

Detecting AI-Generated Product Reviews Using Classical Machine Learning Techniques

Sebastian TONU, Ioana-Alexandra TONU
Gheorghe Asachi Technical University of Iași, Romania
sebastian.tonu@academic.tuiasi.ro, ioana-alexandra.tonu@student.tuiasi.ro

The increased use of online product reviews has made them a key factor in consumer decision-making processes within e-commerce platforms. In this context, recent advances in large language models (LLMs) have enabled the generation of artificial reviews that closely resemble human-written content, raising concerns regarding their impact. This paper proposes a practical approach for detecting AI-generated product reviews based on classical machine learning techniques. A dataset is constructed by combining real user reviews with content generated using a large language model, followed by a preprocessing stage for data standardization. The classification task is addressed using TF-IDF features and a linear Support Vector Machine model. In addition, an explainability component based on SHAP is employed to analyze the contribution of individual linguistic features to the model predictions. The experimental results indicate that the proposed approach achieves high classification performance while also providing interpretable insights into the patterns that differentiate generated and authentic reviews.

Keywords: Fake reviews, LLM, TF-IDF, Support Vector Machine, Explainable AI

DOI: 10.24818/issn14531305/30.3.2026.06

1 Introduction

Online product reviews are a widely used source of information in e-commerce systems, helping users in the absence of direct interaction with products [1], [2], [3]. Empirical studies show that a significant proportion of consumers rely on such reviews during the decision-making process [1], [2], [3], while a non-negligible fraction of the available content is estimated to be misleading or inauthentic [4], [5]. Given the scale at which this data is generated and consumed, even a moderate rate of unreliable reviews can affect a large number of transactions and alter the perceived quality of products [2], [4]. In these circumstances, identifying inauthentic review content becomes a necessary component in maintaining the integrity of online marketplaces [4], [5]. Large language models enable the generation of structured and fluent text that can approximate human-written product reviews [6], [7]. When applied at scale, such systems can produce content that is difficult to distinguish from authentic user feedback. In such a scenario, the presence of artificially generated reviews increases the ambiguity of the available information [7], [8].

Existing methods for detecting deceptive reviews are typically developed for human-generated content and rely on assumptions that do not fully apply to text produced by large language models. Due to the increased regularity and fluency of generated text, the performance of such methods may degrade in this setting [7], [8]. A significant part of the existing work focuses on classification accuracy, while offering limited insight into the factors driving model decisions [5], [8], [9]. This creates the need for approaches that combine effective detection with interpretable representations of the underlying patterns [9], [10]. This paper proposes a practical approach for detecting AI-generated product reviews based on classical machine learning techniques. A dataset is constructed by combining real user reviews [11] with artificially generated content obtained using a large language model, followed by a preprocessing stage aimed at removing noise and standardizing the input data. The classification task is addressed using a *TF-IDF* representation and a *linear Support Vector Machine* model, selected for its simplicity and effectiveness in high-dimensional text spaces. In addition, an explainability

component based on *SHAP* is introduced to analyze the contribution of individual linguistic features to model predictions. The results show that the proposed approach achieves high classification performance, while also providing interpretable insights into the patterns that differentiate generated and authentic reviews.

From a practical perspective, the use of classical machine learning models is motivated not only by their competitive performance in sparse text classification settings, but also by their relative simplicity, reproducibility, and lower computational cost. These characteristics make such models suitable for integration into e-commerce review auditing workflows, where transparent decision support may be preferable to more complex black-box alternatives.

2 Related Work

Online product reviews have been extensively examined in the e-commerce literature due to their influence on consumer judgment and purchasing behavior [1], [2]. Early studies showed that recommendations and review-based signals can affect online choices, while subsequent work demonstrated that both the presence and the perceived quality of reviews may influence product evaluation and sales performance [1], [2], [3]. In this context, review content has been treated not only as a source of consumer opinion, but also as a decision signal with direct relevance for digital marketplaces [2], [3]. This perspective explains the increasing interest in methods for analyzing review text and, more recently, for assessing its credibility [4], [5].

