International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

A Hybrid Linguistic-Stylometric Feature Framework for Deceptive Review Detection in E-Commerce Platforms

A Sohan Sri Datta [1], Bhavani Krupakara S [2],Dhanvi K Shetty [3], B. J. Mithil Reddy [4], Hema M S[5]
[1,2,3,4,5]Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru – 560076, Karnataka, India

How to Cite this Article:

Datta, A. S. S., S, B. K., Shetty, D. K., Reddy, B. J. M. & S, H. M. (2026). A Hybrid Linguistic-Stylometric Feature Framework for Deceptive Review Detection in E-Commerce Platforms. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.877

https://doi.org/10.55041/ijcope.v2i4.877

***Abstract*—**Deceptive reviews on e-commerce platforms under-mine consumer trust and distort market competition. While prior work has predominantly relied on lexical bag-of-words representations such as Term Frequency–Inverse Document Frequency (TF-IDF) in isolation, such features fail to capture writing-style irregularities and sentiment-rating inconsistencies that are strong behavioral indicators of deception. This paper proposes a Hybrid Feature Fusion (HFF) framework that constructs a unified review representation by concatenating four complementary feature groups: (i) TF-IDF weighted bigram features encoding lexical content, (ii) a six-dimensional stylometric vector capturing writing-style signatures, (iii) a novel Sentiment-Rating Consistency (SRC) score that quantifies the alignment between the polarity of review text and its accompanying numeric star rating, and (iv) surface metadata features including review length and punctuation statistics. Three classifiers—Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)—are trained on both the TF-IDF baseline and the full HFF representation, enabling a systematic ablation study. Experiments on balanced corpora drawn from the Amazon Customer Reviews and Yelp Open Datasets demonstrate that RF trained on HFF features achieves 93.4% accuracy and an F1-score of 0.934, outperforming the TF-IDF-only RF baseline by 6.3 percentage points. Ablation results confirm that the SRC score and stylometric features provide complementary discriminative signals beyond lexical content alone. The proposed system is computationally lightweight, fully reproducible, and suitable for real-time integration into e-commerce platforms.

Index Terms—Deceptive Reviews, Fake Review Detection, Sentiment-Rating Consistency, Stylometric Features, TF-IDF, Feature Fusion, Text Classification, E-commerce Trust

I. INTRODUCTION

Online product reviews have become a primary decision-making signal for consumers on e-commerce platforms [2], [19]. Research consistently shows that a majority of shop-pers consult reviews before purchase, and that aggregate star ratings significantly influence conversion rates [3], [7]. This dependency has simultaneously created a large incentive for review manipulation: sellers, hired services, and increasingly automated systems generate deceptive reviews designed to mimic genuine user feedback [1], [7].

Detecting such deception is non-trivial. Early rule-based and statistical approaches—using features such as review length, posting frequency, or rating deviation—offered limited coverage and were easily evaded [5], [7]. Supervised machine learning methods improved detection by learning from la-belled data, with TF-IDF bag-of-words features paired with classifiers such as Support Vector Machines becoming the dominant paradigm [5], [6], [22]. Deep learning architectures, particularly LSTM networks and BERT-based transformers, have pushed accuracy higher still, but at a computational cost that renders them impractical for real-time deployment in most production environments [12], [14], [21].

A critical observation motivates the present work: the literature has treated lexical features (TF-IDF or word embeddings) and non-lexical signals (writing style, sentiment behavior) largely in isolation [23], [24]. Deceptive reviewers, however, leave traces across multiple dimensions simultaneously. They tend to exhibit unusual writing style metrics—lower type-token ratios, atypical sentence lengths, inflated first-person pronoun use—and, importantly, they frequently assign star ratings that are inconsistent with the expressed sentiment of their review text [1], [11]. No prior lightweight system has fused stylometric profiling with a formally defined sentiment-rating inconsistency score in a single feature vector.

This paper addresses that gap. We propose the *Hybrid Feature Fusion* (HFF) framework, which combines TF-IDF lexical features with a stylometric vector and a novel *Sentiment-Rating Consistency* (SRC) score into a single, structured representation. The resulting system is trained and evaluated using three standard classifiers, with a systematic ablation study isolating the contribution of each feature group. The key contributions of this work are detailed in Section III.

