

Detecting Information-Hiding Lead Generation Behaviors in Social Q&A Platforms: A Multi-Modal Deep Learning Approach

Yang Dong

Beijing International Studies University, 100024, Beijing, China.

Abstract The proliferation of consultation-oriented social platforms has created unprecedented opportunities for information sharing, yet has also given rise to strategic information-hiding behaviors that undermine user experience and platform integrity. This paper addresses a critical but understudied problem: the automatic detection of "lead generation" posts that deliberately withhold essential information (such as pricing, resources, or contact details) to force users into private communication channels. We propose a novel multi-modal deep learning framework, termed IH-LGDetector, that integrates textual semantic analysis, visual content understanding, and user behavioral sequence modeling to identify such manipulative content. Our approach introduces an Information Deficiency Quantification Index (IDQI) that mathematically measures the degree of information incompleteness in posts, enabling fine-grained analysis of the relationship between information hiding and user engagement patterns. We construct a comprehensive dataset comprising 127,843 posts from a major Chinese social Q&A platform, with meticulous annotation of lead generation behaviors. Extensive experiments demonstrate that our proposed framework achieves 94.7% accuracy and 0.967 AUC-ROC, significantly outperforming baseline methods. Furthermore, correlation analysis reveals a strong positive relationship ($r=0.78$, $p<0.001$) between IDQI scores and private message trigger rates, validating our theoretical framework. The proposed system offers immediate deployability for platform governance, with inference latency under 15ms per post, representing a substantial contribution to content ecosystem management and user experience optimization.

Keywords Social Q&A platforms; information hiding detection; user behavior modeling; multi-modal learning; anomaly detection; lead generation; platform governance; user experience optimization; deep learning; content moderation

Received: September 3, 2026

Revised: September 21, 2026

Accepted: September 23, 2026

Published: October 22, 2026

Copyright: © 2026 by the authors.

Licensee Axon Academic

Publishing Institute, Hong Kong,

China. This article is an open access

article distributed under the terms

and conditions of the Creative

Commons Attribution (CC BY)

license

(<http://creativecommons.org/licenses/by/4.0/>).

DOI:10.67148/fahss-2026-262

1. Introduction

1.1. Background and Motivation

The emergence of consultation-oriented social platforms, exemplified by Xiaohongshu (The Rednote), Zhihu, and Dianping in China, has fundamentally transformed how users seek and share experiential knowledge. These platforms have evolved into essential information resources where users post detailed guides, reviews, and recommendations spanning diverse domains including real estate, home decoration, educational courses, software tutorials, and lifestyle services. The core value proposition of such platforms lies in their ability to provide comprehensive, actionable information that enables users to make informed decisions without requiring additional information-seeking efforts. According to recent industry reports, Xiaohongshu alone hosts over 300 million monthly active users who generate approximately 10 million new posts daily, underscoring the platform's significance as an information ecosystem.

However, this information-sharing paradigm has been increasingly exploited by content creators seeking to divert platform traffic to private communication channels for commercial purposes. A pervasive phenomenon has emerged wherein posts deliberately withhold critical information—such as pricing details, resource links, contact information, or specific recommendations—while explicitly or implicitly directing users to "private message for details," "contact me for resources," or "say hello to get the price." This strategic information hiding transforms public knowledge-sharing platforms into lead generation funnels, fundamentally contradicting the platforms' stated purpose of open information exchange. Our preliminary analysis of 50,000 randomly sampled posts revealed that approximately 23.4% exhibited some form of intentional information withholding, with particularly high concentrations in commercial categories such as real estate (41.2%), home decoration (38.7%), and educational courses (35.6%).

The technical implications of this phenomenon extend beyond mere user inconvenience. From an information retrieval perspective, such posts pollute search results with incomplete content, forcing users to engage in additional rounds of information seeking. From a recommendation system standpoint, these posts may receive elevated engagement metrics (comments, saves) driven by users seeking missing information, thereby distorting content ranking algorithms. From a platform governance angle, the migration of interactions to private channels reduces platform visibility into user activities, complicating content moderation and quality control efforts. Most critically, this behavior erodes user trust in platform content, potentially accelerating user churn and diminishing platform value.

1.2. Problem Statement

The information-hiding lead generation (IH-LG) behavior presents a multifaceted technical challenge that cannot be adequately addressed by conventional spam detection or content moderation approaches. Unlike traditional spam, IH-LG posts often contain genuinely useful partial information, making binary classification insufficient. The behavior manifests through subtle linguistic patterns, strategic visual content choices, and specific user interaction dynamics that require sophisticated multi-modal analysis to detect accurately.

We identify four core technical research questions that this paper addresses: First, the automatic identification problem: How can we leverage natural language processing, multi-modal learning, and user behavior sequence modeling techniques to automatically distinguish IH-LG posts from legitimate information-sharing content? Second, the quantification problem: How can we mathematically measure the degree of information deficiency in posts and model its correlation with user experience degradation metrics such as search efficiency decline, private message triggers, and user churn propensity? Third, the impact analysis problem: What are the technical consequences of IH-LG behaviors on platform content distribution mechanisms, recommendation system fairness, and overall user engagement patterns? Fourth, the governance problem: How can we construct deployable algorithmic detection and intervention frameworks that can be integrated into existing platform governance systems?

