysis

The research data were obtained through Twitter crawling using a single neutral keyword set (e.g., “emas antam”, “logam mulia”, “butik LM”) designed to capture public discussions about Antam’s physical gold (Logam Mulia) while reducing stock-market noise through stock-related exclusions (e.g., “saham”, “ANTM”, “IHSG”). The crawling process yielded 2,017 raw tweets; after duplicate removal, 1,271 unique tweets remained (746 duplicates removed). This indicates that a substantial portion of the raw stream consisted of reposted or templated content (e.g., recurring daily price updates), a pattern commonly observed in Twitter-based financial conversations. The deduplication step is therefore important to reduce bias from repetitive text patterns and to obtain a more representative sample of public discourse; moreover, duplicate control is applied prior to evaluation to reduce the risk of duplicate/near-duplicate leakage across splits.

Thematically, the discussions are dominated by terms related to price monitoring and physical-gold transactions, such as harga (price), buyback, beli/jual (buy/sell), butik logam mulia, and gram/batangan, reflecting users’ focus on retail trading and daily valuation. TF-IDF features reinforce this pattern by highlighting informative phrases (e.g., “harga hari ini”) and transactional expressions (e.g., “beli/jual”), alongside evaluative terms such as “mahal” and “turun,” suggesting

that conversations are largely reactive to price dynamics. The activity is episodic, with tweet-count spikes on particular dates that may coincide with routine price-update postings and short-lived attention shifts (descriptive observation). Overall, these characteristics motivate the next stages of hybrid relevance filtering and hybrid sentiment classification so that the sentiment index is constructed only from tweets that meaningfully represent discussions about Antam's physical gold [31].

3.2. Performance of Relevance Classification

The relevance pipeline was designed to separate tweets that genuinely discuss Antam's physical gold (Logam Mulia) from content related to ANTM stock, broader capital-market talk, or other irrelevant topics. Using the final dataset of 1,271 unique tweets. This imbalance suggests that the crawling strategy successfully captured the intended physical-gold context while substantially reducing stock-related noise; nevertheless, the rule outputs are treated as silver labels (weak supervision) rather than ground truth, and borderline/ambiguous cases are handled by the subsequent machine-learning refinement stage. For model evaluation, the dataset was split using a stratified 80:20 scheme into 1,016 training tweets and 255 test tweets, and performance was assessed using hold-out testing complemented by 5-fold cross-validation[32].

Table 1. Rule-based Filtering Outcomes

Class	Tweet
Relevant	harga emas Antam hari ini naik Rp 7.000 per gr....
Relevant	antrean panjang terjadi di Butik Emas Antam....
Unrelevant	anak emas kok apa-apa usaha sendiri...
Unrelevant	generasi unggul menuju Indonesia emas...

Although effective as an initial screen, the rule-based approach still has several limitations. First, the domain ambiguity between "Antam" (physical gold) and "ANTM" (the stock ticker) can cause stock-related tweets to be mistakenly labeled as relevant. Second, Twitter language is highly informal and often figurative, which may trigger false positives, for example, the use of "gold/emas" as a metaphor unrelated to physical-gold transactions. Third, slang variants, spelling elongations, and morphological variations in informal Indonesian can produce false negatives when relevant tweets do not match the predefined keyword list. Overall, these observations suggest that rule-based filtering is suitable for coarse relevance screening but is insufficient for high-precision context separation.

To refine the initial filtering and improve context separation, we subsequently applied a machine-learning relevance classifier trained on TF-IDF n-gram features. Five classical models were compared: Random Forest, LinearSVC, Multinomial Naive Bayes, Logistic Regression, and KNN under the same evaluation protocol. Such comparative assessment is important for noisy, unstructured social-media text, where model performance can vary substantially across algorithms and feature interactions; therefore, selecting the most robust relevance model requires empirical evaluation and stability checks [33].

