Vol. 8, No. 2, April 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Detecting Deceptive Online Reviews Using a Semantic Reliability Index and Hybrid Text Representation

Hartatik[1], Andri Syafrianto[2]

[1]Department of Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta, Sleman 55283, Yogyakarta, Indonesia

[2]Infotmatics Department, STMIK EL RAHMA Yogyakarta, Yogyakarta, Indonesia

Received:
September 14, 2025
Revised:
March 16, 2026
Accepted:
March 31 2026
Published:
April 12, 2026

Corresponding Author:
Author Name*:
Hartatik
Email*:
hartatik@amikom.ac.id

DOI:
10.63158/journalisi.v8i2.1576

Abstract. Online review platforms such as Yelp play an important role in consumer decision-making, but the growing prevalence of fake reviews undermines their reliability. This study proposes a hybrid approach for fake review detection by integrating stylometric features, language model signals, and semantic embeddings within a unified classification framework. The proposed method combines linguistic indicators, including GPT-2 perplexity, lexical diversity, sentence burstiness, punctuation ratio, and sentiment intensity, with TF-IDF representations and Sentence-BERT embeddings. A composite feature, namely the Semantic Reliability Index (SRI), is introduced to capture interactions between semantic similarity and linguistic characteristics, serving as an auxiliary feature within the hybrid model rather than a standalone classifier. Experiments on a Yelp hotel review dataset demonstrate that the hybrid model outperforms baseline methods in terms of F1-score and AUC, indicating improved discriminative capability. It should be noted that the classification setting is based on a binary transformation of ordinal labels, which may simplify the underlying label structure and influence performance interpretation. Overall, this work's contribution lies in a systematic feature-integration strategy that enhances fake review detection in the evaluated dataset.

Keywords: opinion spam detection, semantic similarity, stylometric features, XGBoost, hybrid feature representation

1. INTRODUCTION

User opinions have become a primary source of information in consumer decision-making, driven by the growing use of online review platforms such as Yelp. Reviews on these platforms strongly influence perceptions of product quality, business reputation, and customer trust in services. Previous studies show that online reviews significantly affect consumer purchasing behavior and the evaluation of business reputation in the digital ecosystem [1], [2], [3]. However, the growing reliance on online reviews has also introduced the serious problem of fake reviews, which are intentionally generated to manipulate public perception of products or services. Fake reviews can distort information, reduce the effectiveness of recommendation systems, and weaken user trust in online review platforms [2], [4], [5]. Consequently, developing automated methods for fake review detection has become an important research topic in natural language processing (NLP), text mining, and opinion mining [1], [6], [7], [8].

Early approaches to fake review detection relied primarily on machine learning techniques based on lexical features, such as term frequency–inverse document frequency (TF-IDF), combined with traditional classifiers, including Support Vector Machine, Naïve Bayes, and Logistic Regression. TF-IDF enables models to identify word distribution patterns that frequently appear in manipulated reviews and has been widely used in text spam detection studies [9], [10], [11]. However, lexical-based approaches are limited because they depend heavily on vocabulary distribution within a dataset. Fake review authors can easily modify word choices or use synonyms, making lexical patterns more difficult to detect. Furthermore, advances in generative technologies have enabled the production of increasingly natural-looking texts, reducing the effectiveness of frequency-based detection methods [12], [13].

To overcome these limitations, researchers have explored stylometric analysis, which examines linguistic characteristics such as lexical complexity, variation in sentence structure, sentence length distribution, and patterns of punctuation use. Prior studies show that fake reviews often exhibit lower vocabulary diversity, simpler sentence structures, and more extreme sentiment expressions compared with genuine reviews [10], [13], [14]. Stylometric indicators such as lexical diversity, sentence burstiness, punctuation ratio, and sentiment intensity therefore, provide useful signals for distinguishing between

deceptive and genuine reviews [4], [14], [15]. However, stylometric approaches alone still struggle to capture deeper semantic relationships among reviews.

Recent advances in transformer-based representation models have created new opportunities for semantic text analysis. Models such as BERT and Sentence-BERT generate high-dimensional vector representations that capture semantic meaning more effectively than traditional frequency-based methods [16], [17]. These representations enable the measurement of semantic similarity between reviews, even when different vocabularies are used. Research also shows that fake reviews frequently appear in clusters with high semantic similarity because they are generated using similar templates or promotional strategies [9], [18]. Therefore, embedding-based semantic similarity analysis can help identify semantic redundancy patterns associated with manipulated reviews.

In addition to semantic representations, recent studies have also employed language model signals and deep learning architectures to detect linguistic anomalies in deceptive texts [7], [12], [19]. Generative language models can compute perplexity, which measures the probability of a text according to the language distribution learned by the model. Unusual perplexity values may indicate unnatural linguistic structures that deviate from typical language patterns [12], [13]. Integrating language model signals with stylometric and semantic features has therefore been shown to improve performance in fake review detection.