1.3. Research Contributions

This paper makes the following significant contributions to the field of social platform content analysis and governance:

1. **Novel Problem Formulation:** To the best of our knowledge, this is the first systematic study of information-hiding lead generation behaviors in social Q&A platforms from a computational perspective. We provide a rigorous formal definition of the IH-LG detection problem and establish its connections to related research areas including spam detection, information quality assessment, and user behavior modeling.
2. **Multi-Modal Detection Framework:** We propose IH-LGDetector, a novel deep learning architecture that integrates textual semantic features extracted through fine-tuned BERT, visual content features through a Vision Transformer with OCR fusion, and user behavioral sequence patterns through a Transformer-based encoder. The framework achieves state-of-the-art detection performance while maintaining computational efficiency suitable for real-time deployment.
3. **Information Deficiency Quantification Index (IDQI):** We introduce a principled metric for quantifying the degree of information incompleteness in posts, based on information-theoretic principles and domain-specific knowledge graphs. This metric enables fine-grained analysis of the spectrum of information-hiding behaviors and their differential impacts on user engagement.
4. **Comprehensive Dataset:** We construct and release a large-scale annotated dataset comprising 127,843 posts with detailed labels for IH-LG behaviors, information deficiency scores, and user interaction traces. This dataset fills a significant gap in available resources for platform governance research.
5. **Actionable Insights for Platform Governance:** Through extensive experiments and correlation analyses, we provide empirical evidence of the negative impacts of IH-LG behaviors on user experience metrics and offer concrete recommendations for platform governance strategies, including detection thresholds, intervention timing, and fairness considerations.

1.4 Paper Organization

The remainder of this paper is organized as follows: Section II provides a comprehensive review of related work in social platform content analysis, multi-modal

learning, information hiding detection, and user behavior modeling. Section III details our methodology, including dataset construction, feature engineering, and the proposed IH-LGDetector framework. Section IV presents experimental design and results, including comparative evaluations, ablation studies, and correlation analyses. Section V discusses technical considerations including model interpretability, adversarial robustness, and deployment feasibility. Section VI concludes the paper and outlines directions for future research.

2. Related Work

This section reviews relevant literature across six interconnected research domains that provide the theoretical and technical foundations for our work.

2.1 Social Platform Content Analysis

The analysis of user-generated content (UGC) on social platforms has attracted substantial research attention, driven by the need to understand information dissemination patterns, content quality dynamics, and user engagement behaviors. Early work by Agichtein et al. [1] established frameworks for identifying high-quality content in community question-answering platforms through feature engineering combining textual, structural, and community-based signals. More recent approaches have leveraged deep learning techniques for content quality assessment, with Chen et al. [2] proposing a hierarchical neural network that captures both sentence-level and document-level semantic features for answer quality prediction on Stack Overflow.

Research specifically targeting Chinese social platforms has emerged as a distinct subfield, with particular focus on Xiaohongshu's unique content ecosystem. Liu et al. [3] conducted a large-scale analysis of Xiaohongshu content characteristics, identifying distinct content categories and their associated engagement patterns. Their work revealed significant variations in information completeness across categories, with commercial content exhibiting systematically lower information density. Wang et al. [4] extended this line of research by developing a content value assessment framework that explicitly considers information utility, finding that posts with incomplete information received paradoxically higher engagement in certain categories—a phenomenon they attributed to curiosity-driven interaction.

The intersection of content analysis and platform governance has gained prominence as platforms grapple with content quality challenges. Zeng et al. [5] proposed a multi-task learning framework for simultaneously detecting multiple content quality issues, including spam, low-quality content, and clickbait. However, their framework did not address the more subtle phenomenon of strategic information hiding, which presents unique detection challenges due to its partial compliance with platform content standards.

2.2. Multi-Modal Learning for Content Understanding

Multi-modal learning has emerged as a powerful paradigm for content understanding, enabling the integration of heterogeneous information sources including text, images, and user interactions. The introduction of vision-language pre-trained models such as CLIP [6] and ViLT [7] has revolutionized cross-modal representation learning, enabling zero-shot transfer to downstream tasks with remarkable effectiveness.

For social media content analysis, these models have been adapted to capture the complex relationships between textual descriptions and visual elements that characterize modern social posts.

Recent work has explored multi-modal approaches for detecting deceptive or manipulative content. Zhou et al. [8] proposed a multi-modal fake news detection framework that fuses textual and visual features through a cross-modal attention mechanism, achieving significant improvements over text-only baselines. Similarly, Wang et al. [9] developed a multi-modal approach for detecting clickbait posts that analyzes the consistency between post titles and image content, identifying cases where visual elements are used to attract attention without providing substantive information.

For social Q&A platforms specifically, the integration of OCR-extracted text from images has proven crucial for comprehensive content understanding. Huang et al. [10] demonstrated that a significant proportion of information in Xiaohongshu posts is conveyed through text embedded in images, including price tags, product specifications, and contact information. Their OCR-enhanced content analysis framework achieved substantial improvements in information extraction accuracy, highlighting the importance of visual text processing for platform content analysis.

2.3 Information Hiding and Adversarial Detection

Information hiding techniques have been extensively studied in the context of steganography and digital forensics, with recent attention extending to social media contexts. Traditional steganography research [11, 12] focuses on embedding hidden messages within cover media, while detection methods aim to identify statistical anomalies indicative of hidden content. However, the information hiding behaviors addressed in this paper differ fundamentally from classical steganography: rather than concealing messages within content, IH-LG posts strategically omit information that users would reasonably expect to find.

A more relevant body of work addresses adversarial behaviors in online platforms, where users deliberately manipulate content to evade detection systems. Wu et al. [13] studied adversarial content creation in e-commerce reviews, identifying linguistic strategies used to disguise fake reviews as genuine. Their detection framework incorporated adversarial training to improve robustness against evolving evasion tactics. Similarly, Chen et al. [14] analyzed strategic behaviors in social media marketing posts, finding that content creators systematically exploit platform recommendation algorithms through engagement-baiting tactics.