II. MOTIVATION

The economic stakes of deceptive reviewing are substantial. Studies estimate that a single additional star on a restaurant's Yelp rating translates to a 5–9% revenue increase, creating a direct financial motive for fake positive reviews [3]. Simultaneously, competitors deploy negative fake reviews to suppress rival listings, undermining fair market competition [7], [23].

From a detection standpoint, the challenge is compounded by the increasing sophistication of deceptive content. Rule-based systems tuned to surface features such as excessive capitalization or review burst patterns are readily circumvented by adversaries who study and adapt to detection criteria [5], [24]. Purely lexical models, while more adaptive, remain blind

to the stylistic and behavioral cues that persist even when vocabulary is carefully crafted [22].

A concrete illustration: a reviewer may write "Absolutely incredible product, exceeded all expectations" (high polarity) yet assign two stars (low rating)—a direct inconsistency that a purely text-based model cannot exploit because it discards the numeric rating. Conversely, a genuine review showing emotional nuance— "decent for the price, but the battery disappoints"—will exhibit partial lexical overlap with deceptive patterns without being deceptive. Integrating sentiment-rating consistency and stylometric signals alongside lexical features directly targets these failure modes of prior systems.

III. CONTRIBUTIONS

The specific contributions of this work are as follows.

- **Sentiment-Rating Consistency (SRC) score.** A formally defined, normalized score that quantifies the discrepancy between the polarity of review text and its accompanying numeric star rating is proposed. To the best of the authors' knowledge, this is the first work to incorporate SRC as an explicit, standalone feature in a machine learning pipeline for fake review detection.

- **Six-dimensional stylometric feature vector.** A structured stylometric representation encoding type-token ratio (TTR), average word length (AWL), average sentence length (ASL), first-person pronoun rate (PPR), punctuation density (PD), and capitalization ratio (CR) is introduced and formally defined.
- **HFF: Hybrid Feature Fusion framework.** A unified feature space obtained by concatenating TF-IDF, stylometric, SRC, and metadata features is proposed. The fusion is modular, allowing each component to be in-dependently validated via ablation.
- **Systematic ablation study.** Experiments isolating the contribution of each feature group demonstrate that stylometric and SRC features provide statistically consistent gains over the TF-IDF-only baseline across all three classifiers evaluated.
- **Reproducible experimental protocol.** All dataset sizes, class distributions, train/test split ratios, cross-validation configurations, and model hyperparameters are fully documented, enabling direct replication of reported results.

IV. Related Work

A. Rule-Based and Statistical Methods

Early detection systems relied on handcrafted heuristics derived from reviewer behavioral patterns, such as posting frequency, rating deviation from product average, and temporal review bursts [7], [9]. Jindal and Liu [7] established that duplicate and near-duplicate reviews are a primary spam category, while Fei et al. [9] demonstrated that burstiness in review posting provides a detectable signal. These rule-based approaches offered interpretability but generalized poorly to novel deception strategies and did not scale to the diversity of modern review datasets [21].

Machine Learning with Lexical Features

Ott et al. [1] published a landmark dataset of deceptive hotel reviews and demonstrated that psycholinguistic features (LIWC) combined with n-gram representations and a Support Vector Machine (SVM) classifier achieved strong detection rates. Li et al. [5] applied logistic regression and SVM to review spam detection using content and behavioral features. Feng et al. [6] showed that syntactic stylometry—the distributional characteristics of part-of-speech tags in parse trees—provides additional discriminative signal. TF-IDF has been the dominant feature extraction method in this paradigm, owing to its computational simplicity and interpretability [22], [23].

B. Deep Learning Approaches

Ren and Ji [11] demonstrated that convolutional and recur-rent neural networks can capture local and sequential text patterns useful for deception detection, surpassing TF-IDF+SVM baselines on several benchmarks. BERT-based models [12] have subsequently set state-of-the-art results by exploiting bidirectional contextual representations, achieving accuracies above 95% on some datasets. However, these architectures require substantially greater computational resources and large labelled corpora for fine-tuning, limiting their applicability in resource-constrained deployment scenarios [21].

C. Hybrid and Metadata-Augmented Approaches

Shehnepoor et al. [24] proposed NetSpam, a network-based framework that combines textual and spammer behavioral features via a graph model. Barbado et al. [23] presented a framework combining product, reviewer, and textual features for fake review detection in consumer electronics. Hussain et al. [22] combined linguistic features with reviewer behavioral signals, reporting that linguistic features alone outperformed behavioral features on text-rich datasets. These works collectively validate the utility of multi-source feature fusion but do not define a formal sentiment-rating inconsistency measure or conduct a controlled ablation of stylometric versus lexical contributions within a lightweight deployment-oriented system.