Motivated by these observations, this study proposes a hybrid stylometric–semantic feature framework for detecting fake reviews on online review platforms. The proposed approach integrates several linguistic indicators of textual manipulation, including perplexity, lexical diversity, sentence burstiness, punctuation ratio, and sentiment intensity. In addition, semantic representations are generated using Sentence-BERT embeddings to capture semantic relationships between reviews. To model semantic manipulation patterns, semantic centroids representing prototypes of fake and genuine reviews are constructed from training data embeddings, and the similarity between each review and these centroids is used as an additional detection signal.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Despite these advances, existing approaches still face several limitations. Lexical-based methods are sensitive to vocabulary variations and fail to capture deeper semantic relationships, while stylometric approaches provide useful linguistic signals but often lack sufficient discriminative power when used independently. Embedding-based methods capture semantic similarity but do not explicitly model linguistic irregularities, and language model signals, such as perplexity, are rarely systematically integrated with stylometric and semantic features within a unified framework. More importantly, prior studies tend to evaluate these feature groups either in isolation or through loosely coupled combinations, resulting in a limited understanding of how different representation types interact within a single learning framework. This suggests that the core challenge lies not only in identifying informative features but in effectively integrating heterogeneous representations under a unified modeling strategy. Furthermore, this study explicitly focuses on this integration problem within a controlled experimental setting using a transformed binary Yelp review dataset, enabling a more precise evaluation of representation complementarity rather than proposing a broad, domain-independent detection solution.

To address this gap, this study introduces a hybrid feature framework that integrates stylometric features, language model signals, and semantic embeddings. Within this framework, a composite feature, the Semantic Reliability Index (SRI), is proposed to model interactions between semantic similarity and the linguistic characteristics of review texts. Unlike approaches that treat these components independently, the proposed framework emphasizes systematic integration of features within a unified representation space. It is important to note that SRI is not designed as a standalone classifier, but as an auxiliary feature within the hybrid model. Therefore, this work's contribution is positioned as a representation integration strategy rather than a fundamentally new detection paradigm. This study provides three main contributions. First, it introduces the Semantic Reliability Index (SRI) as a composite feature that captures interactions between semantic similarity and stylometric characteristics, rather than functioning as an independent detection mechanism. Second, it proposes a hybrid stylometric–semantic feature-extraction framework that systematically integrates lexical (TF-IDF), stylometric, and semantic representations within a unified classification framework. Third, it demonstrates, through experiments on a transformed binary Yelp review dataset, that such a hybrid representation strategy improves detection performance while providing empirical

insights into the complementary roles of different feature types in distinguishing deceptive and genuine reviews.

2. METHODS

This study develops a fake review detection approach that integrates linguistic stylometric features, signals from generative language models, and embedding-based semantic representations. To ensure methodological rigor, each component of the proposed approach is grounded in recent advances in fake review detection, stylometric analysis, semantic embedding, and hybrid feature learning [1], [17], [20], [21]. The overall research framework of the proposed method is illustrated in Figure 1, which summarizes the main stages of the detection pipeline, including data preprocessing, feature extraction, hybrid feature construction, and classification. For clarity and reproducibility, the proposed framework is implemented as a deterministic sequence of processing stages, where each stage produces a well-defined intermediate representation used by the next stage.

The overall workflow of the proposed method is as follows. First, raw review texts are preprocessed to remove noise and standardize the textual format. Second, multiple feature types are extracted, including stylometric features, language model signals, and semantic embeddings. Third, these features are combined into a unified hybrid representation. Finally, the resulting feature vectors are used as input to a supervised classification model for fake review detection. For reproducibility, the pipeline is formally defined as a sequence of stages: (1) preprocessing, (2) stylometric feature extraction, (3) semantic embedding generation, (4) centroid-based similarity computation, (5) SRI construction, (6) TF-IDF extraction, (7) feature fusion, and (8) classification. All stages are applied to all samples in the same order, and no manual intervention is introduced after preprocessing.

All experiments were conducted using a previously cleaned hotel review dataset from the Yelp platform. The dataset consists of two main files: a review text file and a metadata label file. The first file contains one review per line, and the second file contains numerical labels for the review categories. Since the original dataset contains multiple label categories, these labels are transformed into a binary classification problem

following common practices in fake review detection studies. Recent studies have shown that multi-level annotation schemes in review datasets often reflect varying degrees of credibility or deception, with higher label values indicating stronger indicators of manipulated content [20], [22]. Based on this convention, the highest label value is treated as representing fake reviews, while the remaining labels are considered genuine reviews. This transformation assumes an ordinal reliability structure in the dataset, consistent with recent deception-detection studies. However, we acknowledge that this binarization simplifies the original label distribution and may introduce bias; therefore, its potential impact is evaluated and discussed in the experimental results. More specifically, the transformation is motivated by the ordinal interpretation of the original labels, in which the highest label represents the strongest available indication of deceptive content in the dataset [20], [22]. Under this formulation, the binary task distinguishes the most deceptive reviews from the remaining reviews, thereby creating a clearer supervised decision boundary for evaluating the proposed hybrid representation. This transformation also improves class separability by focusing on the most reliable deceptive instances while reducing ambiguity from intermediate labels, thereby improving supervised learning stability. This design operationalizes the dataset as a binary classification problem by distinguishing the most deceptive instances from the remaining reviews, in accordance with the ordinal structure of the original annotations. Formally, if the original label is denoted by (y_i^{raw}), then the transformed binary label is defined as follows:

$$y_i = \begin{cases} 1, & if\ y_i^{raw} = max(y^{raw}) \\ 0, & otherwise \end{cases} \quad (1)$$

Thus, each data sample can be represented as a pair

$$(x_i, y_i), \text{ with } y_i \in \{0,1\}, \quad (2)$$

where $(y_i = 1)$ denotes a fake review and $(y_i = 0)$ denotes a genuine review. Equations (1) and (2) define the supervised learning target used throughout all experiments, and the same transformed labels are used consistently in the TF-IDF, SRI, and Hybrid model configurations.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