The concept of "dark patterns" in user interface design provides a theoretical framework for understanding IH-LG behaviors. Mathur et al. [15] conducted a large-scale analysis of shopping websites, cataloging various manipulative design patterns including "confirmshaming," "forced continuity," and "obstruction." While their work focused on interface-level manipulation, the underlying principle—deliberately creating friction to drive specific user behaviors—directly parallels the IH-LG phenomenon, where information withholding creates artificial barriers to information access.

2.4 User Behavior Modeling and Sequence Analysis

User behavior modeling has become central to understanding and predicting engagement patterns on social platforms. Sequential recommendation models, pioneered by RNN-based approaches [16] and subsequently enhanced by Transformer architectures [17], capture temporal dependencies in user interaction histories to predict future behaviors. These techniques have proven particularly effective for modeling the complex decision processes underlying user engagement with platform content.

Recent work has extended behavior modeling to detect anomalous or malicious user activities. Wang et al. [18] proposed a graph neural network approach for identifying fake accounts by modeling their interaction patterns within the user-content bipartite graph. Their method achieved high detection accuracy by capturing structural anomalies in interaction networks. For content-level anomaly detection, Liu et al. [19] developed a self-supervised learning framework that identifies unusual posting behaviors by learning normal behavior distributions from large-scale user activity logs.

The relationship between content characteristics and user behavioral responses has been explored in the context of engagement prediction. Zhang et al. [20] found that information completeness significantly affects user satisfaction and subsequent engagement behaviors, with incomplete information triggering additional search activities. This finding provides empirical support for our hypothesis that IH-LG behaviors systematically alter user interaction patterns in measurable ways.

2.5 Anomaly Detection and Fraud Identification

Anomaly detection techniques provide a foundation for identifying content that deviates from expected patterns. Classical approaches including Isolation Forest [21], One-Class SVM [22], and autoencoder-based methods [23] have been successfully applied to various domains including network intrusion detection, financial fraud identification, and manufacturing quality control. For social platform applications, these methods have been adapted to detect fake accounts, spam content, and coordinated manipulation campaigns.

Deep learning approaches have significantly advanced anomaly detection capabilities. Zong et al. [24] proposed a deep autoencoding Gaussian mixture model for unsupervised anomaly detection, achieving state-of-the-art performance on multiple benchmark datasets. For sequential data, Malhotra et al. [25] developed an LSTM-based encoder-decoder framework that learns to reconstruct normal sequences, flagging sequences with high reconstruction error as anomalies. These techniques are directly applicable to modeling normal posting behaviors and identifying deviations indicative of IH-LG strategies.

Graph-based anomaly detection has emerged as a powerful approach for identifying coordinated behaviors. Liu et al. [26] proposed a Graph Convolutional Network framework for detecting fraudulent user groups by analyzing their collective interaction patterns. Their method identified anomalous subgraph structures indicative of coordinated manipulation, achieving significant improvements over node-level detection approaches. This graph-based perspective is particularly relevant for IH-LG detection, as such behaviors often exhibit characteristic patterns in the user-content interaction graph.

The algorithmic governance of platform content has become a critical research area as platforms increasingly rely on automated systems for content moderation and curation. Roberts [27] provided a comprehensive analysis of commercial content moderation practices, highlighting the challenges of scale, accuracy, and fairness that platforms face. From a technical perspective, content moderation systems must balance multiple objectives including detection accuracy, processing latency, and user experience impact.

Recommendation system fairness has emerged as a significant concern, with research demonstrating that algorithmic curation can systematically disadvantage certain content creators or topics. Beutel et al. [28] proposed fairness-aware recommendation methods that ensure equitable exposure across different content categories. Their work highlighted the importance of considering fairness implications when designing content filtering systems, as interventions targeting specific content types may have unintended distributional consequences.

Recent work has begun addressing the intersection of content quality and recommendation dynamics. Wei et al. [29] demonstrated that low-quality content can achieve elevated recommendation rankings through engagement manipulation, creating a feedback loop that amplifies content quality issues. Their proposed solution incorporated content quality signals into recommendation objectives, improving overall content ecosystem health. This line of research directly motivates our work on IH-LG detection, as such behaviors may similarly distort recommendation dynamics through artificially inflated engagement metrics.

3. Methodology

This section presents our comprehensive methodology for detecting information-hiding lead generation behaviors, including dataset construction, feature engineering, the proposed IH-LGDetector framework, and evaluation metrics.

3.1 Dataset Construction

3.1.1 Data Collection

We constructed a comprehensive dataset through systematic collection from Xiaohongshu, one of China's largest social Q&A platforms with over 300 million monthly active users. Data collection was conducted in compliance with platform terms of service, accessing only publicly available content through authorized API endpoints and respecting robots.txt directives. The collection process employed a stratified sampling strategy to ensure coverage across multiple content categories and time periods.

The sampling strategy targeted categories with high prevalence of IH-LG behaviors, as identified through preliminary analysis: real estate (property listings, rental guides), home decoration (renovation guides, furniture recommendations), educational courses (skill training, certification programs), software tutorials (technical guides, resource sharing), and lifestyle services (beauty, fitness, travel). For each category, we collected posts spanning the period from January 2022 to December 2023, enabling analysis of temporal patterns and behavior evolution.

The final dataset comprises 127,843 posts with complete metadata including: post text content, associated images (up to 9 per post), user interaction metrics (likes, comments, saves, shares), user profile information, and temporal attributes. Additionally, we collected user interaction sequences for 45,672 unique users who engaged with the sampled posts, enabling behavioral analysis at the user level.

3.1.2 Annotation Protocol

We developed a rigorous annotation protocol to identify IH-LG behaviors, combining rule-based pre-filtering with human expert annotation. The annotation guidelines defined IH-LG posts according to three criteria: (1) presence of explicit or implicit directives for private communication (e.g., "DM for price," "contact me for details," "say hello to get resources"); (2) absence of critical information that users would reasonably expect based on post context (e.g., price for product recommendations, download links for resource posts, specific recommendations for guide posts); and (3) evidence of commercial intent inferred from post content and user profile characteristics.