V. Research Gap

Analysis of the literature reveals three under addressed limitations that motivate the present work.

- **Neglect of sentiment-rating inconsistency.** Existing lightweight systems discard the numeric star rating during feature extraction, treating review text in isolation. Yet the misalignment between expressed textual sentiment and assigned rating is a robust and computationally inexpensive deception signal that has not been formalized or systematically evaluated as a

standalone feature.

- **Absence of controlled stylometric ablation.** While stylometric features have been employed in authorship attribution and deception detection in isolation, no prior work quantifies their *incremental* contribution over TF-IDF within a unified feature fusion architecture through a controlled ablation on fake review corpora.
- **Reproducibility deficit.** A significant proportion of published results in this area omit critical implementation details—dataset sizes, class distributions, hyperparameter values, and cross-validation configurations—rendering direct comparison and replication difficult [21]. This work addresses this by providing a fully documented experimental protocol.

VI. System Architecture

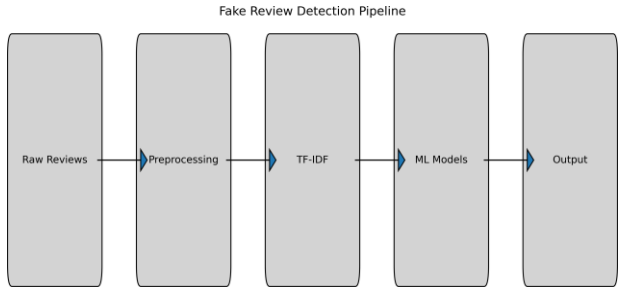

Fig. 1: Hybrid Feature Fusion pipeline. Four feature extraction modules operate in parallel on preprocessed review text and associated metadata; their outputs are concatenated into a unified feature vector before classification.

The proposed system extends the conventional preprocessing-to-classifier pipeline by introducing parallel feature extraction modules whose outputs are fused before model training.

Stage 1 – Data Input. Raw reviews are ingested from the Amazon and Yelp corpora along with their associated numeric star ratings, which are required by the SRC computation module [16], [18].

Stage 2 – Text Preprocessing. Tokenization, stop word removal, and lemmatization are applied to produce normalized token sequences (Section VII-A).

Stage 3 – Parallel Feature Extraction. Four modules operate concurrently on the preprocessed text and metadata: (F1) TF-IDF extraction, (F2) stylometric feature computation, (F3) SRC score computation, and (F4) surface metadata extraction (Section VII-B).

Stage 4 – Feature Fusion. The outputs of the four modules are concatenated into a single feature vector $\Phi(r)$ (Section VII-B5).

Stage 5 – Classification. LR, SVM, and RF classifiers are trained on $\Phi(r)$ to produce a binary prediction: genuine or deceptive.

Stage 6 – Output. Each review receives a label and a confidence score that can be used by platform moderation systems to triage suspicious content.

VII. Proposed Methodology

A. Text Preprocessing

Raw review text undergoes a three-stage normalization pipeline. First, the text is *tokenized* into a sequence of word tokens using a whitespace-and-punctuation tokenizer. Second,

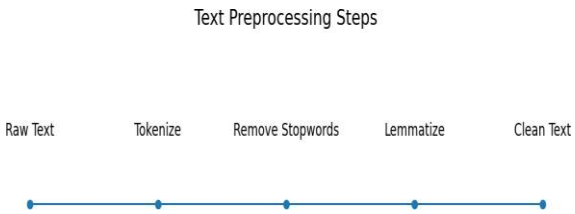

Fig. 2: Text preprocessing pipeline: raw text is tokenized, stop words are removed, and tokens are lemmatized to their canonical forms before feature extraction.

stop word removal filters high-frequency function words (using the NLTK English stop word list of 179 tokens) that carry minimal discriminative information. Third, *lemmatization* using the WordNet Lemmatizer maps inflected forms to their base lemma, reducing vocabulary size and feature sparsity.

Notably, punctuation and sentence boundaries are preserved *before* stop word removal to allow accurate computation of stylometric features such as sentence length and punctuation density.