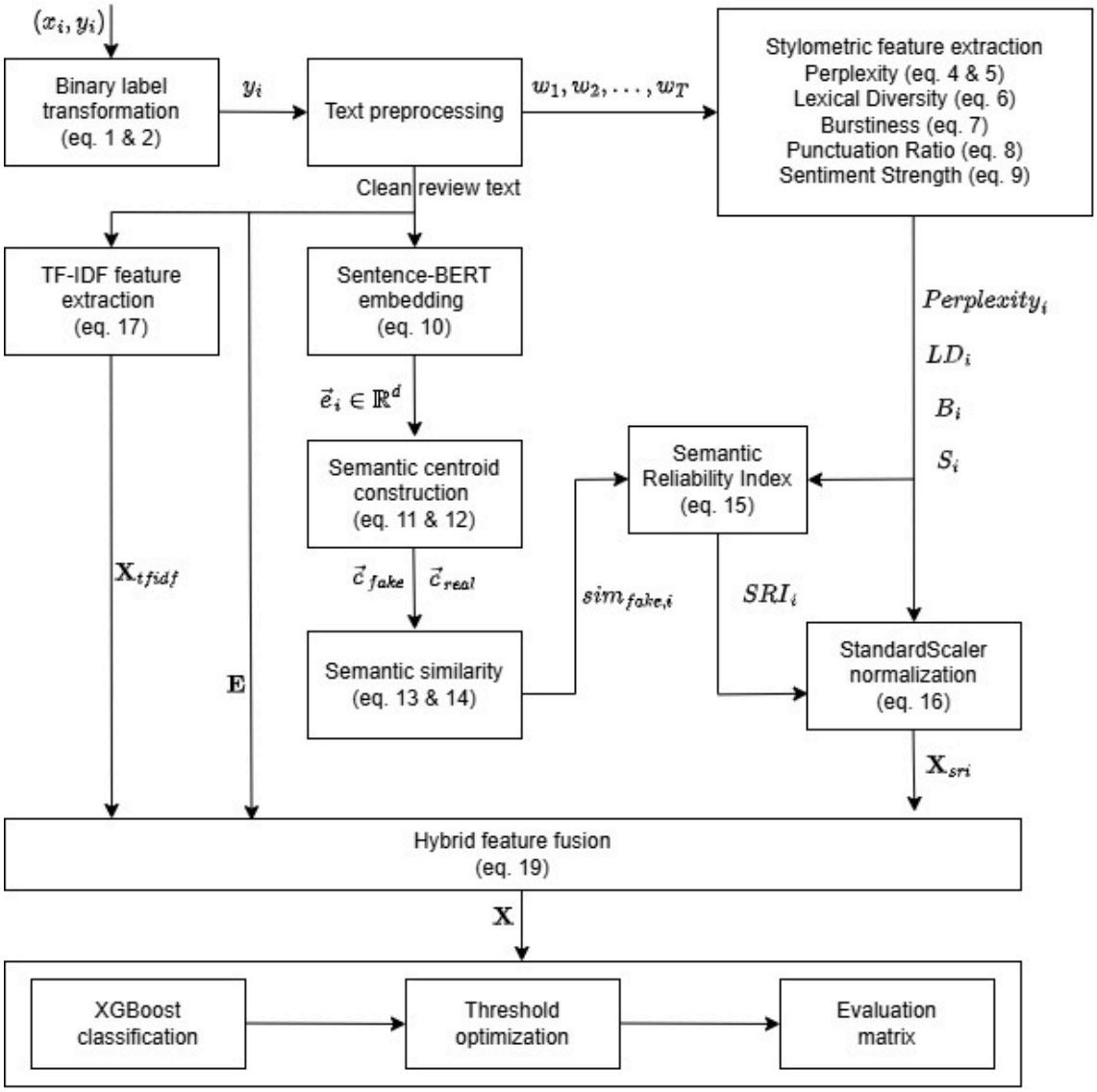

Figure. 1 Proposed framework for hybrid stylometric–semantic fake review detection integrating GPT-2 perplexity, Sentence-BERT embeddings, TF-IDF lexical features, and the Semantic Reliability Index (SRI).

Before feature extraction, all review texts undergo preprocessing. This stage includes converting all characters to lowercase, removing non-alphanumeric characters using regular expressions, and normalizing excessive whitespace. The purpose of this step is to reduce noise in the dataset, thereby making the text's linguistic structure more consistent. After preprocessing, each review is represented as a sequence of tokens.

$$w_1, w_2, \ldots, w_T \,, \tag{3}$$

where (T) denotes the number of tokens in the review. In the implementation, tokenization is performed after text normalization, using whitespace-based tokenization. No stemming or lemmatization is applied, so that lexical variation remains available to both the stylometric features and the TF-IDF representation.

This study extracts several stylometric features that capture the linguistic characteristics of review texts. The first feature is perplexity, computed using the GPT-2 language model. Perplexity measures the unpredictability of a text with respect to the language distribution learned by the model. If the probability of token (w_t) at position (t) is denoted as $(P(w_t \mid w_1,\ldots,w_{t-1}))$, then the cross-entropy is computed as

$$H = -\frac{1}{T}\sum_{t=1}^{T}\log P(w_t \mid w_1,\ldots,w_{t-1}), \tag{4}$$

where (T) is the number of tokens in the review. The perplexity value is then calculated as

$$Perplexity = e^{H}\ . \tag{5}$$

A higher perplexity value indicates that the linguistic structure of the text is more unusual relative to the language model.

