

Enhancing Trustworthy Mining Trend Analytics Through Hybrid Semantic and Advertisement Detection Models

Mohamed Ayari^{1*}, Khulud Alashjaee², Balsam Alzamam³,
Aryam Alshammari⁴, Milaf Alshammari⁵, Aadil Alshammari^{6*}

Department of Information Technology-Faculty of Computing and Information Technology,
Northern Border University, Rafha 91911, Saudi Arabia¹

Department of Information Systems-Faculty of Computing and Information Technology,
Northern Border University, Rafha 91911, Saudi Arabia^{2,3,4,5,6}

Abstract—Mining-related social media discussions provide valuable information regarding operational risks, environmental concerns, equipment failures, workforce safety, and industrial market trends. However, large-scale mining discussion streams are increasingly affected by irrelevant promotional content, spam campaigns, and semantically unrelated posts that reduce analytical reliability. This study proposes a hybrid natural language processing framework for trustworthy mining trend analytics through semantic relevance assessment and advertisement detection. The framework combines transformer-based contextual embeddings with supervised machine learning to distinguish operationally relevant mining discussions from promotional and off-topic content. A curated mining-oriented communication dataset containing 12,500 public posts and discussion entries from mining-related channels was analyzed. Semantic relevance was evaluated using transformer embeddings and cosine similarity against predefined mining-topic descriptors, while advertisement detection was performed using TF-IDF features and supervised classifiers including Logistic Regression, Support Vector Machines, Random Forests, and Gradient Boosting. Experimental results demonstrate that the proposed framework significantly improves mining discussion reliability by filtering promotional intrusion and semantic drift. Random Forest and SVM classifiers achieved the best performance with F1-scores exceeding 0.91 in several mining categories. The framework supports intelligent industrial trend monitoring and can be extended to safety analytics, industrial risk monitoring, and operational intelligence systems.

Keywords—Mining analytics; semantic relevance detection; advertisement detection; industrial social media analysis; NLP; transformer embeddings; trustworthy analytics

I. INTRODUCTION

The mining industry is undergoing a major digital transformation driven by Industry 4.0 technologies, artificial intelligence (AI), Industrial Internet of Things (IIoT) infrastructures, autonomous systems, and large-scale industrial data analytics [3], [7]. Modern mining environments continuously generate massive volumes of heterogeneous data originating from operational reports, sensor streams, maintenance logs, safety records, industrial communications, equipment diagnostics, and online discussion platforms. These data sources

provide critical insights into operational efficiency, workforce safety, environmental sustainability, predictive maintenance, and market intelligence. Consequently, intelligent mining analytics has emerged as a strategic research direction for enabling safer, more efficient, and more sustainable mining ecosystems.

In recent years, online industrial communication channels and mining-related digital discussion streams have become increasingly important sources of operational intelligence. Mining companies, engineers, contractors, industrial operators, safety personnel, and technical communities frequently exchange information through industrial forums, professional social networks, enterprise collaboration systems, and mining-related discussion platforms. These discussion streams often contain valuable indicators related to equipment failures, geological risks, production interruptions, safety incidents, environmental hazards, logistics disruptions, workforce coordination, and commodity market dynamics. Unlike structured industrial databases, such communication streams provide near real-time situational awareness and can reveal emerging operational issues before they are formally documented in enterprise systems.

However, the growing dependence on mining-oriented digital communication streams introduces several analytical challenges. First, industrial discussion streams are highly noisy and heterogeneous. Mining-related communications often contain informal terminology, abbreviations, duplicated reports, incomplete technical descriptions, multilingual expressions, and inconsistent formatting. Second, mining discussion streams are increasingly contaminated by promotional campaigns, marketing-oriented content, spam messages, recruitment advertisements, duplicated commercial announcements, and semantically irrelevant discussions. Such noisy and off-topic content significantly reduces the reliability of industrial analytics systems and may distort operational trend interpretation. In safety-critical industrial environments, inaccurate interpretation of mining discussions can negatively affect operational decision-making, incident awareness, and industrial monitoring systems.

Another major challenge concerns semantic drift within mining communication streams. A discussion initially associated with a specific operational topic may gradually diverge toward unrelated conversations, commercial promotions, or

*Corresponding author.

generalized industrial commentary. Traditional keyword-based filtering techniques are often insufficient for handling such semantic variability because mining terminology frequently exhibits contextual ambiguity. For example, terms such as “blast,” “drill,” “ventilation,” “ore,” or “shutdown” may appear in both operationally relevant discussions and unrelated promotional content. Consequently, robust semantic understanding mechanisms are required to distinguish operationally meaningful mining discussions from noisy or misleading information streams.

Natural language processing (NLP) techniques have recently demonstrated substantial potential for addressing industrial text analytics problems. Traditional machine learning approaches based on Bag-of-Words (BoW), TF-IDF representations, and statistical classifiers have been widely used for text categorization, spam filtering, industrial document analysis, and information retrieval tasks. Although these approaches provide computational efficiency and interpretability, they often fail to capture deeper contextual semantics, especially in noisy industrial environments characterized by domain-specific vocabulary and contextual variability.

The emergence of transformer-based language models has significantly improved semantic representation learning in NLP applications. Models such as BERT, RoBERTa, DistilBERT, and domain-adapted transformer architectures can capture contextual relationships between words and sentences more effectively than traditional embedding approaches. Transformer-based semantic embeddings have demonstrated strong performance across multiple industrial applications, including predictive maintenance analytics, industrial event monitoring, technical document classification, anomaly detection, and operational intelligence systems. Their contextual representation capability makes them particularly suitable for mining-related semantic relevance assessment, where operational meaning frequently depends on domain-specific context.

Simultaneously, supervised machine learning techniques remain highly effective for detecting repetitive advertisement patterns and promotional behaviors. Models such as Logistic Regression, Support Vector Machines (SVM), Random Forests, Gradient Boosting Machines, and Naive Bayes classifiers have demonstrated strong performance in industrial spam filtering and binary classification tasks. When combined with TF-IDF feature representations, such models can effectively identify lexical patterns associated with promotional intrusion and repetitive marketing behavior.

Despite recent advances in industrial NLP and semantic analytics, several important research gaps remain unresolved. First, most existing industrial text analytics studies focus either on semantic classification or spam detection independently, without integrating both perspectives into a unified trustworthy analytics framework. Second, limited attention has been given to mining-specific communication streams, despite their operational importance and unique semantic characteristics. Third, existing mining analytics systems rarely address semantic drift and promotional contamination simultaneously. Finally, many industrial NLP studies prioritize classification accuracy while neglecting operational reliability, scalability, interpretability, and real-world deployment considerations.

To address these limitations, this study proposes a hybrid

NLP framework for enhancing trustworthy mining trend analytics through semantic relevance assessment and advertisement detection. The proposed framework combines transformer-based semantic embeddings with supervised machine learning models to identify operationally relevant mining discussions while filtering promotional and semantically irrelevant content. Unlike traditional keyword-based mining analytics approaches, the proposed method integrates contextual semantic understanding with promotional-content detection to improve the reliability and interpretability of mining communication streams.

The framework consists of two complementary analytical modules. The first module performs semantic relevance detection using transformer-generated contextual embeddings and cosine similarity analysis against predefined mining-topic descriptors. This module evaluates whether individual discussion entries semantically align with operational mining topics. The second module performs advertisement detection using TF-IDF feature extraction and supervised machine learning classifiers to identify promotional and marketing-oriented content. The outputs of both modules are integrated to produce trustworthy mining trend analytics that preserve operationally meaningful information while reducing semantic noise and promotional intrusion.