The annotation process employed a two-stage approach. In the first stage, rule-based pre-filtering identified candidate IH-LG posts using keyword matching for common directive phrases and heuristic scoring for information completeness. This reduced the annotation workload while ensuring comprehensive coverage. In the second stage, trained annotators reviewed candidate posts and a random sample of non-candidate posts to verify labels and identify false negatives from the rule-based filter.

Inter-annotator agreement was measured on a held-out validation set of 5,000 posts annotated independently by three annotators. The resulting Fleiss' kappa of 0.84 indicates substantial agreement, supporting the reliability of the annotation protocol. Disagreements were resolved through discussion and consensus, with ambiguous cases excluded from the final dataset. The final annotated dataset contains 29,847 IH-LG posts (23.3%) and 97,996 normal posts (76.7%), reflecting the prevalence of this behavior on the platform.

3.1.3 Data Splitting

The dataset was partitioned into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain class distribution across splits. Temporal integrity was preserved by ensuring that posts from later time periods appeared only in validation and test sets, enabling evaluation of model generalization to future content. User-level splitting was employed for behavioral modeling tasks, ensuring that all posts from a given user appeared in a single split to prevent information leakage.

3.2 Feature Engineering

Our feature engineering approach extracts comprehensive signals from textual content, visual elements, user behaviors, and structural properties to enable accurate IH-LG detection.

3.2.1 Textual Features

Textual feature extraction operates at multiple granularity levels. At the character level, we extract TF-IDF representations (10,000 dimensions) capturing distinctive vocabulary patterns associated with IH-LG behaviors. At the word level, we employ pre-trained Word2Vec embeddings (300 dimensions) trained on a large corpus of platform posts to capture semantic relationships. At the sentence level, we utilize BERT-base-Chinese to extract contextualized representations (768 dimensions) that capture nuanced semantic patterns.

Beyond embedding-based features, we extract several handcrafted textual features informed by linguistic analysis of IH-LG behaviors: (1) directive phrase presence, identified through a curated lexicon of 237 phrases commonly used to direct users to private channels; (2) information category indicators, marking the presence or absence of expected information types (price, link, contact, recommendation) based on post category; (3) linguistic markers of uncertainty and vagueness, including hedging expressions and placeholder phrases; and (4) sentiment and subjectivity scores computed using a fine-tuned sentiment analysis model.

Topic modeling features are extracted using Latent Dirichlet Allocation (LDA) with 50 topics, capturing the thematic content of posts. We compute topic distribution vectors and derive features measuring topic concentration (entropy of topic distribution) and topic-category alignment (similarity between post topic distribution and category-specific topic priors).

3.2.2 Visual Features

Visual feature extraction addresses both the semantic content of images and text embedded within images. For semantic visual features, we employ a Vision Transformer (ViT-B/16) pre-trained on ImageNet-21k, extracting 768-dimensional feature vectors from the [CLS] token of each image. For posts with multiple images, we compute both individual image features and aggregated representations through attention-weighted pooling.

OCR-based features are extracted using PaddleOCR, a state-of-the-art Chinese OCR system. For each image, we extract all text regions, their spatial coordinates, and confidence scores. From OCR results, we derive: (1) OCR text embeddings using the same BERT encoder as post text, enabling cross-modal comparison; (2) text-image consistency scores measuring the semantic alignment between post text and OCR-extracted text; and (3) information presence indicators derived from OCR text, identifying whether critical information types appear in images.

We additionally extract visual style features capturing aesthetic properties that may correlate with IH-LG behaviors: color histogram features (64 dimensions), image quality metrics (blur, noise levels), and the presence of text-heavy vs. photograph content. These features help identify posts that use visual elements strategically to attract attention without providing substantive information.

3.2.3 Behavioral Features

User behavioral features capture interaction patterns that may indicate IH-LG behaviors. At the post level, we compute engagement metrics including like rate (likes per view), comment rate, save rate, and share rate. Critically, we also measure private

message trigger rate—the proportion of viewers who initiate private messages after viewing the post—obtained through aggregated platform analytics with appropriate privacy protections.

User-level behavioral features capture the posting history and interaction patterns of content creators. We extract: (1) posting frequency and temporal patterns; (2) historical engagement rates across posts; (3) content category distribution in posting history; (4) private message response patterns; and (5) follower growth trajectory. These features help distinguish systematic IH-LG practitioners from occasional or unintentional information withholding.

Interaction sequence features are extracted from user engagement histories using a Transformer-based sequence encoder. For each user, we construct sequences of their interactions with posts, encoding post features and interaction types (view, like, comment, save, share, private message). The encoder produces fixed-dimensional representations capturing behavioral patterns that predict future IH-LG engagement.

3.2.4 Structural Features

Structural features capture the position and connectivity patterns of posts within the user-content interaction graph. We construct a bipartite graph with user nodes and post nodes, connected by interaction edges. From this graph, we extract: (1) node centrality measures (degree, betweenness, PageRank) for both users and posts; (2) community membership features derived from graph clustering; and (3) local neighborhood statistics capturing the density and composition of interactions around each post.

Graph neural network features are pre-computed using a 2-layer Graph Convolutional Network (GCN) trained on the node classification task. The GCN takes initial node features (aggregated content features for posts, aggregated interaction features for users) and produces 128-dimensional embeddings that capture structural context. These embeddings enable the detection model to leverage network-level patterns that may not be apparent from content features alone.

3.3 Information Deficiency Quantification Index (IDQI)

A key contribution of this work is the Information Deficiency Quantification Index (IDQI), a principled metric for measuring the degree of information incompleteness in posts. The IDQI is designed to capture the gap between information expected by users and information actually provided, enabling fine-grained analysis of IH-LG behaviors beyond binary classification.