Table 2. Performance Evaluation Relevance Classification

Model	Accuracy	Test F1-Macro	CV F1-Macro (Mean)	CV F1-Macro (Std)
RandomForest	0.984	0.943	0.928	0.033
LinearSVC	0.961	0.870	0.858	0.041
LogReg	0.949	0.842	0.724	0.044
Naive Bayes	0.925	0.600	0.489	0.023
KNN(k=7)	0.306	0.292	0.353	0.124

Overall, Random Forest achieved the best performance for Antam-gold tweet relevance

classification. On the held-out test set (n = 255), it obtained Accuracy = 0.984 and macro-F1 = 0.943. The model also showed strong stability under 5-fold cross-validation on the training split, with CV macro-F1 = 0.928 ± 0.033 (mean ± std). Based on these results, Random Forest was selected as the final relevance model, and its predictions were used to filter relevant tweets before proceeding to sentiment classification and the construction of a monthly sentiment index.

Compared with other models, LinearSVC delivered competitive performance but remained below Random Forest (Accuracy = 0.961; macro-F1 = 0.870; CV macro-F1 = 0.858 ± 0.041), indicating lower overall effectiveness and slightly higher variability across folds. Logistic Regression produced moderate results (Accuracy = 0.949; macro-F1 = 0.842; CV macro-F1 = 0.724 ± 0.044), suggesting that a linear decision boundary over TF-IDF features was less able to capture relevance patterns under cross-validation. Multinomial Naive Bayes showed weaker class separation (Accuracy = 0.925; macro-F1 = 0.600; CV macro-F1 = 0.489 ± 0.023), which is consistent with its sensitivity to feature independence assumptions in noisy social-media text. KNN (k = 7) performed poorly (Accuracy = 0.306; macro-F1 = 0.292; CV macro-F1 = 0.353 ± 0.124), aligning with known limitations of distance-based methods in high-dimensional sparse TF-IDF spaces. To account for class imbalance, we report macro-F1 alongside accuracy and include precision/recall, confusion matrices, and qualitative error analysis to examine ambiguity-driven errors[34][33]. In detail, Precision and Recall values are reported in the classification report and confusion matrix to facilitate the interpretation of error patterns, particularly in identifying false-positive and false-negative cases arising from term ambiguity and informal language variation on Twitter.

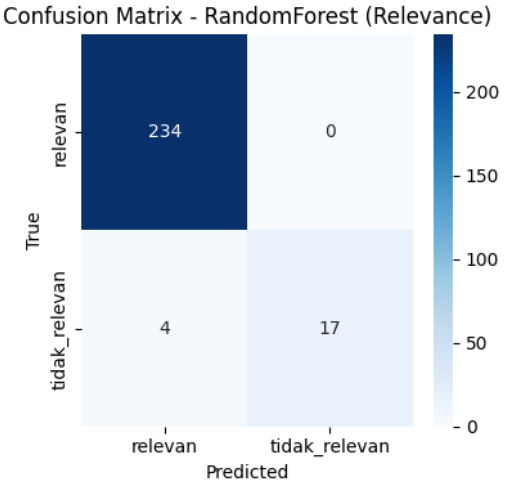


Figure 2. Best Confusion Matrix Model Visualization

The confusion matrix (Figure 2) for Random Forest on the held-out test set (n=255) indicates strong relevance separation. The model correctly classified 234 relevant tweets and 17 non-relevant tweets, with 0 false negatives for the relevant class and 4 false positives (non-relevant tweets predicted as relevant). This corresponds to Accuracy = 0.984 and macro-F1 = 0.943. For the relevant class, precision = 0.983 and recall = 1.000, showing that the model retains full coverage of relevant tweets while keeping leakage from the non-relevant class low. For the non-relevant class, precision = 1.000 and recall = 0.810, indicating that a small portion of stock-/irrelevant tweets still enters the relevant set. Because the non-relevant class is relatively small in this split (21 tweets), these results should be interpreted together with cross-validation stability and qualitative inspection of the remaining false positives.

3.3 Performance of Sentiment Classification Results

This section reports the performance evaluation of the sentiment classification on Indonesian-language Antam gold tweets, conducted after the relevance filtering stage to ensure that only tweets genuinely discussing physical gold are analysed. The evaluation focuses on two-class sentiment classification (positive vs. negative) by excluding the neutral label, to obtain a clearer

polarity signal more suitable for constructing a monthly sentiment index. Multiple machine learning algorithms are assessed using TF-IDF representations, with accuracy and F1-macro used to measure model performance and stability, providing the basis for selecting the most reliable sentiment model to extract a risk-related sentiment signal for future VaR-oriented integration.