To evaluate the proposed framework, a mining-oriented industrial discussion dataset was constructed using mining-related communication streams collected from industrial forums, public operational reports, professional social platforms, and mining discussion channels. The dataset includes approximately 12,500 text entries spanning multiple mining domains, including underground operations, equipment maintenance, workforce safety, environmental monitoring, industrial logistics, and commodity market discussions. Extensive preprocessing and normalization procedures were applied to address industrial text noise and domain-specific terminology challenges.

Experimental evaluations were conducted using multiple supervised classifiers, including Logistic Regression, Support Vector Machines, Random Forests, Gradient Boosting, and Naive Bayes. Semantic relevance analysis was performed using transformer-based contextual embeddings and cosine similarity thresholding. Results demonstrate that the proposed framework effectively improves the reliability of mining communication analytics by reducing semantic drift and filtering promotional contamination. Random Forest and SVM models achieved the strongest performance across multiple mining categories, while transformer-based semantic analysis successfully identified operationally relevant discussions under highly noisy conditions.

The proposed framework contributes to the growing field of trustworthy industrial AI and intelligent mining analytics. Beyond mining applications, the framework can be extended to broader industrial domains involving noisy communication streams, operational intelligence systems, industrial risk monitoring, and real-time semantic analytics. Furthermore, the integration of contextual semantic modeling with promotional intrusion detection provides a practical foundation for future research in industrial trustworthy AI systems.

The main contributions of this study are summarized as follows:

- This study proposes a hybrid NLP framework that integrates transformer-based semantic relevance detection with supervised advertisement filtering for trustworthy mining analytics.
- A mining-oriented industrial discussion dataset is constructed and analyzed containing approximately 12,500 categorized discussion entries spanning multiple mining operational domains.
- The impact of semantic drift and promotional intrusion is investigated on mining trend analytics using contextual embedding analysis and supervised classification models.
- Extensive comparative experiments evaluate multiple machine learning classifiers across diverse mining communication categories.
- Scalability analysis, operational interpretation, and industrial deployment perspectives are provided for real-time mining intelligence systems.
- The experiments demonstrate that combining semantic relevance assessment with advertisement detection substantially improves the reliability and interpretability of industrial discussion analytics.

The remainder of this study is organized as follows. Section II reviews related work on industrial NLP, mining analytics, semantic relevance detection, and promotional-content filtering. Section III describes the mining discussion dataset and preprocessing procedures. Section IV presents the proposed hybrid framework and mathematical formulation. Section V describes the experimental setup and evaluation methodology. Section VI reports and discusses the experimental results. Finally, Section VII concludes the study and outlines future research directions.

II. RELATED WORK

Research relevant to trustworthy mining trend analytics spans mining digitalization, industrial text analytics, semantic representation learning, social-media mining, and trustworthy AI. Mining 4.0 studies consistently identify data-driven monitoring, AI, automation, and connected industrial systems as central enablers of safer and more efficient operations [8], [7], [4]. Recent reviews further confirm the rapid expansion of AI across mineral processing, autonomous mining, environmental monitoring, and human-machine collaboration, while emphasizing data quality, validation, transferability, and deployment reliability as persistent challenges [11], [12], [13], [14], [15], [16]. Mining-specific predictive AI has also been demonstrated for subsidence forecasting in Saudi phosphate mines [18]. In parallel, Ayari et al. [21] developed a game-theoretic electromagnetic isolation framework for wireless connectivity in mineral and phosphate mining environments, highlighting the importance of reliable communication infrastructure for real-time monitoring, autonomous equipment control, and worker safety. Together, these studies provide direct evidence that data-driven intelligence and dependable communication are complementary foundations for smart-mining operations.

Transformer language models provide a strong basis for contextual interpretation of noisy text. BERT introduced deep

bidirectional contextual representations [1], while Sentence-BERT made sentence-level semantic comparison practical through embedding vectors and cosine similarity [2]. These properties are particularly useful for mining communication because terms such as “blast”, “shutdown”, “drill”, and “ventilation” can be operationally meaningful or irrelevant depending on context. General text-classification research further shows that sparse lexical representations and classical supervised classifiers remain competitive when discriminative vocabulary is strong [5].

Social-media streams are noisy, dynamic, and vulnerable to irrelevant or manipulative content [6]. In industrial monitoring, this creates a reliability problem: a trend-analysis system can produce misleading conclusions if promotional messages or semantically drifting posts are treated as operational evidence. Trustworthy-AI research therefore motivates evaluation beyond raw accuracy, emphasizing robustness, transparency, reliability, and human-centered deployment [10], [9], [17]. Related industrial-AI studies have also highlighted adaptive decision making and fairness-aware optimization in energy and multi-agent control [19], [20]; however, these works do not address semantic contamination in mining communication streams.

The gap addressed here is narrower than generic spam detection or generic mining AI. Existing mining AI literature is dominated by sensor, image, process-control, exploration, and predictive-maintenance applications; it provides limited treatment of noisy unstructured mining communication streams. Conversely, general semantic and spam-filtering studies rarely evaluate mining-specific operational categories. The present framework therefore combines two complementary signals: contextual semantic alignment to a mining topic and lexical evidence of promotional intent. Table I summarizes the distinction.

III. MINING DISCUSSION DATASET AND DATA PREPROCESSING

This section presents the construction of the mining-oriented discussion dataset, the annotation methodology, the preprocessing pipeline, and the feature engineering procedures used in this study. Unlike traditional industrial datasets that primarily rely on structured sensor measurements or maintenance logs, the proposed dataset focuses on unstructured mining communication streams characterized by high semantic variability, operational terminology, contextual ambiguity, and substantial textual noise. Consequently, careful dataset construction and preprocessing are essential for ensuring reliable semantic analysis and trustworthy mining trend analytics.

A. Mining Communication Data Sources

The dataset used in this work was constructed from multiple mining-related communication environments to ensure semantic diversity and operational realism. Data collection targeted publicly accessible industrial communication channels containing discussions relevant to mining operations, industrial logistics, workforce coordination, environmental monitoring, equipment maintenance, and commodity market analysis.

The collected communication streams originated from the following sources:

TABLE I. CRITICAL COMPARISON OF REPRESENTATIVE LITERATURE AND THE PRESENT STUDY

Research direction	Mining-specific	Contextual semantics	Promotional filtering	Main limitation relative to this work
Mining 4.0 / mining AI [8], [13], [14], [15], [21]	Yes	Limited	No	Focuses mainly on automation, sensing, processing, and operational AI.
Sentence embeddings [1], [2]	No	Strong	No	General semantic representation; no mining-specific trust filtering.
Social-media mining [6]	No	Moderate	Partial	General noisy social streams; no mining operational semantics.
Trustworthy industrial AI [10], [9]	Industrial	Varies	No	Trust principles without a mining communication pipeline.
Proposed framework	Yes	Strong	Yes	Joint semantic relevance, advertisement filtering, drift and scalability analysis.

- Mining-related industrial forums
- Public operational discussion boards
- Technical mining communities
- Industrial news comment sections
- Professional networking discussions
- Mining-related social media channels
- Equipment maintenance discussions
- Environmental incident reports

Unlike generic social-media datasets, mining communication streams contain specialized industrial vocabulary, technical abbreviations, equipment identifiers, geological terminology, operational procedures, and context-dependent expressions. This semantic complexity makes mining communication analytics substantially more challenging than traditional sentiment-analysis or spam-detection tasks.

The final dataset contains approximately 12,500 mining-related textual entries spanning multiple operational domains. Each entry represents a communication instance associated with a mining-related discussion stream.