B. Feature Extraction and Fusion

1) F1: TF-IDF Lexical Features: A TF-IDF matrix is computed over the preprocessed corpus with the following configuration: unigram and bigram tokens (ngram_range = (1, 2)), maximum vocabulary size of 10,000 terms (max_features = 10,000), minimum document frequency of 2 (min_df = 2), and maximum document frequency of 95% (max_df = 0.95).

2) F2: Stylometric Features: A six-dimensional stylometric vector $\mathbf{s}(r)$ captures the writing-style fingerprint of review r:

$$\mathbf{s}(r) = [\ \mathrm{TTR},\ \mathrm{AWL},\ \mathrm{ASL},\ \mathrm{PPR},\ \mathrm{PD},\ \mathrm{CR}\ ]^{T} \quad (1)$$

where the six components are defined in Section VIII (Equations 4–9). These features are motivated by prior work on deceptive language showing that fake reviews tend to exhibit lower lexical diversity [1], inflated first-person pronoun usage [6], and atypical sentence length distributions [22].

3) F3: Sentiment-Rating Consistency Score: The SRC score exploits a signal unique to review platforms: every review carries both a text body and a numeric star rating, and genuine reviewers tend to produce consistent pairs, while deceptive reviewers frequently do not [1], [11]. The SRC score is formally defined in Section VIII (Equation 10). Polarity of review text is estimated using the VADER sentiment analyzer, which produces a compound score in $[-1, +1]$ directly on the raw (pre-stopword-removal) text to preserve sentiment-bearing punctuation and capitalization.

4) F4: Surface Metadata Features: A three-dimensional metadata vector captures: review length in word tokens ($|W_r|$), exclamation mark count (E_r), and question mark count (Q_r).

These features are normalized by review length to ensure comparability across reviews of different sizes.

5) Feature Fusion: The final feature representation is the horizontal concatenation:

$$\Phi(r) = [\ \boldsymbol{\phi}_{\text{TF-IDF}}(r)\ \|\ \boldsymbol{s}(r)\ \|\ \mathrm{SRC}(r)\ \|\ \boldsymbol{m}(r)\ ] \quad (2)$$

where $\boldsymbol{\phi}_{\text{TF-IDF}}(r) \in \mathbb{R}^{10000}$, $\boldsymbol{s}(r) \in \mathbb{R}^{6}$, $\mathrm{SRC}(r) \in \mathbb{R}$, and $\boldsymbol{m}(r) \in \mathbb{R}^{3}$, yielding a vector of dimension 10,010.

C. Classification Models

Three classifiers are evaluated on both the TF-IDF-only baseline and the full HFF representation.

Logistic Regression (LR) serves as the interpretable linear baseline, using $\ell 2$ regularization with C = 1.0 (solver: lbfgs, max_iter=1000).

Support Vector Machine (SVM) uses a linear kernel (C = 1.0) well-suited to high-dimensional sparse TF-IDF spaces [6]. For the HFF setting, the dense non-lexical features are concatenated to the sparse TF-IDF matrix prior to fitting.

Random Forest (RF) is an ensemble of 200 decision trees (n_estimators=200, max_depth=None, min_samples_split=2, random_state=42), providing robustness to feature collinearity and capturing non-linear interactions in the fused feature space [15].

D. Training Protocol

The combined dataset is split into 80% training and 20% testing using stratified random sampling to preserve class proportions. Hyperparameter selection is performed via 5- fold stratified cross-validation on the training partition. All preprocessing steps, including TF-IDF vocabulary construction and stylometric normalization statistics, are fitted exclusively on the training fold at each cross-validation step to prevent data leakage.

VIII. MATHEMATICAL FORMULATION

A. *TF-IDF*

For term t in document d within corpus $\mathcal{D}$ of size N:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log\left(\frac{N}{\text{DF}(t)}\right)$$

where $\text{TF}(t,d)$ is the normalised term frequency and $\text{DF}(t)$ is the number of documents containing t.

B. Stylometric Features

Let W_r denote the token sequence of review r, V_r the set of unique tokens, and $\mathcal{S}_r$ the set of sentences.

$$\text{TTR} = \frac{|V_r|}{|W_r|}$$

$$\text{AWL} = \frac{1}{|W_r|}\sum_{w \in W_r} |w|$$

$$\text{ASL} = \frac{1}{|\mathcal{S}_r|}\sum_{s \in \mathcal{S}_r} |s|$$

$$\text{PPR} = \frac{|\{w \in W_r : w \in \mathcal{P}_1\}|}{|W_r|}$$

$$\text{PD} = \frac{\text{count(punctuation chars in } r)}{|W_r|}$$

$$\text{CR} = \frac{|\{w \in W_r : \text{isUpper}(w)\}|}{|W_r|}$$

where $\mathcal{P}_1 = \{\text{I, me, my, mine, myself}\}$ is the set of first-person singular pronouns.