The second stylometric feature is lexical diversity, which measures vocabulary richness in a review. If (N) denotes the total number of tokens and (V) denotes the number of unique words, and lexical diversity is computed as

$$LD = \frac{|V|}{N}, \tag{6}$$

where $(|V|)$ is the number of unique words and (N) is the total number of tokens. This feature is computed independently for each review and represented as a scalar value.
Another feature used in this study is burstiness, which measures the variability of sentence lengths within a review. Let the length of each sentence be represented as $(l_1,l_2,\ldots,l_m)$, where (m) denotes the number of sentences. Burstiness is calculated as

$$B = \frac{\sigma(l)}{\mu(l)}, \tag{7}$$

where $(\sigma(l))$ is the standard deviation of sentence lengths and $(\mu(l))$ is the mean sentence length. Higher burstiness values indicate greater variation in sentence length. Sentence segmentation is performed directly on the normalized review text before feature aggregation.

This study also computes the punctuation ratio, defined as the number of punctuation marks per token. If (N_{punct}) represents the number of punctuation marks and (N_{token}) represents the number of tokens, the ratio is calculated as

$$R_{punct} = \frac{N_{punct}}{N_{token}}. \tag{8}$$

In addition, a sentiment strength feature is computed using the VADER lexicon-based sentiment analysis method [23]. The sentiment value used is the absolute value of the compound sentiment score:

$$S = | s_{compound} |, \tag{9}$$

where $(s_{compound})$ is the combined sentiment score produced by VADER. All stylometric features in Eqs. (6)–(9) are therefore scalar descriptors extracted at the review level.

Besides stylometric features, this study also utilizes semantic representations using the Sentence-BERT model [16]. Recent transformer-based embedding models have demonstrated strong performance in capturing semantic similarity and contextual relationships in text [17], [22]. This model maps each review into a (d)-dimensional embedding space. The embedding representation of a review is expressed as a vector.

$$\vec{e}_i \in \mathrm{R}^d, \tag{10}$$

where (d) denotes the embedding dimension generated by Sentence-BERT. In the implementation, a fixed, pre-trained Sentence-BERT model is used to encode all reviews

into embeddings of the same dimension. The embedding model is used without task-specific fine-tuning so that the semantic representation remains consistent across all experiments.

To capture semantic patterns for each review class, semantic centroids are constructed for fake and genuine reviews. Suppose there are (N_f) fake reviews and (N_r) genuine reviews in the training data. The centroid vectors for each class are computed as

$$\vec{c}_{fake} = \frac{1}{N_f} \sum_{i:y_i=1} \vec{e}_i , \tag{11}$$

$$\vec{c}_{real} = \frac{1}{N_r} \sum_{i:y_i=0} \vec{e}_i . \tag{12}$$

The centroids are computed using only the training data to avoid information leakage from the test set.

The semantic similarity between a review and a centroid is calculated using the dot product operation. Embedding-based similarity approaches have been widely adopted in recent fake review detection studies to identify semantically redundant or templated deceptive content [20], [24]. This formulation preserves magnitude information in the embedding space, allowing stronger semantic signals to contribute proportionally to similarity estimation.

$$sim_{fake,i} = \vec{e}_i \cdot \vec{c}_{fake} \tag{13}$$

$$sim_{real,i} = \vec{e}_i \cdot \vec{c}_{real} \tag{14}$$

These similarity values indicate how closely a review resembles the semantic patterns of fake or genuine reviews. For consistency of notation, $sim_{fake,i}$ and $sim_{real,i}$ are scalar similarity scores associated with the i-th review.

Based on these features, this study proposes a composite feature, the Semantic Reliability Index (SRI). Unlike conventional single-representation approaches, recent

studies emphasize the importance of feature fusion and multi-view learning for improving robustness in fake review detection [24]. The SRI value is calculated as

$$SRI_i = \frac{sim_{fake,i}}{1+\log\left(1+Perplexity_i\right)} \times LD_i \times (1+B_i) \times (1+S_i) \tag{15}$$

In this equation, $(sim_{fake,i})$ represents the semantic similarity between the (i)-th review and the fake-review centroid, $(Perplexity_i)$ is the perplexity value of the review, (LD_i) is lexical diversity, (B_i) is burstiness, and (S_i) is sentiment intensity. The constant value (1) in several components is used to prevent division by zero and to stabilize feature scaling. Equation (15) is applied independently to each review after all constituent features have been extracted. This formulation encodes the interaction between semantic proximity to deceptive content and linguistic irregularity indicators into a single scalar feature.

All stylometric features and the SRI feature are then normalized using the StandardScaler method [25]. If a feature is represented as (x), the normalized value is computed as

$$z = \frac{x-\mu}{\sigma}, \tag{16}$$

where (μ) is the feature mean and (σ) is the feature standard deviation. The normalization parameters are estimated from the training data and then applied to the test data using the same transformation.

In addition to stylometric features, this study also employs lexical representations using TF-IDF [26]. If $(TF(t,d))$ denotes the frequency of term (t) in document (d), $(DF(t))$ denotes the number of documents containing term (t), and (N) is the total number of documents in the dataset, the TF-IDF value is computed as

$$TFIDF(t,d) = TF(t,d) \times \log \frac{N}{DF(t)} \tag{17}$$

The TF-IDF representation of all documents is expressed as a feature matrix

$$\mathbf{X}_{tfidf}\,, \tag{18}$$

where each row represents a document, and each column represents a term feature. The TF-IDF matrix is constructed after preprocessing, ensuring that the lexical representation aligns with the cleaned token sequence used in the other stages.