Table 3. Performance Evaluation Sentiment Classification

Model	Accuracy	Test F1-Macro	CV F1-Macro (Mean)	CV F1-Macro (Std)
Random Forest	0.564	0.552	0.567	0.009
LinearSVC	0.543	0.542	0.561	0.022
Naive Bayes	0.538	0.529	0.574	0.021
Logistic Regression	0.526	0.519	0.576	0.013
KNN (k=5)	0.526	0.512	0.553	0.015

The TF-IDF–based sentiment classification results show that separating positive vs. negative polarity on this dataset remains challenging. After excluding the neutral class, the sentiment set contains 1,167 tweets, with a near-balanced distribution (609 positive and 558 negative). On the held-out test split, the best test macro-F1 is achieved by Random Forest (Accuracy = 0.564, macro-F1 = 0.552), followed closely by LinearSVC (Accuracy = 0.543, macro-F1 = 0.542) and Naive Bayes (Accuracy = 0.538, macro-F1 = 0.529).

Cross-validation results further indicate modest and relatively similar stability across models, with mean CV macro-F1 ranging from approximately 0.553 to 0.576 (std ≈ 0.009–0.022). Notably, Logistic Regression yields the highest mean CV macro-F1 (0.576 ± 0.013), while Random Forest provides the strongest hold-out performance. Given this pattern, the sentiment classifier can be selected based on the evaluation priority: Random Forest for slightly better held-out performance, or Logistic Regression for marginally better average cross-validated performance. The selected model's polarity outputs are then aggregated into a monthly sentiment index for downstream risk-oriented exploratory analysis.

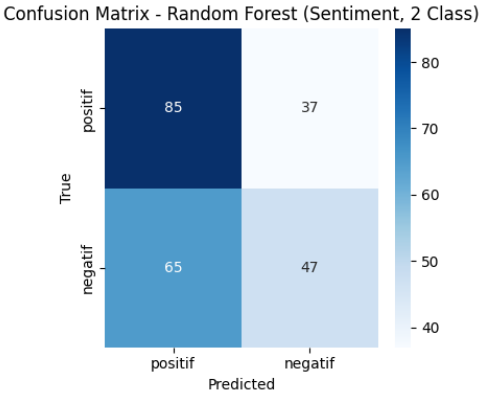


Figure 3. Best Confusion Matrix Model Visualization

The confusion matrix (Figure 3) for the Random Forest sentiment classifier on the held-out test set (n = 234) indicates moderate performance (Accuracy = 0.564, macro-F1 = 0.552). The model correctly classified 85 positive tweets and 47 negative tweets. For the positive class, precision is 0.567 (85/150) and recall is 0.697 (85/122), showing that the model retrieves a substantial portion of positive tweets but still yields many false positives (65 negative tweets predicted as positive). For the negative class, precision is 0.560 (47/84) and recall is 0.420 (47/112), suggesting that negative polarity is harder to detect and is frequently shifted into the positive class. Overall, the errors are consistent with subtle/implicit polarity cues in Indonesian Twitter text; therefore, these results are interpreted alongside cross-validation stability and brief qualitative checks, especially because the sentiment labels are derived from lexicon-based weak

supervision rather than manual gold annotations.

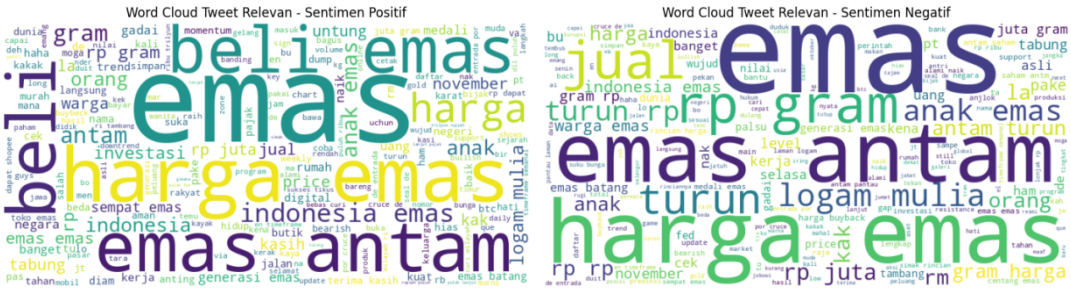


Figure 4. Semantic Interpretation of Sentiment Polarity

The positive-sentiment word cloud (Figure 4) is dominated by domain-specific terms such as “emas”, “antam”, “harga”, “gram”, “rp”, “juta”, alongside transaction-related words like “jual”. This pattern suggests that positive opinions frequently occur in the context of price updates and transactional discussions (buy/sell/save gold), which are often informative but can also carry an evaluative tone. In addition, the presence of explicitly positive cues such as “untung” (profit) and social expressions like “makasih/terima kasih” indicates that some positive tweets reflect satisfactory experiences (e.g., helpful information, responsiveness, or smooth transactions).