Formally, let C denote a content category and $I = \{i_1, i_2, \dots, i_k\}$ denote the set of information types expected for posts in category C . For example, for real estate posts, I might include {price, location, area, contact}. For each information type i_j , we define a presence function $p(i_j, \text{post}) \in [0, 1]$ indicating the degree to which that information is present in the post, computed through a combination of rule-based extraction and neural classification.

The IDQI is then computed as a weighted combination of information absence scores:

$$IDQI(\text{post}) = \sum_j w_j \times (1 - p(i_j, \text{post})) \times s(i_j, \text{post})$$

where w_j represents the importance weight for information type i_j (learned from user survey data on information priorities), and $s(i_j, \text{post})$ represents a specificity score capturing whether the present information is specific (e.g., exact price) or vague (e.g., "affordable"). The IDQI ranges from 0 (complete information) to 1 (maximal information deficiency).

The information type weights are derived from a user survey ($n=2,500$) asking users to rate the importance of different information types for decision-making in each category. The specificity scoring function is implemented as a neural classifier trained on annotated examples of specific vs. vague information presentations. This approach enables the IDQI to capture nuanced cases where information is technically present but presented in an unhelpful manner.

3.4 IH-LGDetector Framework

We propose IH-LGDetector, a multi-modal deep learning framework that integrates textual, visual, behavioral, and structural features for comprehensive IH-LG detection. The framework architecture comprises four main components: text encoder, visual encoder, behavior encoder, and multi-modal fusion module.

3.4.1 Text Encoder

The text encoder processes post textual content through a fine-tuned RoBERTa-base-Chinese model. We initialize with pre-trained weights and fine-tune on our IH-LG detection task, enabling the model to learn task-specific representations. The encoder takes tokenized post text as input and produces a 768-dimensional [CLS] representation capturing the semantic content of the entire post.

To incorporate handcrafted textual features, we employ a feature fusion layer that concatenates the [CLS] embedding with normalized handcrafted feature vectors, followed by a fully-connected layer that produces the final text representation $h_{\text{text}} \in \mathbb{R}^{512}$. This hybrid approach combines the representational power of pre-trained language models with the interpretability and domain knowledge encoded in handcrafted features.

3.4.2 Visual Encoder

The visual encoder processes post images through a dual-stream architecture. The first stream employs a Vision Transformer (ViT-B/16) to extract semantic visual features from each image. For posts with multiple images, we apply attention-weighted pooling to aggregate image features into a single 768-dimensional vector. The attention weights are learned to emphasize images most relevant to IH-LG detection.

The second stream processes OCR-extracted text through the same RoBERTa encoder used for post text, enabling cross-modal comparison. We compute text-image consistency features by measuring the semantic similarity between post text and OCR text embeddings. The visual representation $h_{\text{visual}} \in \mathbb{R}^{512}$ is produced by a fusion layer combining ViT features, OCR text features, and consistency scores.

3.4.3 Behavior Encoder

The behavior encoder models user interaction sequences using a Transformer architecture. For each user, we construct a sequence of their historical interactions, where each interaction is represented by a feature vector combining post features and interaction type embeddings. The Transformer encoder applies multi-head self-attention to capture temporal dependencies and interaction patterns.

The encoder produces a behavioral representation $h_{\text{behavior}} \in \mathbb{R}^{256}$ that captures the user's characteristic interaction patterns. For new users with limited interaction history, we employ a cold-start strategy using population-level behavioral priors. The behavior encoder is particularly important for identifying systematic IH-LG practitioners whose behavioral patterns differ from typical content creators.

3.4.4 Multi-Modal Fusion Module

The multi-modal fusion module integrates representations from all encoders through a cross-modal attention mechanism. We employ a Transformer-based fusion architecture where text, visual, and behavior representations serve as queries, keys, and values in alternating attention layers. This enables the model to learn complex interactions between modalities—for example, identifying cases where visual content contradicts textual claims or where behavioral patterns suggest systematic information hiding.

The fused representation is passed through a classification head comprising two fully-connected layers with ReLU activation and dropout ($p=0.3$). The final output is a probability score indicating the likelihood of IH-LG behavior. The model is trained end-to-end using binary cross-entropy loss with label smoothing ($\epsilon=0.1$) to improve generalization.

3.4.5 Multi-Task Learning Extension

To improve feature learning and enable additional analytical capabilities, we extend the framework with multi-task learning objectives. In addition to the primary IH-LG classification task, the model is trained to predict: (1) IDQI score (regression task); (2) private message trigger probability (regression task); and (3) information type presence indicators (multi-label classification). These auxiliary tasks share the same encoder weights but have task-specific prediction heads.

The multi-task loss function is:

$$L_{\text{total}} = L_{\text{class}} + \lambda_1 L_{\text{IDQI}} + \lambda_2 L_{\text{PM}} + \lambda_3 L_{\text{info}}$$

where L_{class} is the classification loss, L_{IDQI} is the IDQI regression loss (MSE), L_{PM} is the private message prediction loss (MSE), L_{info} is the information presence multi-label loss (BCE), and $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters controlling task weighting (set to 0.5, 0.3, 0.2 respectively based on validation performance).

3.5 Baseline Models

We compare our proposed framework against a comprehensive set of baseline models spanning traditional machine learning, deep learning, and multi-modal approaches:

1. Traditional ML: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), XGBoost, and LightGBM using handcrafted features only.
2. Text-only Deep Learning: TextCNN, BiLSTM with attention, and fine-tuned BERT-base-Chinese.
3. Multi-modal Baselines: Early fusion (concatenating text and visual features before classification), Late fusion (averaging predictions from separate text and visual models), and CLIP-based zero-shot classification.
4. Anomaly Detection: Isolation Forest, One-Class SVM, and Autoencoder-based anomaly detection trained on normal posts only.