This study employs a hybrid feature representation by combining stylometric features, TF-IDF representations, and Sentence-BERT semantic embeddings. Recent research highlights that hybrid and multi-representation learning approaches significantly improve detection performance by capturing complementary information across feature domains [24]. If stylometric features are represented as matrix $\mathbf{X}_{sri}$, TF-IDF representations as $\mathbf{X}_{tfidf}$, and semantic embeddings as $\mathbf{E}$, the final feature representation is defined as

$$\mathbf{X} = \left[X_{sri}, X_{tfidf}, \mathbf{E} \right] \tag{19}$$

For clarity, X_{sri}, X_{tfidf}, and $\mathbf{E}$ are concatenated along the feature dimension so that each review is represented by one unified feature vector in the final matrix $\mathbf{X}$.

The dataset is split into training and test sets at an 80:20 ratio, with stratified sampling to preserve the class distribution. The same split protocol is used for all compared model configurations to ensure a fair comparison. A fixed random state is used in the implementation to exactly reproduce the data partition.

For classification, the Extreme Gradient Boosting (XGBoost) algorithm is employed [27]. Recent studies show that gradient boosting models remain highly effective for tabular and hybrid feature learning due to their ability to model non-linear feature interactions [21]. This model minimizes the objective function

$$L = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{20}$$

where $(l(y_i, \hat{y}_i))$ is the loss function between the true label (y_i) and predicted label $(\hat{y}_i)$, and $(\Omega(f_k))$ is the regularization term for the (k)-th decision tree. The regularization function is defined as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w \|^2 \tag{21}$$

where (T) is the number of leaves in the decision tree, (w) represents leaf weights, (γ) is the tree complexity penalty parameter, and (λ) is the weight regularization parameter. In the implementation, the classifier is trained on the final feature representation generated in Eq. (19), and the same learning setup is used across repeated experiments. To address class imbalance, the parameter

$$\text{scale\_pos\_weight} = \frac{N_{\text{negative}}}{N_{\text{positive}}} \tag{22}$$

is used, where $(N_{negative})$ is the number of genuine reviews and $(N_{positif})$ is the number of fake reviews in the training data. This weighting strategy increases the influence of the minority class during optimization and is computed solely from the training split.

After training, predicted probabilities are converted into class labels using an optimal threshold that maximizes the F1-score. This strategy is consistent with recent studies on imbalanced classification, which show that threshold tuning is necessary to balance precision and recall [28]. The threshold (t) is searched within the interval $0.1 \le t \le 0.9$ with a step size of 0.01. The threshold search is conducted systematically over the specified grid so that the reported F1-score reflects an explicitly optimized decision rule rather than the default probability cutoff.

Model performance is evaluated using several classification metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) computed from the ROC curve. Accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

Precision is calculated as

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

Recall is calculated as

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

and the F1-score is calculated as

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

These evaluation metrics are consistently used across all three model configurations to enable direct comparison of lexical-only, SRI-only, and hybrid feature representations.
For comparative analysis, this study evaluates three model configurations: a TF-IDF-based model, an SRI-based model, and a hybrid model. Such comparative evaluation is commonly used in recent literature to assess the contribution of individual feature groups in hybrid frameworks [20], [24].

3. RESULTS AND DISCUSSION

3.1. Performance Evaluatiom

The experiments were conducted to evaluate the effectiveness of the proposed approach for detecting fake reviews on the Yelp hotel review dataset. The dataset used in this study comprises 5854 reviews, including 4516 genuine and 1338 fake reviews. Therefore, the dataset exhibits an imbalanced class distribution. The class imbalance ratio is 4516/1338 = 3.37, indicating that the number of genuine reviews is approximately 3.37 times that of fake reviews. To reduce bias toward the majority class, the XGBoost model uses the scale_pos_weight parameter with a value of 3.3766, giving the minority class (fake reviews) greater weight during training. This imbalance handling strategy is crucial because, without proper weighting, the model would tend to favor the majority class, leading to poor detection of fake reviews, which are the primary target of this study.

The performance of the models was evaluated using F1-score and Area Under the ROC Curve (AUC). The F1-score was selected because it balances precision and recall for imbalanced datasets, while AUC measures the model's ability to distinguish between the two classes across different classification thresholds. The combination of these two metrics provides a more comprehensive evaluation: the F1-score reflects classification performance at a specific threshold, while AUC captures the model's overall ranking performance across all thresholds.

The experimental results are presented in Table 1, which compares the performance of the three approaches used in this study.

Table 1. Performance comparison of fake review detection models

Methods	F1-score	AUC
TF-IDF	0.5655	0.8185
SRI	0.5035	0.7548
TF-IDF + SRI	0.5965	0.8396

Based on Table 1, the Hybrid approach achieved the best performance with an F1-score of 0.5965 and an AUC of 0.8396. Compared with the TF-IDF baseline, the Hybrid model improved the F1-score by 0.5965 − 0.5655 = 0.031, which corresponds to a relative improvement of approximately 0.031/0.5655 = 5.48%. In addition, the AUC value increased by 0.8396 − 0.8185 = 0.0211, indicating that the Hybrid model has a stronger discriminative capability in distinguishing fake reviews from genuine ones.