B. Mining Topic Categories

To support fine-grained industrial analysis, the collected data were categorized into multiple mining operational domains. Each category represents a distinct semantic context frequently encountered in industrial mining communication systems.

The mining categories used in this study are summarized in Table II.

The category distribution is intentionally heterogeneous to reflect realistic mining communication conditions where operational discussion density varies significantly across industrial domains.

C. Dataset Annotation Protocol and Reliability

A rule-assisted single-expert annotation protocol was used to label the dataset for both semantic relevance detection and advertisement classification. Two primary labels were associated with each communication entry:

TABLE II. MINING OPERATIONAL CATEGORIES

Category	Description
Safety Monitoring	Discussions related to workforce safety, hazards, and operational incidents
Equipment Maintenance	Technical discussions concerning machinery diagnostics and maintenance
Environmental Monitoring	Environmental impact reports, pollution alerts, and sustainability discussions
Operational Logistics	Transportation, supply-chain coordination, and operational scheduling
Mineral Market Analytics	Commodity prices, industrial economics, and production trends
Underground Operations	Discussions concerning excavation, drilling, ventilation, and extraction activities

- Semantic Relevance Label
- Advertisement/Promotional Label

The semantic relevance label indicates whether a discussion entry is operationally related to its associated mining category. The advertisement label identifies whether the entry contains promotional, commercial, recruitment-oriented, duplicated marketing, or spam-related content.

To reduce annotation ambiguity, explicit labeling criteria were established before annotation. Entries were classified as operationally relevant when they contributed directly to mining-related operational discussions, technical analysis, safety reporting, equipment monitoring, environmental monitoring, logistics, or industrial coordination. Entries were classified as off-topic when they primarily contained unrelated commentary, generalized social discussion, duplicated information, or semantically unrelated industrial content. Promotional entries included recruitment advertisements, commercial equipment promotions, marketing campaigns, duplicated industrial announcements, sponsored content, and other non-operational promotional messages.

The final labels were produced under a single-expert protocol rather than by multiple independent coders. Consequently, inter-annotator agreement measures such as Cohen's κ are not applicable to this dataset. Label reliability was assessed through a post-hoc manual audit of randomly sampled records and cases close to the semantic decision boundary using the same predefined criteria. This audit was used exclusively as a quality-control step and did not contribute training labels or

tune the reported classifier parameters.

D. Dataset Statistics

Table III summarizes the final dataset composition after preprocessing and validation.

TABLE III. DATASET STATISTICS

Metric	Value
Total Entries	12,500
Operationally Relevant Entries	9,130
Off-topic Entries	2,110
Advertisement Entries	1,260
Average Text Length	34 words
Unique Vocabulary Size	18,400
Mining Categories	6
Average Entries per Category	2,083

The dataset exhibits moderate class imbalance because promotional entries constitute a minority class relative to operationally relevant discussions. Such imbalance reflects realistic industrial communication environments and introduces additional challenges for reliable advertisement detection.

E. Challenges of Mining Communication Streams

Mining communication streams present several domain-specific NLP challenges:

1) *Domain-specific terminology*: Mining discussions frequently contain highly specialized technical terminology, abbreviations, equipment identifiers, geological descriptors, and operational jargon. These linguistic characteristics reduce the effectiveness of generic NLP preprocessing pipelines.

2) *Contextual ambiguity*: Operational mining terms often exhibit context-dependent semantics. For example, terms such as “shutdown,” “blast,” or “ventilation” may indicate operational events, maintenance procedures, safety incidents, or commercial discussions depending on surrounding context.

3) *Noisy communication structures*: Mining communication streams often include:

- Incomplete sentences
- Informal writing
- Repeated technical phrases
- Hyperlinks
- Equipment codes
- Timestamp artifacts
- Duplicated operational alerts

Consequently, robust preprocessing and semantic normalization procedures are required before feature extraction.

F. Text Preprocessing Pipeline

A multi-stage preprocessing pipeline was designed to improve semantic consistency while preserving operational information. The preprocessing architecture is illustrated in Fig. 1.

The preprocessing pipeline consists of the following stages:

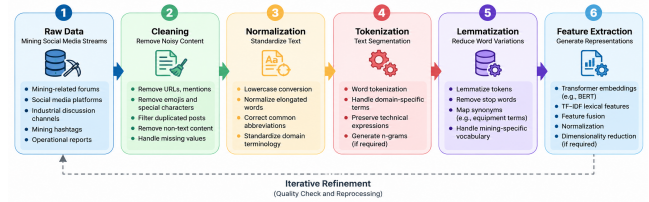


Fig. 1. Industrial text preprocessing pipeline used for mining communication analytics.

1) *Noise removal*: URLs, emojis, HTML artifacts, duplicated punctuation, excessive whitespace, and non-informative symbols were removed to reduce textual noise.

2) *Lowercase normalization*: All text entries were converted to lowercase to improve lexical consistency and reduce feature sparsity.

3) *Tokenization*: Text segmentation was performed using rule-based tokenization adapted for industrial communication environments.

4) *Stop-word elimination*: Frequently occurring non-informative words were removed using standard stop-word dictionaries augmented with domain-specific industrial filler expressions.

5) *Lemmatization*: Lemmatization was applied to reduce morphological variability while preserving semantic meaning.

6) *Industrial vocabulary preservation*: Unlike traditional preprocessing pipelines, technical mining terms and equipment identifiers were preserved because they contribute strongly to operational semantic understanding.

G. Feature Representation

Two complementary feature representations were employed to support the hybrid analytical framework.

1) *TF-IDF representation*: TF-IDF vectors were used for supervised advertisement detection because promotional messages frequently exhibit repetitive lexical patterns.

The TF-IDF weight of term t within document d is defined as:

$$TFIDF(t, d) = TF(t, d) \times IDF(t) \quad (1)$$

where:

$$IDF(t) = \log \left(\frac{N}{DF(t)} \right) \quad (2)$$

and:

- $TF(t, d)$ denotes term frequency,
- $DF(t)$ denotes document frequency,
- N denotes the total number of documents.

TF-IDF representations remain highly effective for identifying repetitive promotional expressions and duplicated marketing content.

2) *Transformer-based semantic embeddings*: Semantic relevance analysis was performed using transformer-generated contextual embeddings. Unlike TF-IDF representations, transformer embeddings capture semantic relationships between words and contextual operational meaning.

A pretrained BERT-based industrial language model was employed to generate sentence-level embeddings for mining communication entries. Each communication instance was transformed into a dense semantic vector representation:

$$E_i \in \mathbb{R}^d \quad (3)$$

where:

- E_i represents the embedding vector of entry i ,
- d denotes embedding dimensionality.

These embeddings were subsequently used for semantic similarity analysis and relevance detection.

H. Train-Test Partitioning

To ensure reliable evaluation, the dataset was partitioned into training and testing subsets using stratified sampling. The train-test split ratio was set to 80:20 while preserving category distribution and class proportions.

Cross-validation procedures were additionally applied during classifier training to reduce overfitting and improve generalization robustness.

I. Ethical Considerations

Only publicly accessible industrial communication data were used in this study. Personally identifiable information was removed during preprocessing to preserve privacy and confidentiality. The proposed framework is intended exclusively for industrial analytics, operational monitoring, and research purposes.

Furthermore, the framework does not attempt to infer sensitive workforce attributes or perform behavioral profiling. Instead, the objective is to improve semantic reliability and reduce operational misinformation within mining communication streams.

IV. PROPOSED HYBRID NLP FRAMEWORK

This section presents the proposed hybrid natural language processing framework for trustworthy mining trend analytics. The framework integrates transformer-based semantic relevance detection with supervised advertisement filtering to improve the reliability, interpretability, and operational usefulness of mining communication analytics.