3.6. Evaluation Metrics

We employ multiple evaluation metrics to comprehensively assess model performance:

- Classification Metrics: Accuracy, Precision, Recall, F1-score, and AUC-ROC. For imbalanced evaluation, we additionally report AUC-PR (Area Under Precision-Recall Curve).
- Regression Metrics: For IDQI prediction, we report Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson correlation coefficient (r) between predicted and ground-truth IDQI scores.
- Correlation Analysis: Pearson and Spearman correlation coefficients between IDQI scores and user engagement metrics, with statistical significance testing.
- Efficiency Metrics: Model inference time (ms per post), memory footprint, and throughput (posts per second).

4. Experimental Design and Results

This section presents comprehensive experimental evaluation of the proposed IH-LGDetector framework, including comparative analysis, ablation studies, correlation analysis, and deployment feasibility assessment.

4.1 Experimental Setup

All experiments were conducted on a server equipped with dual NVIDIA A100 GPUs (80GB each), 512GB RAM, and AMD EPYC 7742 CPU. The deep learning models were implemented in PyTorch 2.0 with CUDA 11.8. For traditional machine learning baselines, we used scikit-learn 1.3 with default hyperparameters unless otherwise specified.

Training hyperparameters for IH-LGDetector were selected through grid search on the validation set: learning rate = $2e-5$ (with linear warmup for first 10% of steps), batch size = 64, maximum sequence length = 512, dropout rate = 0.3, weight decay = 0.01. Training proceeded for 20 epochs with early stopping based on validation loss (patience = 3 epochs). The AdamW optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.999$.

4.2 Main Results

Table I presents the comparative performance of all models on the test set. Our proposed IH-LGDetector achieves the best performance across all metrics, with 94.7% accuracy, 93.2% F1-score, and 0.967 AUC-ROC. The improvement over the best baseline

(BERT with late fusion) is statistically significant ($p < 0.001$, McNemar's test).

Table 1 Performance Comparison of Different Models on IH-LG Detection

Model	Acc.	Prec.	Rec.	F1	AUC
LR	78.3%	72.1%	68.4%	70.2%	0.812
SVM	80.1%	74.8%	71.2%	72.9%	0.834
XGBoost	84.2%	79.6%	76.3%	77.9%	0.871
LightGBM	85.4%	80.9%	77.8%	79.3%	0.882
TextCNN	86.7%	82.3%	79.1%	80.7%	0.894
BiLSTM	87.9%	83.8%	81.2%	82.5%	0.903
BERT	90.4%	87.2%	84.6%	85.9%	0.931
Early Fusion	91.2%	88.4%	86.1%	87.2%	0.939
Late Fusion	92.1%	89.7%	87.3%	88.5%	0.945
IH-LGDetector (Ours)	94.7%	92.4%	94.1%	93.2%	0.967

The results demonstrate the effectiveness of our multi-modal approach. Traditional ML models achieve reasonable performance (78-85% accuracy) using handcrafted features alone, indicating that linguistic and behavioral patterns provide discriminative signals. Deep learning models substantially improve performance, with BERT achieving 90.4% accuracy through its learned semantic representations. Multi-modal fusion approaches further improve performance, validating the importance of visual features for IH-LG detection. Our proposed IH-LGDetector achieves the best performance through its integrated architecture that jointly models textual, visual, and behavioral signals.

4.3 Ablation Study

To understand the contribution of each component, we conduct ablation experiments by removing individual modules from IH-LGDetector. Table II presents the results.

Table 2 Ablation Study Results

Model Variant	F1	AUC	Δ Performance
Full Model	93.2%	0.967	-
w/o Visual Features	89.7%	0.943	-3.5% / -0.024
w/o Behavior Features	91.4%	0.952	-1.8% / -0.015
w/o Handcrafted Features	91.8%	0.958	-1.4% / -0.009
w/o Multi-Task Learning	92.1%	0.961	-1.1% / -0.006

Model Variant	F1	AUC	Δ Performance
w/o Cross-Modal Attention	91.6%	0.954	-1.6% / -0.013

The ablation study reveals that all components contribute to overall performance, with visual features providing the largest individual contribution (-3.5% F1 when removed). This confirms that visual content analysis is crucial for IH-LG detection, as images often contain information that either complements or contradicts textual claims. Behavior features contribute significantly (-1.8% F1), validating our hypothesis that user interaction patterns provide discriminative signals. The cross-modal attention mechanism and multi-task learning provide modest but consistent improvements, demonstrating the value of our integrated architecture.

4.4 IDQI Validation and Correlation Analysis

We validate the Information Deficiency Quantification Index through correlation analysis with user engagement metrics. Figure 1 illustrates the relationship between IDQI scores and private message trigger rates across the test set. The analysis reveals a strong positive correlation (Pearson $r = 0.78$, $p < 0.001$; Spearman $\rho = 0.82$, $p < 0.001$), confirming that higher information deficiency is associated with increased private message seeking behavior.

We further analyze the correlation between IDQI and other engagement metrics. Interestingly, IDQI shows a positive correlation with comment rate ($r = 0.34$, $p < 0.001$) but a negative correlation with save rate ($r = -0.21$, $p < 0.001$). This pattern suggests that users engage with IH-LG posts through comments seeking missing information, but are less likely to save such posts for future reference—consistent with the interpretation that these posts provide incomplete utility.

The IDQI prediction task achieves MAE = 0.087 and RMSE = 0.112 on the test set, with Pearson correlation of $r = 0.91$ between predicted and ground-truth IDQI scores. This demonstrates that the model effectively learns to quantify information deficiency, enabling fine-grained analysis beyond binary classification.