In contrast, the negative-sentiment word cloud remains dominated by core domain vocabulary (“emas”, “antam”, “harga”, “gram”, “rp”, “juta”, “logam mulia”), but more strongly highlights complaint-related terms such as “turun” (down), “mahal” (expensive), “antri/antrian” (queue), and “jual”. This semantic pattern indicates that negative sentiment is commonly associated with concerns about price declines, perceived high prices, and service/access friction (e.g., queues at boutiques). Overall, the contrast in prominent terms across the two classes supports that the polarity separation captures two main contexts: (i) price/transaction narratives perceived as beneficial (positive) and (ii) price drops/high prices/queue-related complaints (negative). Since word clouds are descriptive, these findings are presented as qualitative validation of the classifier outputs rather than causal evidence.

3.4 Baseline Pretrained Embedding

This study includes a baseline using pretrained sentence embeddings that map each tweet into a dense semantic vector. The baseline is reported as an additional benchmark alongside the TF-IDF models to assess whether semantic representations help sentiment polarity classification in Indonesian Twitter data, where posts are often informative (e.g., price updates) and may express sentiment implicitly. We use a SentenceTransformer to generate embeddings and train a lightweight classifier on top, evaluated under the same 80:20 split and 5-fold cross-validation to ensure comparability.

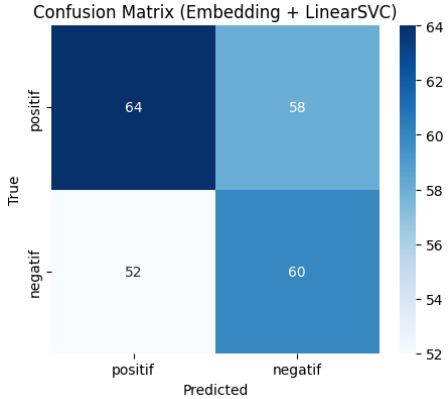


Figure 5. Embedding baseline (SentenceTransformer)

Using sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, the Embedding + LinearSVC baseline achieves a test macro-F1 of approximately 0.530 and a 5-fold CV macro-F1 of approximately 0.537 ± 0.007, indicating relatively stable performance across folds. The confusion matrix shows notable overlap between the positive and negative classes, suggesting that many tweets contain weak affective cues or mixed informational content, which makes polarity separation challenging even with semantic embeddings. Overall, this baseline provides a reference point for the proposed hybrid pipeline and highlights that label noise and implicit sentiment remain key difficulties in this dataset.

3.5 Implication for Market Risk Signal Extraction

This section examines the potential use of a social-media–based sentiment index as a risk-monitoring proxy by descriptively comparing its temporal movement with Antam gold prices. Rather than treating sentiment as a direct price-prediction variable, we interpret the index as an exogenous, risk-related indicator that may complement historical price–based risk signals. Given the limited number of time months in the observation window, correlation and lag analyses are reported for descriptive/illustrative purposes only and should not be interpreted as predictive validation.