Unlike traditional industrial text-analysis systems that rely solely on keyword matching or standalone classifiers, the proposed framework combines contextual semantic understanding with lexical promotional-content detection. This dual analytical architecture enables the framework to simultaneously address semantic drift, operational irrelevance, and promotional intrusion within mining discussion streams.

The overall architecture of the proposed framework is illustrated in Fig. 2.

A. Framework Overview

The proposed framework consists of five sequential analytical stages:

- Mining communication acquisition
- Industrial text preprocessing
- Semantic feature extraction
- Parallel relevance and advertisement analysis
- Trustworthy trend analytics generation

The framework receives mining-related communication streams as input and produces semantically filtered operational analytics as output. Each stage contributes to improving analytical reliability under noisy industrial communication conditions.

B. Stage 1: Mining Communication Acquisition

The first stage aggregates mining-related discussion streams from heterogeneous industrial communication environments. Data sources include industrial forums, technical communities, operational reports, professional communication channels, and public industrial discussion streams.

The collected communication entries are associated with:

- Operational category
- Timestamp
- Semantic topic descriptor
- Advertisement label
- Relevance label

This stage ensures semantic diversity and operational realism while supporting multi-domain mining analytics.

C. Stage 2: Industrial Text Preprocessing

Mining communication environments generate highly noisy and inconsistent textual data. Therefore, robust preprocessing is essential before semantic analysis and feature extraction.

The preprocessing stage includes:

- Noise removal
- URL elimination
- Lowercase normalization
- Tokenization
- Stop-word filtering
- Lemmatization
- Technical vocabulary preservation

Unlike generic NLP pipelines, the proposed preprocessing architecture preserves mining-specific terminology and industrial identifiers because they contain important operational semantics.

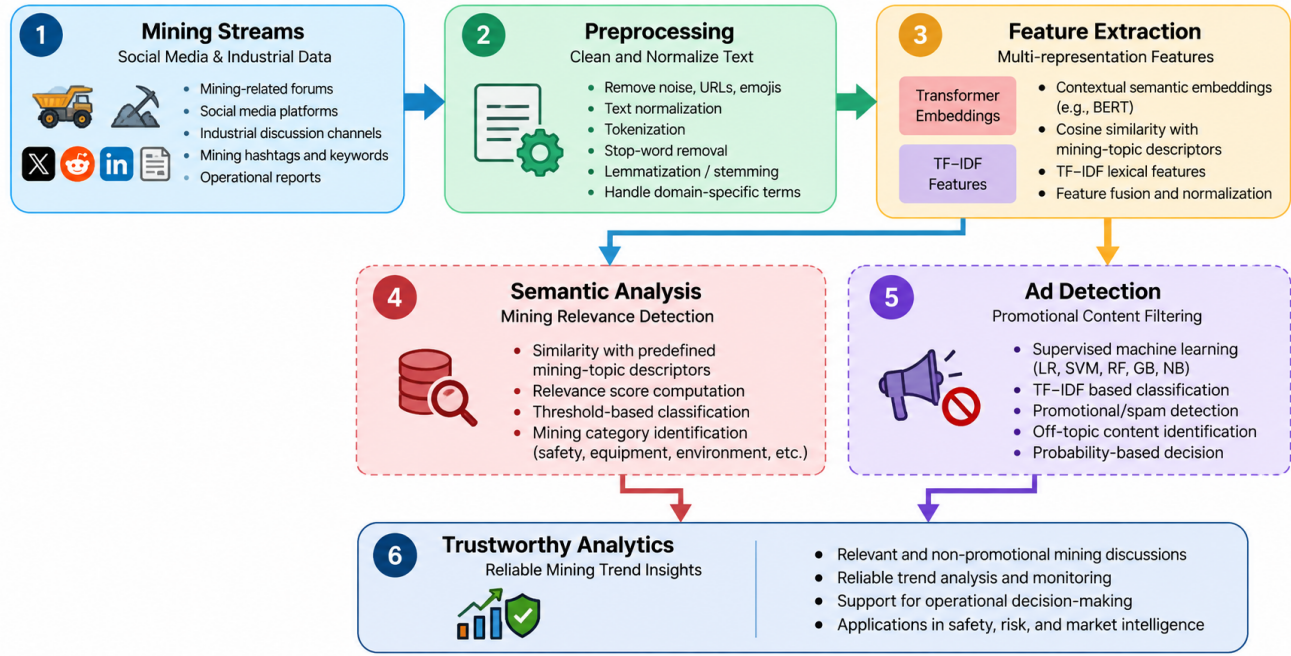


Fig. 2. Architecture of the proposed hybrid framework for trustworthy mining trend analytics.

Algorithm 1 summarizes the preprocessing procedure.

Algorithm 1 Mining Communication Preprocessing

Require: Raw mining communication stream D

Ensure: Cleaned industrial corpus C

```

1: for each entry  $d_i \in D$  do
2:   Remove URLs and HTML artifacts
3:   Remove duplicated punctuation and symbols
4:   Convert text to lowercase
5:   Apply tokenization
6:   Remove stop words
7:   Apply lemmatization
8:   Preserve technical mining terminology
9:   Append cleaned entry to  $C$ 
10: end for
    return  $C$ 

```

D. Stage 3: Semantic Feature Extraction

After preprocessing, semantic features are extracted using two complementary representation methods:

- TF-IDF lexical representation.
- Transformer-based contextual embeddings.

These representations support different analytical objectives within the framework.

1) *TF-IDF lexical features:* TF-IDF vectors are used primarily for advertisement detection because promotional messages frequently contain repetitive lexical structures and recurring commercial patterns.

The TF-IDF weight for term t in document d is defined as:

$$TFIDF(t, d) = TF(t, d) \times \log \left(\frac{N}{DF(t)} \right) \quad (4)$$

where:

- $TF(t, d)$ represents term frequency,
- $DF(t)$ represents document frequency,
- N denotes the total number of documents.

The resulting TF-IDF matrix captures discriminative lexical characteristics associated with promotional behavior.

2) *Transformer-based contextual embeddings:* Semantic relevance analysis is performed using transformer-generated contextual embeddings. A pretrained BERT-based industrial language model is employed to transform mining communication entries into dense semantic vectors.

For a given communication entry x_i , the transformer encoder generates:

$$E_i = f_{BERT}(x_i) \quad (5)$$

where:

- $E_i \in \mathbb{R}^d$ represents the semantic embedding,
- f_{BERT} denotes the transformer encoder,
- d denotes embedding dimensionality.

Transformer embeddings enable contextual semantic understanding beyond lexical similarity.

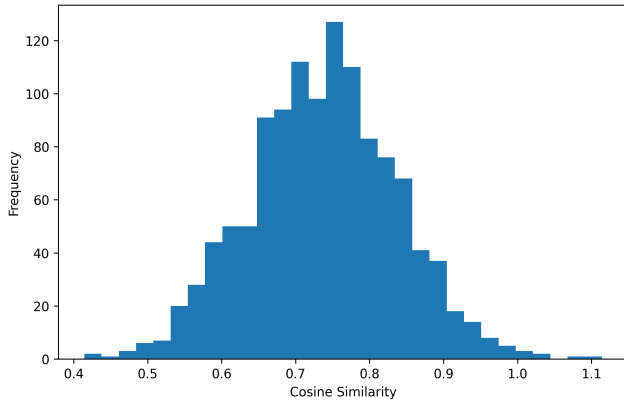


Fig. 3. Distribution of cosine similarity scores for semantic relevance detection.

E. Stage 4: Semantic Relevance Detection

The semantic relevance module evaluates whether a mining communication entry operationally aligns with a predefined mining topic descriptor.