The TF-IDF baseline approach achieved an F1-score of 0.5655 and an AUC of 0.8185. These results indicate that lexical patterns remain an important indicator in detecting manipulated reviews. The TF-IDF representation can capture word distribution patterns that frequently appear in fake reviews, such as repeated promotional expressions and strong sentiment. However, TF-IDF relies solely on word frequency and does not consider semantic relationships among reviews or stylistic characteristics of the writing.

The approach based on the Semantic Reliability Index (SRI) achieved an F1-score of 0.5035 and an AUC of 0.7548, the lowest among the three evaluated methods. This result suggests that stylometric and semantic features alone are insufficient to replace traditional lexical representations in text classification tasks.

The discriminative capability of the three models can be further observed in Figure 2, which presents the ROC curves for each approach. Based on Figure 2, the ROC curve of the Hybrid model (green line) consistently lies above the ROC curves of the TF-IDF model (blue line) and the SRI model (orange line) across most ranges of the false positive rate (FPR). This indicates that the Hybrid model has better classification performance in

distinguishing between fake and genuine reviews. The performance difference is most evident in the low FPR range (approximately 0–0.3), where the Hybrid model achieves a higher true positive rate (TPR), indicating better detection of fake reviews while maintaining a relatively low false positive rate. Interestingly, the improvement in AUC (0.0211) is more pronounced than the improvement in F1-score (0.031), suggesting that the Hybrid model enhances overall class separability rather than significantly improving classification decisions at a fixed threshold.

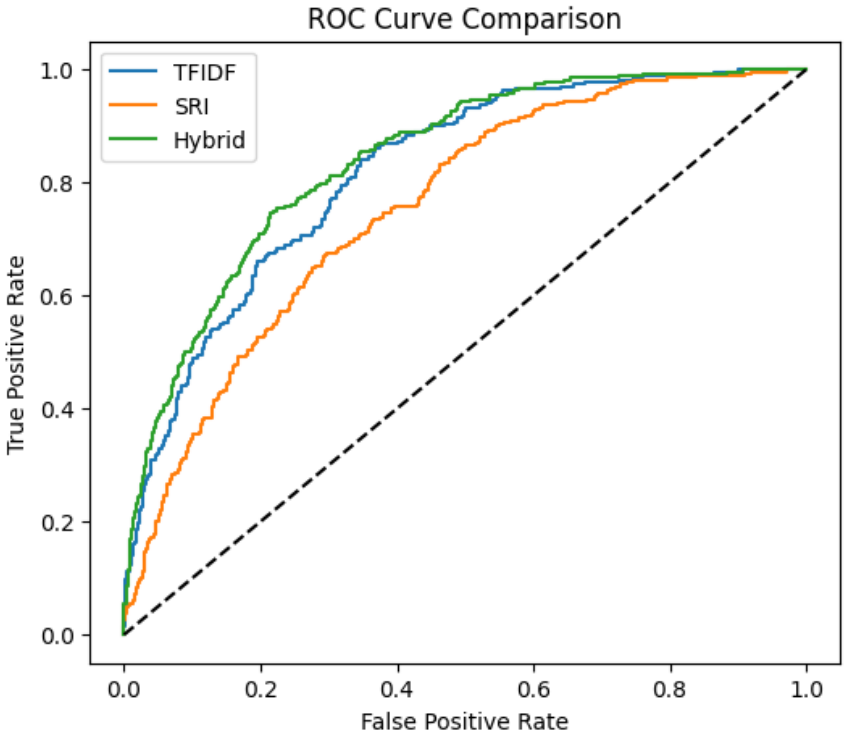

Figure 2. ROC curves for TF-IDF, SRI, and Hybrid models

The observed F1-score of 0.5965 indicates that while the proposed hybrid model improves detection performance, there is room for further improvement in classification accuracy. In practical deployment scenarios, a moderate F1-score suggests that the model is well-suited as a supportive component within a moderation pipeline, particularly in assisting human moderators rather than fully replacing them. In this context, false positives (genuine reviews incorrectly flagged as fake) may influence user trust and platform credibility, while false negatives (fake reviews incorrectly classified as genuine) may allow deceptive content to persist. Therefore, the balance between precision and recall becomes an important consideration in real-world applications, where the relative

impact of these errors depends on platform-specific moderation policies and risk tolerance.

Furthermore, the relatively modest improvement in F1-score compared to the more noticeable gain in AUC indicates that the proposed hybrid model primarily enhances global ranking quality rather than threshold-dependent classification decisions. This suggests that the model provides stronger probabilistic separation between classes, which can be further leveraged through threshold optimization or cost-sensitive strategies in deployment scenarios that require discrete decisions.

Another important consideration is cross-domain robustness. The current evaluation is conducted on a single Yelp hotel review dataset, where linguistic patterns, writing styles, and semantic structures are relatively consistent. In real-world settings, however, fake reviews may vary across domains such as e-commerce, restaurants, or social media platforms. Such variations may influence both stylometric and semantic signals, particularly for embedding-based similarity and SRI computation. As a result, evaluating the generalizability of the proposed hybrid representation under cross-domain distribution shifts represents an important direction for future work.

In addition, the binary label transformation used in this study has important implications for interpreting the reported performance metrics. By treating only the highest-label value as representing fake reviews, the model is trained to distinguish strongly deceptive instances from all other reviews. This formulation improves class separability and provides a clearer learning signal during training. At the same time, grouping intermediate-label reviews into a single class may introduce ambiguity, which can affect threshold-based metrics such as F1-score. This perspective also helps explain why the model exhibits greater improvements in AUC, which reflects ranking performance, compared to F1-score, which depends on a fixed decision threshold.