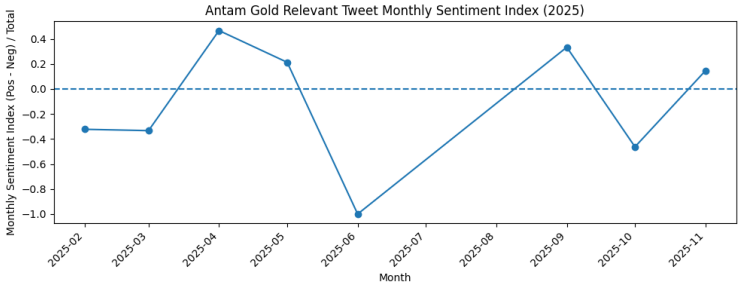


Figure 6. Monthly Sentiment Index of Antam Gold Relevant Tweets

Figure 6 presents the monthly sentiment index computed from tweets predicted as relevant to Antam's physical gold. The index varies across months and crosses the neutral line (0), indicating that aggregate public sentiment alternates between relatively optimistic and pessimistic periods rather than remaining consistently positive. In the observed months, the series shows positive phases (e.g., around April–May and September–November) and negative phases (e.g., around February–March and October), suggesting that sentiment may shift with changing information flows, user experiences, and price-related discussion intensity. This pattern aligns with prior findings that social-media sentiment can reflect investor perceptions and market reactions promptly [10], [35].

Because the index is monthly aggregated, it smooths short-term noise in tweet-level polarity and yields a more interpretable, higher-level indicator. The resulting monthly sentiment series is used as an exploratory, risk-related signal for descriptive comparison with market variables (e.g., gold price movements). Nevertheless, the interpretation remains preliminary: the number of observed months is limited, and the signal may be affected by episodic events and reposted/templated content. Therefore, stronger validation, such as longer-horizon data, lagged association tests, and integration into VaR/volatility frameworks [36], [37].

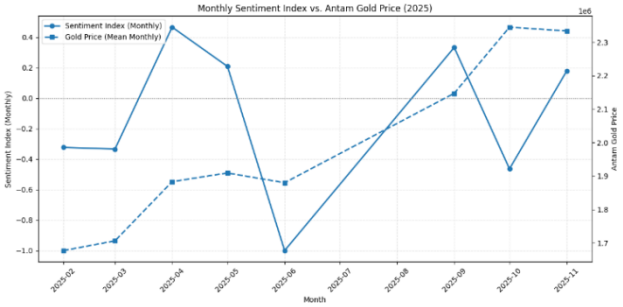


Figure 7. Monthly Sentiment Index vs Antam Gold Price

The implications of sentiment for market risk signal extraction were examined by comparing the monthly sentiment index derived from Antam-gold–relevant tweets with monthly Antam gold prices, using both visual inspection and quantitative association measures. The time-series plot suggests that sentiment and price may occasionally move in the same direction, but the pattern is not consistently linear across the observation window. To complement the visual comparison, we computed Pearson's correlation (linear association) and Spearman's rank correlation (monotonic association) for multiple lags (Lag 0–Lag 3), where Lag k evaluates sentiment_{t−k} against price_t (Table 4).

Tabel 4. Correlation Summary by Lag

Correlation Analysis	ρ (p-value)	Interpretation	Conclusion
Lag 0 (sentiment at month t vs price at month t)	0.1317 (0.7558)	very weak (+), ns	No same-month association (exploratory).
Lag 1 (sentiment at t−1 vs price at t)	−0.1622 (0.7283)	very weak (−), ns	No evidence that sentiment leads the next-month price.
Lag 2 (sentiment at t−2 vs price at t)	−0.5218 (0.2883)	moderate (−), ns	Possible negative indication, but not statistically supported (a few months).
Lag 3 (sentiment at t−3 vs price at t)	0.2000 (0.7471)	weak (+), ns	No lag-3 association (highly exploratory).

Note: not significant

Table 4 indicates that the associations are generally weak and statistically non-significant across lags (all p-values > 0.05). For example, the same-month association is very small (Pearson's $r = 0.058$, $p = 0.891$; Spearman's $\rho = 0.132$, $p = 0.756$). The lagged correlations remain weak (e.g., Lag 1 Pearson's $r = -0.193$, $p = 0.679$; Spearman's $\rho = -0.162$, $p = 0.728$). Although Spearman's ρ at Lag 2 shows a moderate negative indication ($\rho = -0.522$, $p = 0.288$), it is not statistically supported, likely reflecting the limited number of observed months and potential episodic effects. Overall, these results suggest that, within this sample, social-media sentiment is unlikely to function as a reliable short-horizon predictor of gold price levels[38][39]. Instead, the sentiment index is more appropriately interpreted as an exploratory, risk-related signal reflecting public reactions and perceived uncertainty, while stronger validation is required.