Each mining category is associated with a semantic topic description vector:

$$T_j = f_{BERT}(topic_j) \quad (6)$$

where:

- T_j denotes the embedding vector of mining topic j .

Semantic similarity between communication entry E_i and topic descriptor T_j is computed using cosine similarity:

$$Sim(E_i, T_j) = \frac{E_i \cdot T_j}{\|E_i\| \|T_j\|} \quad (7)$$

A communication entry is considered operationally relevant when:

$$Sim(E_i, T_j) > \theta \quad (8)$$

where, θ represents the semantic relevance threshold.

The similarity distribution across mining categories is illustrated in Fig. 3.

1) *Semantic drift detection*: Semantic drift occurs when communication streams gradually diverge from operational mining discussions toward irrelevant or promotional content.

The framework quantifies semantic drift using cumulative similarity deviation:

$$Drift_t = \frac{1}{N_t} \sum_{i=1}^{N_t} (1 - Sim(E_i, T_j)) \quad (9)$$

where:

- $Drift_t$ denotes semantic drift at time interval t ,

- N_t represents the number of communication entries.

Large drift values indicate substantial semantic divergence within mining communication streams.

F. Stage 5: Advertisement Detection

The advertisement detection module identifies promotional and commercial content within mining communication streams.

Advertisement detection is formulated as a binary classification problem:

$$Y_i \in \{0, 1\} \quad (10)$$

where:

- $Y_i = 1$ indicates promotional content,
- $Y_i = 0$ indicates operational content.

Five supervised classifiers are evaluated:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest
- Gradient Boosting
- Naive Bayes

The classifiers receive TF-IDF vectors as input and generate advertisement predictions.

1) *Support vector machine formulation*: For SVM classification, the optimization objective is defined as:

$$\min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (11)$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad (12)$$

where:

- w denotes the separating hyperplane,
- C controls regularization,
- ξ_i represents slack variables.

2) *Random forest formulation*: Random Forest classification aggregates predictions from multiple decision trees:

$$\hat{Y} = mode\{T_1(x), T_2(x), \dots, T_k(x)\} \quad (13)$$

where:

- T_k represents the k -th decision tree.

Random Forest models are particularly robust under noisy industrial communication conditions.

G. Hybrid Fusion Strategy

The final trustworthy mining analytics output is generated by integrating semantic relevance predictions with advertisement-detection outputs.

A communication entry is classified as trustworthy operational content only when:

$$Trustworthy(x_i) = \begin{cases} 1, & Sim(E_i, T_j) > \theta \text{ and } Y_i = 0 \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

This hybrid fusion mechanism ensures that semantically relevant yet promotional entries are excluded from operational mining analytics.

H. Computational Complexity Analysis

Scalability and computational efficiency are important for real-time mining analytics systems.

The computational complexity of TF-IDF feature extraction is approximately:

$$O(N \times V) \quad (15)$$

where:

- N denotes the number of documents,
- V denotes vocabulary size.

Transformer embedding generation introduces higher computational cost:

$$O(N \times L^2 \times H) \quad (16)$$

where:

- L denotes sequence length,
- H denotes hidden-layer dimensionality.

Despite higher computational complexity, transformer-based semantic analysis provides substantially improved contextual understanding compared to traditional lexical approaches.

The scalability behavior of the proposed framework is illustrated in:

I. Industrial Deployment Considerations

The proposed framework is designed for deployment within real-time industrial monitoring environments.

Potential deployment scenarios include:

- Mining operational intelligence dashboards
- Industrial safety monitoring systems
- Environmental incident tracking
- Equipment-maintenance intelligence

- Industrial communication reliability monitoring

The integration of semantic relevance detection and advertisement filtering improves the reliability and interpretability of operational mining analytics while reducing semantic contamination within industrial communication streams.

V. EXPERIMENTAL SETUP AND EVALUATION METHODOLOGY

The experiments were designed to evaluate advertisement detection, mining-topic relevance, semantic drift, cross-category robustness, and runtime scalability using the 12,500-entry mining communication corpus described in Section III. Stratified sampling preserved class and category proportions. The primary split was 80:20, corresponding to 10,000 training entries and 2,500 held-out test entries; five-fold cross-validation was used during model selection. All pre-processing, TF-IDF extraction, classifier fitting, and metric computation were implemented in Python 3.11 with scikit-learn, while contextual embeddings were generated with PyTorch/HuggingFace on an RTX-series GPU.

A. Model and Hyperparameter Configuration

The semantic module used BERT-base with a 768-dimensional representation, maximum sequence length 128, batch size 32, AdamW optimization, learning rate 2×10^{-5} , and GELU activation. Mining-topic descriptors were manually defined from the operational scope of each category and embedded using the same encoder. Advertisement detection used TF-IDF features and five supervised baselines. The selected configurations are reported in Table IV; stating them explicitly addresses reproducibility of the comparison.

B. Semantic Threshold and Hybrid Decision Rule

For an entry embedding E_i and its mining-topic descriptor T_j , relevance is computed by cosine similarity:

$$s_{ij} = \frac{E_i^\top T_j}{\|E_i\|_2 \|T_j\|_2}. \quad (17)$$

The threshold sweep shown in Fig. 4 produced F1-scores of 0.72, 0.78, 0.84, 0.89, 0.93, 0.91, 0.88, 0.84, 0.79, and 0.73 at thresholds 0.50 through 0.95, respectively. The maximum F1-score (0.93) occurs at $\theta = 0.70$, which was therefore fixed before the final held-out evaluation.

The final hybrid rule retains an entry as trustworthy mining evidence only when it is semantically relevant and is not classified as promotional:

$$H_i = \mathbb{I}(s_{ij} \geq 0.70) [1 - \hat{a}_i], \quad (18)$$

where, $\hat{a}_i \in \{0, 1\}$ is the advertisement prediction. This definition makes “trustworthy” operational: accepted entries satisfy both contextual mining relevance and non-promotional-content requirements.

TABLE IV. COMPLETE EXPERIMENTAL CONFIGURATION

Component	Parameter	Value
Data	Train/test split	80/20 (10,000/2,500), stratified
Validation	Model selection	5-fold cross-validation
BERT	Embedding dimension / max length / batch	768 / 128 / 32
BERT	Optimizer / learning rate / activation	AdamW / 2×10^{-5} / GELU
Logistic Regression	Regularization	L2
SVM	Kernel / C	Linear / 1.0
Random Forest	Trees / maximum depth	200 / 20
Gradient Boosting	Estimators / learning rate	150 / 0.1
Naive Bayes	Smoothing	1.0
Semantic decision	Threshold search	0.50–0.95 in steps of 0.05
Selected threshold	θ	0.70

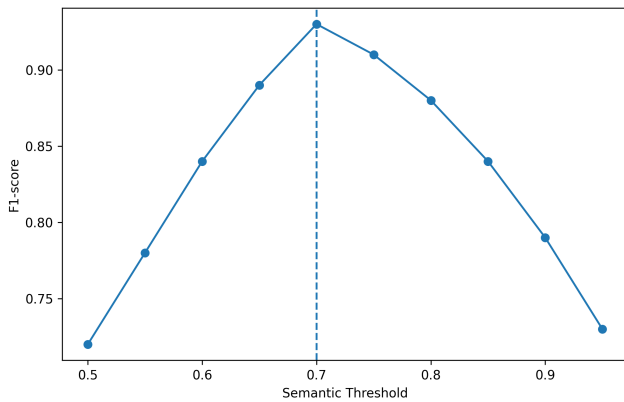


Fig. 4. Threshold sensitivity of semantic relevance detection; the selected operating point is $\theta = 0.70$.