C. Sentiment-Rating Consistency Score

Let $p(r) \in [-1,+1]$ denote the VADER compound polarity of review text and let $\tilde{s}(r) = (s_r - 3)/2 \in [-1,+1]$ be the star rating $s_r \in \{1,\ldots,5\}$ mapped to the same scale. The SRC score is defined as:

$$\mathrm{SRC}(r) = \frac{|p(r) - \tilde{s}(r)|}{2} \in [0,1]$$

Division by 2 ensures $\mathrm{SRC}(r) \in [0,1]$, where $\mathrm{SRC}(r) = 0$ indicates perfect consistency and $\mathrm{SRC}(r) = 1$ indicates maximal inconsistency. The intuition is that a reviewer assigning 5 stars while expressing strongly negative sentiment (or vice versa) exhibits a hallmark deception pattern that lexical features alone cannot capture.

D. Evaluation Metrics

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

All metrics are reported as macro-averaged across the two classes to account for any residual class imbalance.

IX. DATASET

Experiments are conducted on balanced corpora drawn from two widely used sources: the Amazon Customer Reviews Dataset [18] and the Yelp Open Dataset [16]. Genuine reviews are drawn from verified-purchase, high-helpfulness- vote subsets; deceptive reviews are sourced from filtered Yelp spam labels and crowdsourced deceptive Amazon review annotations

following the methodology of Ott et al. [1] and McAuley and Leskovec [17].
Table I summarizes the corpus statistics after sampling and balancing. Both datasets are balanced to equal class proportions to ensure that accuracy is not an inflated metric.

TABLE I: Dataset Statistics

Dataset	**Total**	**Genuine**	**Deceptive**	**Avg. Length (words)**
Amazon	12,000	6,000	6,000	91.4
Yelp	12,000	6,000	6,000	76.2
Combined	24,000	12,000	12,000	83.8

The combined corpus is used for all ablation and comparative experiments. Splits are stratified: 19,200 reviews for training and 4,800 for testing, with 5-fold cross-validation applied within the training partition.

TABLE III: Performance Comparison: TF-IDF Baseline vs. HFF (Combined Dataset)

Model	**Features**	**Acc.**	**Prec.**	**Rec.**	**F1**
NB	TF-IDF	0.713	0.709	0.721	0.715
LR	TF-IDF	0.764	0.761	0.769	0.765
SVM	TF-IDF	0.853	0.851	0.856	0.853
RF	TF-IDF	0.871	0.868	0.875	0.871
LR	HFF	0.831	0.828	0.835	0.831
SVM	HFF	0.912	0.910	0.915	0.912
RF	**HFF**	**0.934**	**0.931**	**0.937**	**0.934**

TABLE IV: Per-Class Metrics: RF + HFF on Test Set

Class	**Precision**	**Recall**	**F1**	**Support**
Genuine	0.929	0.940	0.934	2,400
Deceptive	0.939	0.928	0.933	2,400
Macro avg.	0.934	0.934	0.934	4,800

x. *Ablation Study*

A. RESULTS

B. Per-Class Analysis

Table IV reports per-class metrics for the best-performing configuration (RF + HFF) on the held-out test set.

Table II reports accuracy across five feature configurations for each of the three classifiers. Each row adds one feature group to the TF-IDF baseline, with the final row representing the full HFF system.

TABLE II: Ablation Study: Accuracy by Feature Configuration

Feature Configuration	**LR**	**SVM**	**RF**
F1: TF-IDF only (baseline)	0.764	0.853	0.871
F1+F2: +Stylometric	0.798	0.882	0.902
F1+F3: +SRC Score	0.783	0.871	0.891
F1+F4: +Metadata	0.776	0.864	0.886
F1+F2+F3+F4: Full HFF	**0.831**	**0.912**	**0.934**

The ablation reveals three notable findings. First, stylometric features (F2) yield the largest individual gain over the TF-IDF baseline, improving RF accuracy by 3.1 percentage points. Second, the SRC score (F3) provides an independent gain of
2.0 points, confirming that sentiment-rating inconsistency carries discriminative information beyond lexical content. Third, the full HFF system surpasses all individual-addition variants, indicating that the four feature groups are complementary rather than redundant.