4.5 Category-wise Performance Analysis

Table III presents detection performance across different content categories. Performance varies significantly across categories, with highest accuracy in real estate (96.2%) and lowest in software tutorials (91.8%). This variation reflects differences in IH-LG behavior patterns across categories: real estate posts typically exhibit clear directive phrases and systematic information withholding, while software tutorials may present more ambiguous cases where information omission is less clearly intentional.

Table 3 Category-wise Detection Performance

Category	Acc.	F1	AUC	IH-LG Rate
Real Estate	96.2%	95.1%	0.978	41.2%
Home Decoration	95.4%	94.2%	0.974	38.7%

Category	Acc.	F1	AUC	IH-LG Rate
Education	94.1%	92.8%	0.968	35.6%
Lifestyle	93.8%	92.1%	0.962	28.3%
Software Tutorials	91.8%	89.7%	0.951	32.1%

4.6 Efficiency and Deployment Analysis

For practical deployment, we evaluate model efficiency in terms of inference latency and resource consumption. On a single A100 GPU, IH-LGDetector processes posts at an average rate of 67 posts per second with batch size 64, corresponding to approximately 15ms per post. This latency is well within acceptable bounds for real-time content moderation systems, which typically operate on timescales of hundreds of milliseconds.

Model compression through knowledge distillation yields a student model with 4x fewer parameters while maintaining 92.8% of original performance (F1 = 91.4%). Quantization to INT8 precision further reduces memory footprint by 75% with negligible performance degradation (F1 = 92.9%). These optimizations enable deployment on resource-constrained edge devices for on-device content analysis.

5. Discussion

5.1. Model Interpretability

To understand model decision-making, we employ SHAP (SHapley Additive exPlanations) analysis on the handcrafted feature components. The analysis reveals that directive phrase presence, price information absence, and user historical IH-LG rate are the top three contributing features, together accounting for 47% of prediction variance. This aligns with intuitive understanding of IH-LG behaviors and provides actionable insights for platform governance.

For neural components, we visualize attention weights from the cross-modal attention module. Case studies reveal that the model learns to attend to image regions containing text (as identified by OCR) when assessing information completeness, and to user profile elements when evaluating behavioral patterns. These attention patterns provide transparency into the model's reasoning process and support trust in deployment contexts.

5.2. Adversarial Robustness

Content creators may adapt their strategies to evade detection. We identify several potential adversarial approaches: (1) paraphrasing directive phrases to avoid keyword detection; (2) embedding information in images in ways that evade OCR; (3) gradually building reputation before engaging in IH-LG behaviors; and (4) using indirect language that implies rather than explicitly requests private communication.

To assess robustness, we conduct adversarial testing with paraphrased directive phrases generated by GPT-4. The model maintains 87.3% accuracy on paraphrased examples, demonstrating reasonable robustness to linguistic variation. However, performance degrades more significantly (72.1% accuracy) on examples using indirect

implication strategies, highlighting an area for future improvement. We recommend periodic model retraining with adversarial examples to maintain detection effectiveness as evasion tactics evolve.

5.3. Algorithmic Fairness

Detection systems must balance effectiveness with fairness to avoid disproportionate impact on legitimate content creators. We analyze false positive rates across user subgroups defined by account age, follower count, and posting frequency. The analysis reveals slightly elevated false positive rates for new accounts (8.2% vs. 5.1% overall), reflecting the model's reliance on behavioral history features.

To mitigate this bias, we implement a calibrated threshold strategy that applies lower detection thresholds for accounts with established positive histories. Additionally, we incorporate a human review step for borderline cases involving new accounts, ensuring that legitimate new content creators are not unfairly penalized. These fairness considerations are essential for maintaining platform creator ecosystem health.

5.4. Deployment Feasibility

The proposed framework is designed with deployment feasibility in mind. The modular architecture enables incremental deployment, starting with text-only detection and progressively adding visual and behavioral components. The model can be integrated into existing content moderation pipelines through a REST API, with asynchronous processing for non-time-critical analysis.

For platforms with existing moderation infrastructure, we recommend a hybrid approach: automated detection for high-confidence cases, combined with human review for borderline decisions. This approach maximizes detection coverage while maintaining quality control. The IDQI score provides a natural threshold for routing decisions, with high-IDQI posts flagged for immediate review and moderate-IDQI posts queued for batch processing.

5.5. Limitations

This work has several limitations that warrant acknowledgment. First, the dataset is collected from a single platform (Xiaohongshu), and generalization to other platforms with different content norms and user behaviors requires further validation. Second, the annotation process, while rigorous, may reflect annotator biases in identifying IH-LG behaviors, particularly for borderline cases. Third, the model's reliance on behavioral history features may disadvantage new users, as discussed in the fairness analysis.

Additionally, the current framework does not address the intent behind information hiding—some cases may reflect genuine uncertainty or inability to provide complete information rather than strategic lead generation. Distinguishing these cases remains an open challenge. Finally, the cross-platform transferability of detection models and the evolution of adversarial strategies over time represent ongoing challenges that require continued research attention.

6. Conclusion and Future Work

This paper addresses the critical but understudied problem of detecting

information-hiding lead generation behaviors in social Q&A platforms. We propose IH-LGDetector, a multi-modal deep learning framework that integrates textual, visual, and behavioral signals to identify posts that strategically withhold information to drive private communication. Our approach introduces the Information Deficiency Quantification Index (IDQI), enabling fine-grained analysis of information hiding behaviors and their correlation with user engagement patterns.

Extensive experiments on a dataset of 127,843 posts demonstrate that IH-LGDetector achieves 94.7% accuracy and 0.967 AUC-ROC, significantly outperforming baseline methods. Correlation analysis validates the IDQI metric, revealing a strong positive relationship ($r=0.78$) between information deficiency and private message trigger rates. The framework demonstrates practical deployability with inference latency under 15ms per post.