4. CONCLUSION

This study proposes a hybrid relevance–sentiment classification framework to analyze public opinion on physical Antam gold from Indonesian Twitter data and to support exploratory market risk signal extraction. Tweets collected during February–November 2025 were preprocessed and deduplicated on normalized text, yielding 1,271 unique tweets. Using weak supervision (rule-/lexicon-based silver labels) and TF-IDF–based machine learning, the framework performs (1) relevance filtering to separate physical-gold discussion from stock/irrelevant contexts and (2) two-class sentiment classification (positive vs. negative) on relevance-filtered tweets. Random Forest delivered the strongest relevance performance (macro-F1 = 0.943; 5-fold CV macro-F1 = 0.928 ± 0.033). For sentiment classification, performance was moderate and close across models; while Random Forest achieved the best hold-out macro-F1 (0.552), Logistic Regression/Naive Bayes showed slightly better cross-validation stability and is therefore preferred

for downstream aggregation. Aggregating tweet-level polarity into a monthly sentiment index yielded only a weak association with gold prices, indicating the index is better interpreted as a risk-perception proxy rather than a direct price predictor. Future work will strengthen duplicate-aware evaluation, extend the observation horizon, and integrate the sentiment signal into VaR/tail-risk frameworks.

REFERENCES

[1] D. Korsah, Lord Mensah, K. A. Osei, and G. Amewu, "Dynamic connectedness, hedge and safe-haven effects: cryptocurrencies, precious metals and African stock markets," *Int. J. Emerg. Mark.*, 2025, doi: 10.1108/IJOEM-06-2024-1090.

[2] M. Zaenal Asikin, "Peran Emas sebagai Lindung Nilai terhadap Ketidakpastian Pasar Keuangan Global," *Hawalah Kaji. Ilmu Ekon. Syariah*, vol. 3, no. 3, pp. 123–133, 2024, doi: 10.57096/hawalah.v3i3.54.

[3] U. Yeniley, "The Role of Gold ETFs in Market Stability, Price Discovery, and Economic Dynamics during Crises," *Asian J. Econ. Bus. Account.*, vol. 25, no. 1, pp. 60–74, 2025, doi: 10.9734/ajeba/2025/v25i11633.

[4] M. H. Rustam, D. Handoko, A. Rustam, and M. April, "the Dynamics of Court Decisions in the Civil Dispute Over Gold Sales Between Pt Antam Tbk and Budi Said and Its Implications for the Application of Civil Law in bertnesia," vol. 23, no. 158, pp. 165–176, 2025.

[5] F. Febriansyah and M. A. Adlan, "Saham Sebagai Instrumen Investasi: Peluang, Risiko, dan Strategi Investasi Di Tengah Dinamika Ekonomi Global 2024 – 2025," vol. 3, no. 2, pp. 207–220, 2025.

[6] Muhammad Hisam, "Menavigasi Volatilitas Pasar: Wawasan Tentang Instrumen Keuangan Dan Strategi Investasi," *Curr. J. Ekon. dan Perbank. Syariah Vol.*, vol. 02, no. April, pp. 315–328, 2024.

[7] S. I. Blihaq and R. A. Muthmainah, "Value-at-Risk (VaR) Modeling of LQ45 Stocks Using the GARCH Approach," vol. 6, no. 6, pp. 509–518, 2025.

[8] I. Akgül, M. Bildirici, and S. Özdemir, "Evaluating the Nonlinear Linkage between Gold Prices and Stock Market Index Using Markov-Switching Bayesian VAR Models," *Procedia - Soc. Behav. Sci.*, vol. 210, pp. 408–415, 2015, doi: 10.1016/j.sbspro.2015.11.388.

[9] Nida Shaffitry Hafidz, "Factors Influencing Stock Valuation: A Systematic Review of the Role of Market Fundamentals and Sentiment Nida," vol. I, pp. 13–20, 2024.

[10] F. Aminarty, N. Amina, A. Indrijawati, and M. I. Ferdiansah, "Interaksi Antara Perilaku Investor dan Tren Pasar Modal," *Adv. Manag. Financ. Report.*, vol. 3, no. 3, pp. 335–362, 2025, doi: 10.60079/amfr.v3i3.531.

[11] A. Wibowo, *AI & ML Di Bidang Keuangan: Mengatasi Masalah Kompleks dan Aplikasi Lingkungan, Sosial dan Tatakelola(ESG)*. Semarang: Prima Agust Teknik.