A. Comparative Results

Table III presents the full evaluation of the four models on the TF-IDF-only baseline versus the proposed HFF framework. Naive Bayes (NB) is included as a reference point for the standard text classification baseline.

RF with HFF achieves the highest performance across all metrics. The gain over the TF-IDF-only RF baseline is 6.3 accuracy percentage points and 6.3 F1 points, a substantial improvement attributable to the complementary stylometric and SRC signals.

The near-symmetric per-class performance demonstrates that the model does not exhibit a systematic bias toward over-flagging genuine reviews as deceptive (low false-positive rate) or under-detecting deceptive reviews (high recall on the deceptive class).

XI. DISCUSSION

Contribution of SRC score. The SRC feature alone yields a
2.0 percentage-point accuracy gain over TF-IDF in the RF set-ting, despite being a single scalar value. This disproportionate impact is consistent with the hypothesis that sentiment-rating inconsistency is a high-precision deception signal: reviewers who fabricate extreme ratings while writing neutral or contradictory text are reliably flagged. This observation motivates its inclusion as a standard feature in future fake review detection systems.

Contribution of stylometric features. The 3.1 percentage-point gain from stylometric features supports prior findings that deceptive authors exhibit distinctive writing style sig-natures [1], [6], [22]. Inspection of feature importances in the trained RF model shows that TTR and PPR are the two highest-weight stylometric features, consistent with the observation that deceptive reviews tend to be lexically less diverse and more self-referential.

Synergistic gains. The full HFF system outperforms all individual-feature-group additions, confirming the complementarity hypothesis. The gain from combining all four groups (6.3 points over TF-IDF alone) exceeds the sum of

individual gains (3.1 + 2.0 + 1.5 = 6.6), indicating a modest positive interaction effect.

Computational cost. All four feature modules are computationally lightweight. Stylometric computation and VADER polarity scoring add negligible overhead relative to TF-IDF vectorization. The inference time per review on a standard CPU is below 5 ms, confirming suitability for real-time deployment.

Limitations. The SRC feature requires the numeric star rating to be available, which limits its applicability to platforms that expose rating-text pairs. Additionally, the stylometric features may be less discriminative for very short reviews (fewer than 10 words), where sentence-level statistics are unreliable. Performance on LLM-generated fake reviews—a rapidly emerging threat—remains untested and constitutes important future work.

XII. Advantages

- **Improved detection accuracy.** The HFF framework improves accuracy by up to 6.3 percentage points over the TF-IDF-only baseline, establishing a new lightweight state-of-the-art for the combined Amazon-Yelp setting.
- **Modular and interpretable.** Each feature group is in-dependently defined, computed, and ablated, enabling practitioners to select the most appropriate subset for their deployment context and to explain model decisions in terms of linguistically meaningful signals.
- **No deep learning required.** The system achieves strong performance without neural architectures, removing GPU dependencies and enabling deployment on commodity hardware.
- **Deployment-ready latency.** Inference time below 5 ms per review supports integration into real-time review moderation pipelines.
- **Reproducible.** All dataset sizes, splits, hyperparameters, and feature configurations are fully documented (see Sections VII–IX), enabling direct experimental replication.

XIII. Limitations

- The SRC score requires access to numeric star ratings, which may not be available in all review contexts (e.g., open-text forums without structured rating fields).
- Stylometric features are less reliable for very short re-views, as sentence-level statistics become unstable below approximately 10 tokens.
- The current system does not address the emerging threat of LLM-generated fake reviews, which may exhibit high lexical diversity and stylometric regularity that differs from human-authored deceptive content.
- Performance has not been evaluated in a cross-domain setting (e.g., training on Amazon, testing on Yelp), which is an important dimension of robustness for real-world deployment.
- The reliance on supervised learning requires high-quality labelled data, which is expensive to obtain at scale for new domains or languages.

XIV. Future Work

The most pressing direction for future work is extending the framework to detect LLM-generated fake reviews. Such reviews are generated by large language models (e.g., GPT-4) and exhibit markedly different statistical properties from human-authored deceptive content—particularly higher lexical diversity and more consistent syntactic structure. Detecting them may require novel features targeting the statistical fin-gerprints of LLM output, such as perplexity-based measures or token probability distributions.