When compared with recent studies on fake review detection, particularly those using transformer-based architectures or hybrid feature integration, the proposed model's performance is consistent with the general observation that multi-representation approaches improve over single-feature baselines. In this study, the relatively strong performance of the TF-IDF baseline suggests that lexical signals already capture a

substantial portion of the dataset's discriminative information. The additional improvement achieved by the hybrid model, therefore, reflects the complementary contributions of stylometric and semantic features. In contrast, recent deep learning approaches that incorporate richer contextual modeling or additional behavioral signals often report larger performance gains, particularly when trained on more diverse datasets. This indicates that the proposed hybrid feature integration strategy provides a solid and effective foundation that can be further extended by incorporating additional modalities, such as user behavior or temporal interaction patterns.

3.2. Discussion

The experimental results provide important insights into how different feature representations contribute to fake review detection. The superior performance of the Hybrid model can be explained by the complementary nature of lexical, stylometric, and semantic features. TF-IDF captures surface-level lexical patterns that frequently appear in deceptive reviews, such as repetitive promotional language or exaggerated expressions. In contrast, Sentence-BERT embeddings model the semantic relationships between reviews, allowing the system to detect paraphrased or semantically similar deceptive content. Meanwhile, stylometric features capture writing style characteristics, including sentence variability, punctuation usage, and sentiment intensity. The integration of these heterogeneous representations enables the model to analyze deceptive reviews from multiple perspectives simultaneously, thereby improving discriminative performance.

In contrast, the SRI-only model's relatively lower performance indicates that high-level aggregated features may lose discriminative granularity when not supported by explicit lexical representations. Although stylometric and semantic signals capture important characteristics of deceptive writing, they are insufficient on their own to fully distinguish fake reviews, particularly in datasets where lexical cues remain strong indicators of manipulation.

Compared with prior studies on fake review detection, this study's findings are consistent with recent literature, showing that hybrid and multi-feature approaches generally outperform single-feature models. Previous work has shown that combining linguistic, semantic, and contextual features yields more robust detection performance, especially

in scenarios where deceptive reviews are intentionally diversified through paraphrasing or stylistic variation. The results of this study further support this perspective by showing that no single feature type is sufficient to capture the full complexity of deceptive behavior.

Despite these improvements, the relatively modest gain in F1-score compared to the more noticeable improvement in AUC suggests that the proposed model primarily enhances ranking quality rather than classification threshold decisions. This indicates that while the model is better at separating classes globally, further optimization may be required to improve threshold-based classification performance in practical deployment settings.

Moreover, the results highlight several limitations that have important implications for real-world fake review detection systems. The moderate performance levels indicate that distinguishing deceptive reviews remains challenging, particularly when fake reviews closely resemble genuine ones in both lexical and semantic aspects. This suggests that additional sources of information, such as user behavioral patterns, temporal dynamics, or metadata, may be necessary to further improve detection performance. Furthermore, the current evaluation is limited to a single dataset, and therefore, the generalizability of the proposed approach across different domains cannot yet be fully confirmed.

The findings demonstrate that integrating lexical, stylometric, and semantic representations provides a more comprehensive framework for modeling deceptive review patterns. However, further research is required to enhance robustness and generalization before the approach can be reliably applied in real-world online review platforms.

4. CONCLUSION

This study presents a hybrid feature integration framework for fake review detection that combines lexical, stylometric, and semantic representations within a unified classification framework. The results demonstrate that such integration provides complementary information, thereby improving the model's ability to distinguish between deceptive and genuine reviews in the evaluated dataset. However, the findings should be

interpreted in the context of several limitations. In particular, the binary-label transformation used in this study, which maps ordinal annotation levels to a two-class setting, may simplify the underlying label structure and affect the interpretation of classification performance. In addition, the evaluation is conducted on a single-domain dataset, which may limit the generalizability of the results across different application contexts. Future work may extend this approach by incorporating richer contextual and behavioral signals and by evaluating performance across more diverse datasets and real-world conditions.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Universitas AMIKOM Yogyakarta for providing financial support for the Article Processing Charge (APC) of this publication. The authors also appreciate the institutional support that facilitated the completion of this research.

REFERENCES

[1] Z. K. Nimra Mughal, Ghulam Mujtaba, Muhammad Hussain Mughal, Abdul Manaf, "Fake Reviews Detection on E-Commerce Websites Using Novel User Behavioral Features : An Experimental Study," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 24, no. 9, pp. 0–44, 2026, doi: 10.1145/3748493.

[2] A. Jakhar and A. Indian, "Explainable fake review detection : A hybrid deep learning model for E-commerce platforms to enhance customer trust," *J. Retail. Consum. Serv.*, vol. 92, no. March, pp. 1–15, 2026.

[3] P. Sun *et al.*, "Fake Review Detection Model Based on Comment Content and Review Behavior," *Electronics*, vol. 13, pp. 1–17, 2024.

[4] E. Elmurngi and A. Gherbi, "Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques," in *DATA ANALYTICS 2017 : The Sixth International Conference on Data Analytics*, IARIA, 2017, pp. 65–72.