[12] Christophorus Bintang Saputra and D. P. Koesrindartoto, "Pemanfaatan Analisis Sentimen Youtube untuk Prediksi Harga Saham: Studi pada Investor Retail Indonesia," *J. Manaj.*, vol. 21, no. 1, pp. 1–17, 2024, doi: 10.25170/jm.v21i1.5184.

[13] M. Li, S. Guo, and R. Liu, "BERT-based Consumer Sentiment Analysis for Personalized Marketing Strategies," *Inform.*, vol. 49, no. 28, pp. 211–226, 2025, doi: 10.31449/inf.v49i28.8232.

[14] F. Kurniawan, H. Agustian, D. Dermawan, R. Nurdin, N. Ahmadi, and O. Dinaryanto, "Hybrid Rule-Based and Reinforcement Learning for Urban Signal Control in Developing Cities: A Systematic Literature Review and Practice Recommendations for Indonesia," *Appl. Sci.*, vol. 15, no. 19, p. 10761, 2025, [Online]. Available: https://doi.orgorg/10.3390/app151910761

[15] M. M. Yousuf and S. A. Sofi, "Bug Classification in quantum software: a rule-based framework and its evaluation," *Autom. Softw. Eng.*, vol. 33, no. 1, pp. 1–25, 2026, doi: 10.1007/s10515-025-00585-7.

[16] N. Rahayu, S. Indri Yani, M. Marwah, and A. Pratama, "Analisis Sentimen terhadap

Terorisme pada Platform Twitter menggunakan Support Vector Machine," *J. Manaj. Inform. Sist. Inf. dan Teknol. Komput.*, vol. 4, no. 1, pp. 430–441, 2025, doi: 10.70247/jumistik.v4i1.152.

[17] B. K. Shrivash, D. K. Verma, and P. Pandey, "A Novel Framework for Text Preprocessing using NLP Approaches and Classification using Random Forest Grid Search Technique for Sentiment Analysis," *Econ. Comput. Econ. Cybern. Stud. Res.*, vol. 59, no. 2, pp. 91–107, 2025, doi: 10.24818/18423264/59.2.25.06.

[18] W. Y. Melhem, A. Abdi, and F. Meziane, "Deep Learning Classification of Traffic-Related Tweets: An Advanced Framework Using Deep Learning for Contextual Understanding and Traffic-Related Short Text Classification," *Appl. Sci.*, vol. 14, no. 23, 2024, doi: 10.3390/app142311009.

[19] Z. Shahbazi, R. Jalali, and Z. Shahbazi, "AI-Driven Sentiment Analysis for Discovering Climate Change Impacts," *Smart Cities*, vol. 8, no. 4, pp. 1–20, 2025, doi: 10.3390/smartcities8040109.

[20] A. Alotaibi, F. Nadeem, and M. Hamdy, "Weakly Supervised Deep Learning for Arabic Tweet Sentiment Analysis on Education Reforms: Leveraging Pre-Trained Models and LLMs with Snorkel," *IEEE Access*, vol. 13, no. February, pp. 30523–30542, 2025, doi: 10.1109/ACCESS.2025.3541154.

[21] M. Alizadeh and A. Seilsepour, "A novel self-supervised sentiment classification approach using semantic labeling based on contextual embeddings," *Multimed. Tools Appl.*, vol. 84, no. 12, pp. 10195–10220, 2025, doi: 10.1007/s11042-024-19086-y.

[22] A. M. K. Mohammed, G. G. M. N. Ali, and S. S. Khairunnesa, "GSAF: An ML-Based Sentiment Analytics Framework for Understanding Contemporary Public Sentiment and Trends on Key Societal Issues," *Inf.*, vol. 16, no. 4, 2025, doi: 10.3390/info16040271.

[23] C. P. T. Mendoza, "Analyzing Historical Trends and Predicting Market Sentiment in Digital Currency Using Time Series Decomposition and ARIMA Models on Crypto Fear and Greed Index Data," *J. Digit. Mark. Digit. Curr.*, vol. 2, no. 3, pp. 270–297, 2025, doi: 10.47738/jdmdc.v2i3.37.