Future research directions include: (1) developing adversarial detection methods to address evolving evasion strategies; (2) investigating cross-platform transfer learning to enable model deployment on diverse platforms; (3) building real-time detection systems with streaming data processing architectures; and (4) incorporating user feedback loops to continuously improve detection accuracy. Additionally, exploring the intent behind information hiding—distinguishing strategic lead generation from genuine information limitations—represents an important direction for more nuanced content governance.

References

- [1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. WSDM*, 2008, pp. 183–194.
- [2] T. Chen, Y. Zhang, and M. Zhang, "Hierarchical neural networks for answer quality prediction in community question answering," in *Proc. ACL*, 2022, pp. 4521–4532.
- [3] Y. Liu, W. Zhang, and X. Li, "Large-scale analysis of content characteristics on Xiaohongshu platform," *Information Processing & Management*, vol. 59, no. 3, pp. 102–115, 2022.
- [4] H. Wang, J. Liu, and S. Chen, "Content value assessment framework for social media platforms: A case study of Xiaohongshu," *IEEE Access*, vol. 10, pp. 45678–45692, 2022.
- [5] J. Zeng, M. Chen, and L. Wang, "Multi-task learning for content quality assessment in social platforms," in *Proc. KDD*, 2023, pp. 3345–3355.
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.
- [7] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. ICML*, 2021, pp. 5583–5594.
- [8] Y. Zhou, L. Zheng, and C. Wang, "Multi-modal fake news detection via cross-modal attention," in *Proc. ACL*, 2023, pp. 2341–2352.
- [9] X. Wang, Y. Chen, and Z. Liu, "Detecting clickbait posts through multi-modal consistency analysis," *IEEE Transactions on Multimedia*, vol. 25, pp. 1234–1246, 2023.

-
- [10] J. Huang, W. Li, and M. Zhang, "OCR-enhanced content analysis for Chinese social media platforms," in *Proc. EMNLP*, 2022, pp. 5678–5689.
- [11] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, 2009.
- [12] A. Westfeld and A. Pfitzmann, "Attacks on steganographic systems," in *Proc. Information Hiding Workshop*, 1999, pp. 61–76.
- [13] F. Wu, J. Huang, and X. Chen, "Adversarial content detection in e-commerce reviews," in *Proc. WWW*, 2022, pp. 2345–2356.
- [14] L. Chen, Y. Wang, and H. Zhang, "Strategic behaviors in social media marketing: Detection and analysis," *ACM Transactions on Information Systems*, vol. 41, no. 2, pp. 1–28, 2023.
- [15] A. Mathur, G. Acar, M. J. Friedman, et al., "Dark patterns at scale: Findings from a crawl of 11K shopping websites," *Proc. ACM on Human-Computer Interaction*, vol. 3, pp. 1–32, 2019.
- [16] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *Proc. ICLR*, 2016.
- [17] W. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proc. ICDM*, 2018, pp. 197–206.
- [18] D. Wang, J. Lin, and P. Li, "Graph neural networks for fake account detection," in *Proc. KDD*, 2022, pp. 1987–1997.
- [19] Y. Liu, X. Zhang, and W. Chen, "Self-supervised learning for anomalous posting behavior detection," in *Proc. NAACL*, 2023, pp. 1234–1245.
- [20] H. Zhang, Y. Li, and J. Wang, "Information completeness and user engagement: A large-scale analysis," *Information Sciences*, vol. 612, pp. 456–471, 2022.
- [21] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *Proc. ICDM*, 2008, pp. 413–422.
- [22] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Proc. NIPS*, 2000, pp. 582–588.
- [23] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proc. MLSDA Workshop*, 2014, pp. 4–11.
- [24] L. Zong, B. Chen, and Q. Zhang, "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," in *Proc. ICLR*, 2018.
- [25] P. Malhotra, A. Ramaswamy, V. Sujatha, et al., "LSTM-based encoder-decoder for multi-sensor anomaly detection," in *Proc. ICML Anomaly Detection Workshop*, 2016.
- [26] Z. Liu, Y. Wang, and X. Chen, "Graph convolutional networks for fraudulent user group detection," in *Proc. WWW*, 2023, pp. 1567–1578.
- [27] S. T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019.
- [28] A. Beutel, J. Chen, T. Doshi, et al., "Fairness in recommendation ranking through pairwise comparisons," in *Proc. KDD*, 2019, pp. 2212–2220.

-
- [29] J. Wei, Y. Liu, and X. He, "Quality-aware recommendation: Combating low-quality content through algorithmic intervention," in *Proc. RecSys*, 2023, pp. 234–245.
- [30] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.
- [31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
- [32] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016, pp. 785–794.
- [34] G. Ke, Q. Meng, T. Finley, et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. NIPS*, 2017, pp. 3146–3154.
- [35] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.
- [36] Z. Yang, D. Yang, C. Dyer, et al., "Hierarchical attention networks for document classification," in *Proc. NAACL*, 2016, pp. 1480–1489.
- [37] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. NIPS*, 2017, pp. 4765–4774.
- [38] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. KDD*, 2016, pp. 1135–1144.
- [39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017.
- [40] P. Veličković, G. Cucurull, A. Casanova, et al., "Graph attention networks," in *Proc. ICLR*, 2018.
- [41] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.
- [42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.
- [43] H. Touvron, M. Cord, M. Douze, et al., "Training data-efficient image transformers & distillation through attention," in *Proc. ICML*, 2021, pp. 10347–10357.
- [44] J. Li, R. Socher, and S. C. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," in *Proc. ICLR*, 2020.
- [45] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learning Workshop*, 2015.