[5] J. Wang and J. Chen, "WF-CFRB : A Deep Learning Approach for Fake Review Detection Based on Weighted Fusion of Contextual Features and Reviewer Behaviors," *J SYST SCI SYST ENG*, vol. 34, no. 5, pp. 558–575, 2025.

[6] M. J. Abd and M. H. Hussein, "Fake reviews detection in e-commerce using machine learning techniques : a comparative survey," in *BIO Web of Conferences 97*, ISCKU 2024, 2024, pp. 1–12. doi: 10.1051/bioconf/20249700099.

[7] R. Mohawesh, H. Bany, Y. Jararweh, and M. Alkhalaileh, "International Journal of Cognitive Computing in Engineering Fake review detection using transformer-based enhanced LSTM and RoBERTa," *Int. J. Cogn. Comput. Eng.*, vol. 5, no. June, pp. 250–258, 2024, doi: 10.1016/j.ijcce.2024.06.001.

[8] J. Kumar, "Fake Review Detection Using Behavioral and Contextual Features Fake Review Detection Using Behavioral and Contextual Features," QUAID-I-AZAM UNIVERSITY, 2018.

[9] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a General Rule for Identifying Deceptive Opinion Spam," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA: Association for Computational Linguistics (ACL), 2014, pp. 1566–1576.

[10] C. Chen, H. Zhao, and Y. Yang, "Deceptive Opinion Spam Detection Using Deep Level Linguistic Features," in *International Joint Conference on Natural Language Processing (IJCNLP)*, ACL Anthology (Association for Computational Linguistics), 2015, pp. 465–474. doi: 10.1007/978-3-319-25207-0.

[11] S. Morgan and B. Liu, "Spotting Fake Reviewer Groups in Consumer Reviews," in *the International World Wide Web Conference Committee (IW3C2)*, Lyon, France: ACM, 2026, pp. 191–200. doi: 10.1145/2187836.2187863.

[12] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, "Detecting Deceptive Reviews using Generative Adversarial Networks," in *2018 IEEE Symposium on Security and Privacy Workshops*, 2018, pp. 89–95. doi: 10.1109/SPW.2018.00022.

[13] G. Bathla, P. Singh, R. Kumar, Erik Cambria, and Rajeev Tiwari, "Intelligent fake reviews detection based on aspect extraction and analysis using deep learning," *Neural Comput. Appl.*, vol. 34, no. 22, pp. 20213–20229, 2022, doi: 10.1007/s00521-022-07531-8.

[14] Y. C. Song Feng, Ritwik Banerjee, "Syntactic Stylometry for Deception Detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (ACL), 2012, pp. 171–175.

[15] K. K. Poojary, "Deciphering Deception - Detecting Fake Review using NLP by analysis of stylistic, sentiment-based, and semantic features," Dublin Business School, 2024.

[16] N. Reimers and I. Gurevych, "Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China: Association for Computational Linguistics (ACL), 2019, pp. 3982–3992.

[17] T. Gao, X. Yao, and D. Chen, "SimCSE : Simple Contrastive Learning of Sentence Embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics (ACL), 2021, pp. 6894–6910.

[18] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," *WWW'12 - Proc. 21st Annu. Conf. World Wide Web*, pp. 191–200, 2012, doi: 10.1145/2187836.2187863.

[19] M. Ennaouri and A. Zellou, "Enhancing Fake Review Detection Using Linguistic Exaggeration, BERT Embeddings, and Fuzzy Logic," *IEEE Access*, vol. 13, no. August, pp. 135957–135968, 2025, doi: 10.1109/ACCESS.2025.3594629.

[20] Y. Guo, S. Ji, N. Cao, D. K. W. Chiu, N. Su, and C. Zhang, "MDG : Fusion learning of the maximal diffusion, deep propagation and global structure features of fake news," *Expert Syst. Appl.*, vol. 213, no. November 2022, pp. 1–15, 2023, doi: 10.1016/j.eswa.2022.119291.

[21] S. Sarafian and Y. Aperstein, "Improving Deep Tabular Learning," 2025.

[22] J. Chen, G. Zhou, M. Lan, S. Wang, S. Li, and J. Lu, "Semantic-aware fake news detection with heterogeneous graph attention," *J. Intell. Inf. Syst.*, vol. 63, pp. 1865–1890, 2025.

[23] C. J. Hutto and E. Gilbert, "VADER : A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, Association for the Advancement of Artificial Intelligence, 2014, pp. 216–225.

[24] J. Wang, H. Kan, F. Meng, Q. Mu, G. Shi, and X. Xiao, "Fake Review Detection Based on Multiple Feature Fusion and Rolling Collaborative Training," *IEEE Access*, vol. 8, pp. 182625–182639, 2020, doi: 10.1109/ACCESS.2020.3028588.

[25] J. F. Trevor Hastie, Robert Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. California: Springer, 2017.

[26] Hanafi and B. Mohd Aboobaider, "Word Sequential Using Deep LSTM and Matrix Factorization to Handle Rating Sparse Data for E-Commerce Recommender System," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, 2021, doi: 10.1155/2021/8751173.

[27] T. Chen and C. Guestrin, "XGBoost : A Scalable Tree Boosting System," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2026, pp. 785–794. doi: 10.1145/2939672.2939785.

[28] D. Zhang, W. Li, B. Niu, and C. Wu, "A deep learning approach for detecting fake reviewers : Exploiting reviewing behavior and textual information," *Decis. Support Syst.*, vol. 166, no. November 2022, p. 113911, 2023, doi: 10.1016/j.dss.2022.113911.