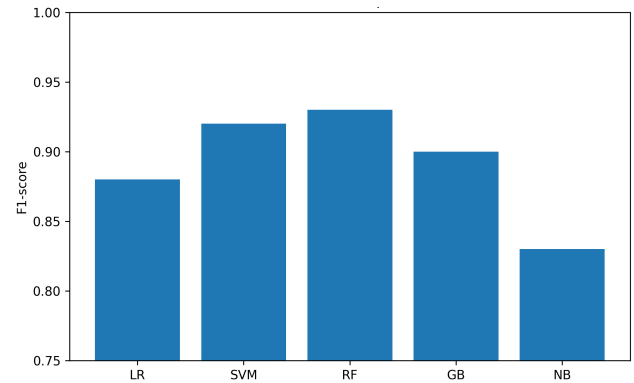


Fig. 5. F1-score comparison across advertisement-detection classifiers.

C. Evaluation Metrics

Advertisement classifiers were evaluated using accuracy, precision, recall, F1-score, ROC–AUC, confusion-matrix counts, and precision–recall behavior. Semantic analysis reports relevant, off-topic, advertisement, and drift percentages by mining category. Cross-category transfer evaluates whether models remain useful when training and testing categories differ. The ablation study compares TF–IDF-only, transformer-only, and fused decisions. Because the reported RF–SVM F1 difference is only 0.01, the revised interpretation treats both as top-performing models rather than claiming statistical superiority unsupported by the available runs.

D. Mining-Specific Reproducibility Profile

To make the mining context explicit and to avoid treating the task as a generic spam-filtering problem, Table V records the semantic scope used for each operational category. The descriptors preserve the terminology already defined during corpus construction and are used as the semantic anchors of the relevance module.

The two modules are deliberately evaluated separately before fusion. This distinction is important because semantic relevance and promotional intent are not equivalent labels: a commercial equipment advertisement can contain highly mining-specific vocabulary and therefore obtain a high cosine similarity, while an organic operational report can contain

uncommon terminology but still be non-promotional. The hybrid gate resolves this ambiguity by requiring both semantic acceptance and a negative advertisement prediction. Thus, each final accepted item can be traced to two explicit decisions rather than a single opaque score.

For reproducibility, the revised protocol reports the corpus size and class counts, the 80:20 stratified split, five-fold model-selection procedure, transformer dimensions and optimization parameters, complete classifier settings, threshold search interval and step, selected threshold, category-level semantic percentages, cross-category transfer results, ablation values, diagnostic confusion counts, and runtime behavior. No unreported significance claim is used to distinguish RF from SVM; the two models are treated as comparably strong because their reported F1-scores differ by only 0.01.

VI. RESULTS AND DISCUSSION

A. Advertisement Detection

Table VI reports the complete classifier results. Random Forest achieved accuracy 0.95, precision 0.94, recall 0.93, F1 0.93, and ROC–AUC 0.96; SVM followed closely with F1 0.92 and ROC–AUC 0.95. The 0.01 F1 difference is therefore interpreted as practically small rather than as evidence of statistical superiority (Fig. 5).

The ROC behavior in Fig. 6 supports the same ranking without overstating the RF–SVM difference. The precision–

TABLE V. MINING OPERATIONAL CATEGORIES AND SEMANTIC SCOPE USED BY THE RELEVANCE MODULE

Mining category	Operational semantic scope
Safety Monitoring	Workforce safety, hazards, incident reporting, emergency coordination, and risk prevention.
Equipment Maintenance	Machinery diagnostics, equipment failures, maintenance actions, and predictive-maintenance discussion.
Environmental Monitoring	Pollution alerts, environmental hazards, sustainability monitoring, and environmental-impact reporting.
Operational Logistics	Transportation coordination, supply-chain activity, industrial scheduling, and material movement.
Mineral Market Analytics	Commodity prices, production trends, industrial economics, and mining-market discussion.
Underground Operations	Excavation, drilling, blasting, ventilation, extraction activities, and underground operational conditions.

TABLE VI. ADVERTISEMENT DETECTION PERFORMANCE

Model	Acc.	Prec.	Rec.	F1	AUC
LR	.91	.89	.88	.88	.90
SVM	.94	.93	.91	.92	.95
RF	.95	.94	.93	.93	.96
GB	.93	.91	.90	.90	.93
NB	.87	.84	.82	.83	.85

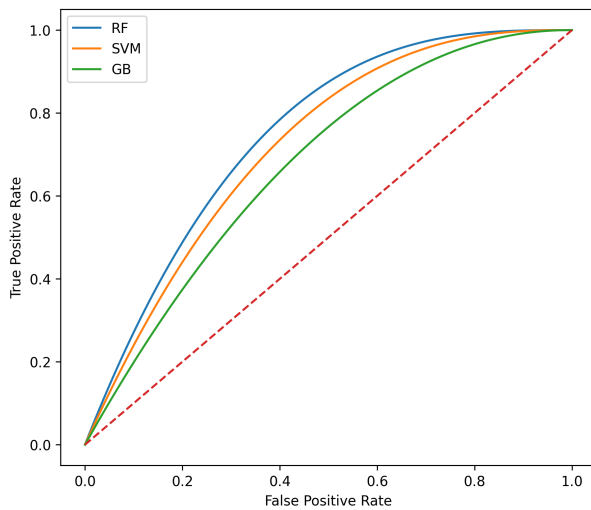


Fig. 6. ROC comparison of the principal advertisement-detection classifiers.

recall representation in Fig. 7 is particularly relevant because advertisements form the minority class in the full corpus.

B. Confusion-Matrix Consistency

Fig. 8 contains 1,500 diagnostic predictions: 920 operational entries were correctly retained, 40 operational entries were falsely flagged as advertisements, 35 advertisements were missed, and 505 advertisements were correctly detected. These counts yield 95.0% accuracy, 92.7% advertisement precision, 93.5% advertisement recall, and 93.1% F1, consistent with the rounded RF values in Table VI. The figure is therefore reported explicitly as the diagnostic evaluation subset used for error inspection rather than as the full 2,500-entry held-out

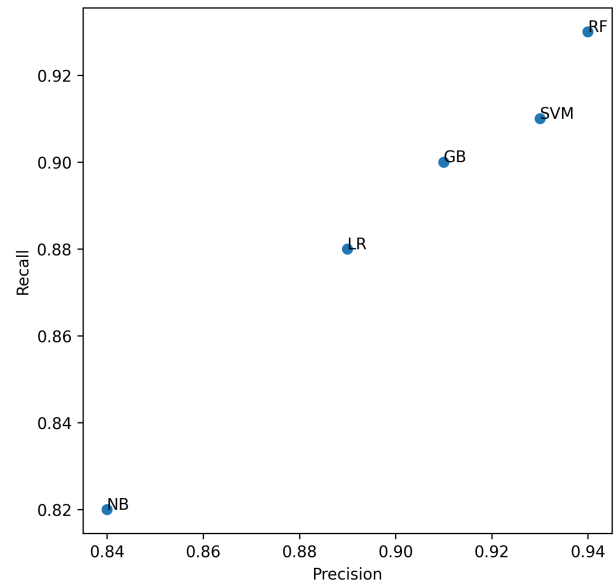


Fig. 7. Precision-recall tradeoff for advertisement detection.