[24] M. Halousková and Š. Lyócsa, "Forecasting U.S. equity market volatility with attention and sentiment to the economy," pp. 1–51, 2025, [Online]. Available: http://arxiv.org/abs/2503.19767

[25] Z. Zhang *et al.*, "Interpretable machine learning model to predict 90-day radiographically confirmed pneumonia after chemotherapy initiation in non-Hodgkin lymphoma: development and internal validation of a single-center cohort," *Front. Med.*, vol. 12, no. September, pp. 1–19, 2025, doi: 10.3389/fmed.2025.1674896.

[26] P. R. Alapati *et al.*, "Improving English Writing Skills Through NLP-Driven Error Detection and Correction Systems," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 2, pp. 1098–1110, 2025, doi: 10.14569/IJACSA.2025.01602109.

[27] R. Mühlhoff and T. Willem, "Social media advertising for clinical studies: Ethical and data protection implications of online targeting," *Big Data Soc.*, vol. 10, no. 1, 2023, doi: 10.1177/20539517231156127.

[28] M. Levine, R. Philpot, S. J. Nightingale, and A. Kordoni, "Visual Digital Data, Ethical Challenges, and Psychological Science," *Am. Psychol.*, vol. 79, no. 1, pp. 109–122, 2024, doi: 10.1037/amp0001192.

[29] B. R. Babu, S. Ramakrishna, and S. K. Duvvuri, "Advanced Sentiment and Trend Analysis of Twitter Data Using CNN-LSTM and Word2Vec," *4th Int. Conf. Sentim. Anal. Deep Learn. ICSADL 2025 - Proc.*, pp. 1536–1543, 2025, doi: 10.1109/ICSADL65848.2025.10933031.

[30] K. Wang, "Multifactor prediction model for stock market analysis based on deep learning techniques," *Sci. Rep.*, vol. 15, no. 1, pp. 1–20, 2025, doi: 10.1038/s41598-025-88734-6.

[31] I. Moustati and N. Gherabi, "Bridging Behavioral Insights and Automated Trading: An Internet of Behaviors Approach for Enhanced Financial Decision-Making," *Inf.*, vol. 16, no. 5, 2025, doi: 10.3390/info16050338.

[32] K. Balog, *Entity-Oriented Search*, vol. 39. Stavanger: Springer, 2018. [Online]. Available:

https://eos-book.org

[33] Y. Liscano, L. A. Anillo Arrieta, J. F. Montenegro, D. Prieto-Alvarado, and J. Ordoñez, "Early Warning of Infectious Disease Outbreaks Using Social Media and Digital Data: A Scoping Review," *Int. J. Environ. Res. Public Health*, vol. 22, no. 7, pp. 1–34, 2025, doi: 10.3390/ijerph22071104.

[34] F. Islamiah and I. Muthahharah, "Pendekatan Analisis Sentimen Terhadap Persepsi Pengguna Fintech Syariah Ammana Berbasis Algoritma Machine Learning," vol. 4, no. 2, pp. 359–369, 2025.

[35] H. Kim-Hahm, A. S. Abou-Zaid, and A. Mohd, "News vs. Social Media: Sentiment Impact on Stock Performance of Big Tech Companies," *J. Risk Financ. Manag.*, vol. 18, no. 12, p. 660, 2025, doi: 10.3390/jrfm18120660.

[36] A. Odunaike, "Integrating Real-Time Financial Data Streams to Enhance Dynamic Risk Modeling and Portfolio Decision Accuracy," *Int. J. Comput. Appl. Technol. Res.*, vol. 14, no. 08, pp. 1–16, 2025, doi: 10.7753/ijcatr1408.1001.

[37] P. K. Bhowmik *et al.*, "AI-Driven Sentiment Analysis for Bitcoin Market Trends: A Predictive Approach to Crypto Volatility," *J. Ecohumanism*, vol. 4, no. 4, pp. 266–288, 2025, doi: 10.62754/joe.v4i4.6729.

[38] D. Léber and B. Egyed, "The Sentiment Augmented GARCH-LSTM Hybrid Model for Value-at-Risk Forecasting," *Comput. Econ.*, 2025, doi: 10.1007/s10614-025-11042-8.

[39] K. V. Thomas and K. Santhosh Kumar, "Investor Sentiment and Stock Market Volatility: A Behavioral Finance Perspective," *GS WOW Wisdom Worthy Res. J.*, vol. 4, no. 1, pp. 37–50, 2025, doi: 10.62078/grks.2025.v04i01.005.