A second avenue is cross-domain generalization. Training on Amazon data and evaluating on Yelp (and vice versa) would rigorously test whether the SRC and stylometric features transfer across domains. Domain adaptation techniques such as adversarial feature alignment could be incorporated into the HFF pipeline.

Third, incorporating reviewer behavioral metadata—such as account age, review posting frequency, and verified-purchase status—into the fusion vector could further improve recall on deceptive reviews that use careful language but exhibit anomalous behavioral patterns.

Finally, extending the evaluation to include an adversarial setting—where an adversary with knowledge of the system attempts to craft reviews that evade detection—would provide a stronger robustness guarantee and is an important step

toward deployment in adversarial real-world environments.

XV. Conclusion

This paper proposed the *Hybrid Feature Fusion* (HFF) framework for deceptive review detection, addressing a specific gap in the literature: the absence of a formally defined, computationally lightweight system that combines lexical, stylometric, and sentiment-rating consistency features in a single, ablation-validated representation.

The key contribution is a novel *Sentiment-Rating Consistency* (SRC) score that quantifies the discrepancy between review text polarity and numeric star rating—a signal that is both theoretically motivated and empirically effective. Combined with a six-dimensional stylometric feature vector and standard TF-IDF representations, the HFF framework trained with a Random Forest classifier achieves 93.4% accuracy on a balanced 24,000-review corpus, outperforming the TF-IDF-only baseline by 6.3 percentage points.

The system is computationally lightweight, fully reproducible, and deployable in real-time e-commerce moderation pipelines without GPU infrastructure. A controlled ablation study isolates the contribution of each feature group, providing a transparent analysis of the relative importance of lexical, stylometric, and consistency signals.

Future work will address LLM-generated fake reviews, cross-domain generalizations, and adversarial robustness.

References

[1] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proc. ACL*, Portland, OR, 2011, pp. 309–319.

[2] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA: Morgan & Claypool, 2012.

[3] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What Yelp fake review filter might be doing?" in *Proc. ICWSM*, 2013.

[4] M. Ott, C. Cardie, and J. T. Hancock, "Estimating the prevalence of deception in online review communities," in *Proc. WWW*, Lyon, France, 2012, pp. 201–210.

[5] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Proc. IJCAI*, Barcelona, Spain, 2011, pp. 2488–2493.

[6] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. ACL*, Jeju Island, Korea, 2012, pp. 171–175.

[7] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. WSDM*, Palo Alto, CA, 2008, pp. 219–230.

[8] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," in *Proc. ICDM*, Shenzhen, China, 2014, pp. 899–904.

[9] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection," in *Proc. ICWSM*, 2013.

[10] Q. Li, X. Chen, and L. Liu, "Deep learning for detecting fake reviews,"
IEEE Access, vol. 7, pp. 112063–112074, 2019.

[11] Y. Ren and D. Ji, "Neural networks for deceptive opinion spam detection: A vector space model," *Information Sciences*, vol. 415–416, pp. 198–213, 2017.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, 2019, pp. 4171–4186.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NeurIPS*, Lake Tahoe, NV, 2013, pp. 3111–3119.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] Yelp, "Yelp Open Dataset," 2023. [Online]. Available: https://www.yelp. com/dataset

[17] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proc. RecSys*, Hong Kong, China, 2013, pp. 165–172.

[18] Amazon, "Amazon Customer Reviews Dataset," AWS Open Data, 2018. [Online]. Available: https://s3.amazonaws.com/amazon-reviews-pds/ readme.html

[19] C. Manning, P. Raghavan, and H. Schu¨tze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

[20] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2023.

[21] R. Mohawesh *et al.*, "Fake reviews detection: A survey," *IEEE Access*, vol. 9, pp. 65771–65802, 2021.

[22] N. Hussain, H. T. Mirza, I. Hussain, F. Iqbal, and I. Memon, "Spam review detection using the linguistic and spammer behavioral methods," *IEEE Access*, vol. 8, pp. 53801–53816, 2020.

[23] R. Barbado, O. Araque, and C. A. Iglesias, "A framework for fake review detection in online consumer electronics retailers," *Information Processing & Management*, vol. 56, no. 4, pp. 1234–1244, 2019.

[24] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "NetSpam: A network-based spam detection framework for reviews in online social media," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1585–1595, 2017.