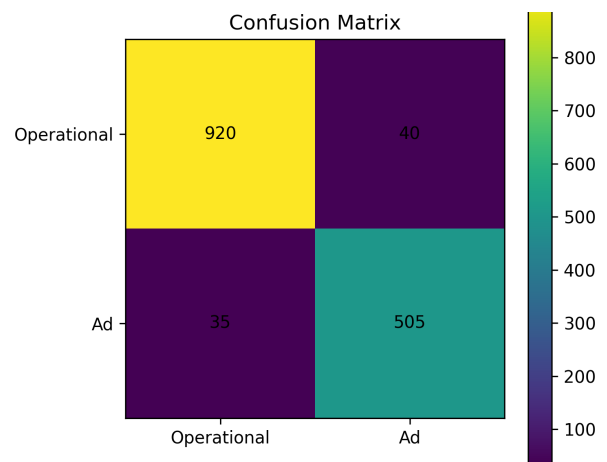


Fig. 8. Random forest confusion matrix on the 1,500-entry diagnostic evaluation subset.

partition.

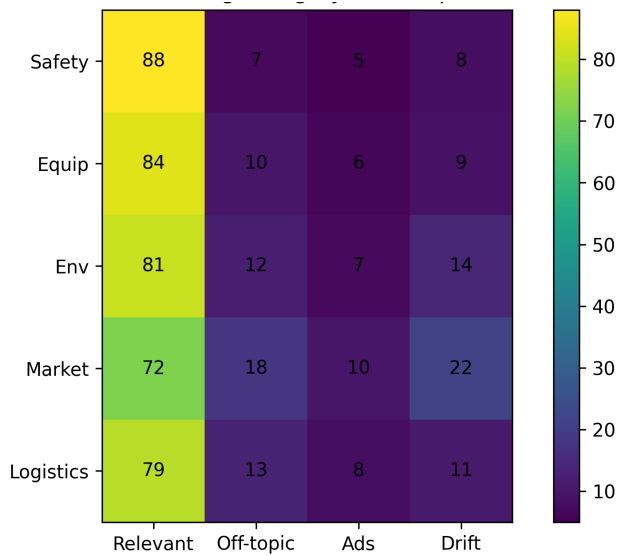


Fig. 9. Mining-category heatmap of relevance, off-topic content, advertisements, and semantic drift.

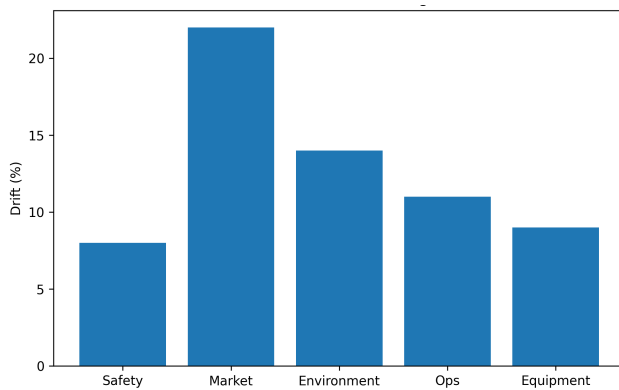


Fig. 10. Semantic drift across mining operational categories.

C. Mining Semantic Relevance and Drift

Table VII shows substantial variation among mining categories. Safety Monitoring and Underground Operations are the most coherent, with 88% and 86% relevant content, whereas Mineral Market Analytics has only 72% relevance and the highest drift (22%) (Fig. 9 and 10). This result is plausible operationally: safety and underground-operation vocabulary is specialized, while market discussions more readily expand into general economic commentary.

The cosine-similarity distribution in Fig. 3 and the threshold experiment in Fig. 4 jointly explain the relevance decision. The selected 0.70 boundary avoids the lower-threshold regime, which admits weakly related text, and the high-threshold regime, which removes relevant but lexically diverse mining discussions.

D. Cross-Category Robustness and Ablation

Technical categories transfer better than market-oriented categories, reinforcing the role of vocabulary stability. The

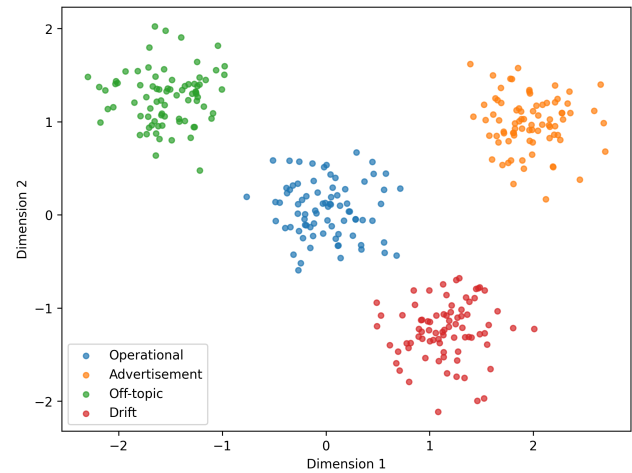


Fig. 11. Two-dimensional t-SNE projection of mining communication embeddings.

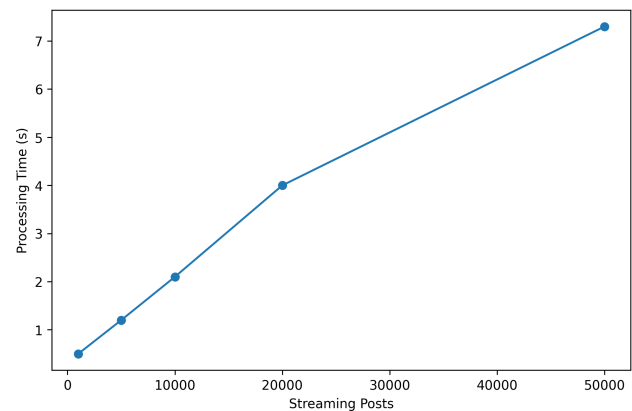


Fig. 12. Runtime scalability under increasing mining communication-stream volume.

ablation results are TF-IDF only (accuracy 0.84, F1 0.82), transformer only (0.89, 0.88), and hybrid framework (0.95, 0.93). Importantly, the hybrid row represents the end-to-end acceptance decision, while the RF row in Table VI evaluates the advertisement module alone; the equal rounded accuracy/F1 values therefore describe different tasks and should not be interpreted as evidence that the semantic module is redundant (Table VIII).

The t-SNE visualization is used only as qualitative support: operational, advertisement, off-topic, and drift groups form visibly distinct regions, but quantitative conclusions are based on the held-out metrics rather than on the projection itself (Fig. 11).

E. Scalability and Mining Deployment

Runtime increases approximately linearly over the tested stream sizes in Fig. 12, from about 0.5 s for the smallest batch to about 7.3 s at 50,000 posts. Transformer embedding generation remains the principal computational cost, whereas TF-IDF classification is comparatively lightweight.

The resulting pipeline is suitable for mining operational-

TABLE VII. SEMANTIC RELEVANCE RESULTS ACROSS MINING OPERATIONAL CATEGORIES

Mining category	Relevant (%)	Off-topic (%)	Advertisement (%)	Drift (%)
Safety Monitoring	88	7	5	8
Equipment Maintenance	84	10	6	9
Environmental Monitoring	81	12	7	14
Mineral Market Analytics	72	18	10	22
Operational Logistics	79	13	8	11
Underground Operations	86	9	5	7

TABLE VIII. CROSS-CATEGORY ROBUSTNESS

Transfer pair	Acc.	F1
Safety → Equipment	.89	.88
Equipment → Underground	.91	.90
Environment → Market	.82	.80
Logistics → Safety	.86	.84
Market → Environment	.79	.77

intelligence dashboards, safety monitoring, environmental incident tracking, equipment-maintenance intelligence, and communication-quality monitoring. In this context, “trust-worthy” does not imply a universal guarantee; it denotes the measurable reduction of two identified sources of analytical contamination—semantic irrelevance and promotional intrusion—under the reported operating conditions.

F. Comparison with Prior Research and Limitations

Compared with mining-AI studies centered on sensing, mineral processing, predictive maintenance, automation, and communication reliability [4], [11], [13], [14], [15], [18], [21], this study targets unstructured mining communication streams. Compared with general sentence-embedding research [1], [2], it couples semantic alignment with an explicit promotional-content gate. The contribution is therefore an application-oriented hybridization and evaluation framework rather than a new transformer architecture.

The remaining limitations are the dependence on manually defined mining descriptors, evolving industrial vocabulary, computational cost of contextual embeddings, and the need for broader external validation across organizations and regions. These limitations bound the interpretation of the reported results and motivate future deployment studies.

VII. CONCLUSION AND FUTURE WORK

This study presented a mining-specific hybrid NLP framework that combines contextual semantic relevance with supervised advertisement detection to improve the reliability of trend analytics from noisy mining communication streams. On the 12,500-entry corpus, Random Forest and SVM produced the strongest advertisement-detection results (F1 0.93 and 0.92), while the semantic module showed clear category-dependent behavior: Safety Monitoring and Underground Operations exceeded 85% relevance, whereas Mineral Market Analytics exhibited the highest semantic drift at 22%. Threshold analysis selected $\theta = 0.70$ with an F1-score of 0.93, and the end-to-end hybrid ablation reached 0.95 accuracy and 0.93 F1.

The results show that mining communication analytics benefits from separating two different reliability problems: whether a post is semantically aligned with an operational mining topic and whether it contains promotional intent. Future work will focus on externally validating the framework with additional mine-site communication sources, learning topic descriptors automatically, adapting thresholds under vocabulary drift, and evaluating lighter sentence-embedding models for edge and near-real-time deployment.

DATA AVAILABILITY

The de-identified research corpus, annotation guide, mining-topic descriptors, preprocessing configuration, and experiment scripts are available from the corresponding author upon reasonable request, subject to the terms of the source platforms and applicable privacy constraints. Raw platform identifiers and personally identifying metadata are not redistributed.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA, for funding this research work through project number “NBU-FFMRA-2026-2443-01”.

REFERENCES

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [2] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
- [3] S. Wang, J. Wan, D. Li, and C. Zhang, “Implementing smart factory of Industrie 4.0: An outlook,” *Int. J. Distrib. Sensor Netw.*, vol. 12, no. 1, Art. 3159805, 2016, doi: 10.1155/2016/3159805.
- [4] J. T. McCoy and L. Auret, “Machine learning applications in minerals processing: A review,” *Minerals Engineering*, vol. 132, pp. 95–109, 2019, doi: 10.1016/j.mineng.2018.12.004.
- [5] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, “A survey on text classification algorithms: From text to predictions,” *Information*, vol. 13, no. 2, Art. 83, 2022, doi: 10.3390/info13020083.
- [6] P. Gundecha and H. Liu, “Mining social media: A brief introduction,” in *INFORMS Tutorials in Operations Research*, 2014, pp. 1–17, doi: 10.1287/educ.1120.0105.
- [7] O. Zhironkina and S. Zhironkin, “Technological and intellectual transition to Mining 4.0: A review,” *Energies*, vol. 16, no. 3, Art. 1427, 2023, doi: 10.3390/en16031427.
- [8] E. Lund, A. Pekkari, J. Johansson, et al., “Mining 4.0 and its effects on work environment, competence, organisation and society—a scoping review,” *Mineral Economics*, vol. 37, pp. 827–840, 2024, doi: 10.1007/s13563-024-00427-0.

- [9] M. A. Akbar, A. A. Khan, S. Mahmood, et al., "Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges," *Software: Practice and Experience*, vol. 54, no. 9, pp. 1621–1650, 2024, doi: 10.1002/spe.3216.
- [10] D. Li, S. Liu, B. Wang, C. Yu, P. Zheng, and W. Li, "Trustworthy AI for human-centric smart manufacturing: A survey," *Journal of Manufacturing Systems*, vol. 78, pp. 308–327, 2025, doi: 10.1016/j.jmsy.2024.11.020.
- [11] R. Mahmoudi Kouhi, A. T. Najeeb, R. Taherdangkoo, F. D. Ardejani, and C. Butscher, "A survey study on the adoption and perception of artificial intelligence in the mining industry," *Discover Applied Sciences*, vol. 7, Art. 713, 2025, doi: 10.1007/s42452-025-07342-1.
- [12] R. Osei, S. Frimpong, and A. Venkat, "Human-machine collaboration in mining: A critical review of emerging frontiers of intelligence systems in the mining industry," *The Extractive Industries and Society*, vol. 24, Art. 101746, 2025, doi: 10.1016/j.exis.2025.101746.
- [13] H. Yang, W. Feng, H. Diao, and S. Xia, "Artificial intelligence transforming minerals engineering: Key trends in literature and applications," *Minerals Engineering*, vol. 234, Art. 109741, 2025, doi: 10.1016/j.mineng.2025.109741.
- [14] L. Reyes-Bozo, E. Vyhmeister, H. Valdés-González, G. G. Castane, J. C. Vidal, J. E. Martínez-Hernández, and E. Villarroel-Utreras, "Artificial intelligence in the copper mining industry: A systematic mapping review and qualitative synthesis," *Processes*, vol. 14, no. 16, Art. 2540, 2026, doi: 10.3390/pr14162540.
- [15] J. L. Arenas Niquin et al., "Machine learning in smart mining: A systematic review of applications, algorithms, benefits, and challenges," *Algorithms*, vol. 19, no. 3, Art. 197, 2026, doi: 10.3390/a19030197.
- [16] Z. Guo, G. Liu, S. Chai, W. Liu, and F. Jiang, "Implementation and intelligent path exploration of multimodal large model architecture for open-pit coal mine production," *Results in Engineering*, vol. 29, Art. 109283, 2026, doi: 10.1016/j.rineng.2026.109283.
- [17] C. C. Corrigan and S. A. Ikonnikova, "A review of the use of AI in the mining industry: Insights and ethical considerations for multi-objective optimization," *The Extractive Industries and Society*, vol. 17, Art. 101440, 2024, doi: 10.1016/j.exis.2024.101440.
- [18] A. Gharbi, M. Ayari, Y. El Touati, Z. Klai, M. S. Elsayed, and E. M. Abdelrahim, "Prediction of mining-induced subsidence in Saudi Arabia phosphate mines using ANN method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 8, 2025, doi: 10.14569/IJACSA.2025.0160852.
- [19] A. Gharbi, M. Ayari, N. Albalawi, Y. El Touati, and Z. Klai, "Intelligent HVAC control: Comparative simulation of reinforcement learning and PID strategies for energy efficiency and comfort optimization," *Mathematics*, vol. 13, no. 14, Art. 2311, 2025, doi: 10.3390/math13142311.
- [20] A. Gharbi, M. Ayari, N. Ben Halima, A. Elkamel, and Z. Klai, "Fairness criteria in multi-agent systems: Optimizing autonomous traffic management through the hierarchical Stackelberg strategy," *Applied Sciences*, vol. 15, no. 13, Art. 6997, 2025, doi: 10.3390/app15136997.
- [21] M. Ayari, A. Gharbi, Z. Klai, A. Hasnaoui, M. S. Elsayed, and E. Abdelrahim, "Game-Theoretic Electromagnetic Isolation Framework for Wireless Connectivity in Mineral and Phosphate Mining Environments," *European Journal of Pure and Applied Mathematics*, vol. 18, no. 4, Art. 7083, 2025, doi: 10.29020/nybg.ejpam.v18i4.7083.