Research on review authenticity has developed mainly under the broader framework of review spam and deceptive review detection. Early work formalized the notion of opinion spam and showed that false or manipulative reviews require dedicated detection strategies rather than direct reuse of methods developed for other forms of spam [4], [12]. Other studies approached deceptive review detection as a supervised text classification problem and showed that linguistic features can provide substantial discriminatory power, even when

human judgment remains unreliable [13]. Later work extended this direction by incorporating reviewer behavior and behavioral footprints, moving beyond purely lexical approaches [14], [15]. More recent studies have further explored richer textual representations and e-commerce-specific settings to improve the detection of false reviews across diverse product categories [5], [16]. The recent development of large language models has introduced a distinct layer of complexity to the problem of review authenticity [8], [17]. Unlike traditional deceptive reviews, which were typically written by humans and often retained observable stylistic irregularities, machine-generated reviews can be produced at scale and may exhibit a higher degree of fluency, regularity, and contextual coherence [6], [7]. This shift has motivated a growing body of work on AI-generated text detection, including studies that address the specific case of machine-generated fake reviews in e-commerce environments [8], [17]. At the same time, recent research has placed increasing emphasis on model interpretability, arguing that high classification performance alone is not sufficient when the objective is to understand the linguistic patterns associated with manipulated content [9], [10], [18]. Interest in the broader role of AI in e-commerce has also been noted in the Romanian context, where recent work has examined user trust in AI-based systems [19], while the expansion of digitally mediated commerce has further increased the relevance of reliable automated content assessment [20].

Despite these advances, several limitations remain visible in the current literature. A substantial part of existing work addresses either traditional deceptive reviews or AI-generated text in a more general setting, while fewer studies examine AI-generated product reviews in a focused e-commerce context [5], [6], [7], [8], [17]. Many contributions remain primarily oriented toward classification performance, whereas the linguistic factors underlying model decisions are discussed less systematically [9], [10], [18]. Against this background, the present study is positioned at the intersection of e-commerce review

analytics, AI-generated content detection, and explainable machine learning, with an emphasis on simple and reproducible text classification models applied to a dataset combining real reviews and *LLM*-generated content [11].

3 Dataset Construction

3.1 Real and AI Data

The dataset used in this study consists of both real user-generated product reviews and artificially generated content. The real reviews were collected from an existing dataset [11] and filtered to remove entries with missing values or extremely short text, ensuring a minimum level of content consistency.

Artificial reviews were generated using a large language model, namely Gemma 2B instruction-tuned model (*google/gemma-2-2b-it*) [21], with the objective of producing short product reviews that resemble typical user feedback. The generation process was guided by prompts designed to produce reviews of 3 to 5 sentences, combining positive aspects with mild negative observations in order to approximate realistic writing patterns [8]. Additional constraints were applied to avoid meta-textual elements and template-like artifacts.

Table 1. Dataset distribution

Class	Count
Real	3000
AI	2893
Total	5893

The final dataset was obtained by combining the two sources into a unified collection of

labeled samples, where real reviews were assigned label 0 and generated reviews label 1. The resulting dataset (Table 1) contains 5,893 samples, including 3,000 real reviews and 2,893 AI-generated reviews.

The dataset is approximately balanced, allowing a fair evaluation of the classification models without requiring additional balancing techniques.

3.2 Data Cleaning

A preprocessing stage was applied to standardize the textual data and remove artifacts that could introduce bias in the classification process [5]. *HTML* elements and residual tokens such as *br* were eliminated, as they do not represent meaningful linguistic information and may act as unintended discriminative features. Duplicate entries were removed to avoid redundancy, and reviews with extremely short or inconsistent content were filtered out. Basic normalization steps were also applied to ensure a consistent format across all samples.

The resulting dataset contains clean and homogeneous text data, suitable for feature extraction and subsequent classification. To provide a compact overview of the cleaned corpus, several descriptive statistics were computed separately for the two classes. The results are presented in Table 2.

The statistics show that AI-generated reviews are, on average, longer than authentic reviews, both in terms of word count and sentence count. These differences suggest that the two classes vary not only in lexical content, but also in basic structural properties of the text.

Table 2. Descriptive statistics of the cleaned dataset

Class	Avg. word count	Avg. sentence count	Avg. char count
Real	36.80	3.19	188.75
AI-generated	42.21	4.11	201.25

4 Methodology

4.1 Feature Extraction

The textual data was transformed into numerical representations using two feature extraction strategies (Fig. 1). The first strategy was based on simple surface features (Table 3),

intended to capture general structural characteristics of the reviews, such as text length, punctuation-related information, and other basic stylistic indicators. This representation was used for the baseline configuration.

Table 3. Surface features for baseline

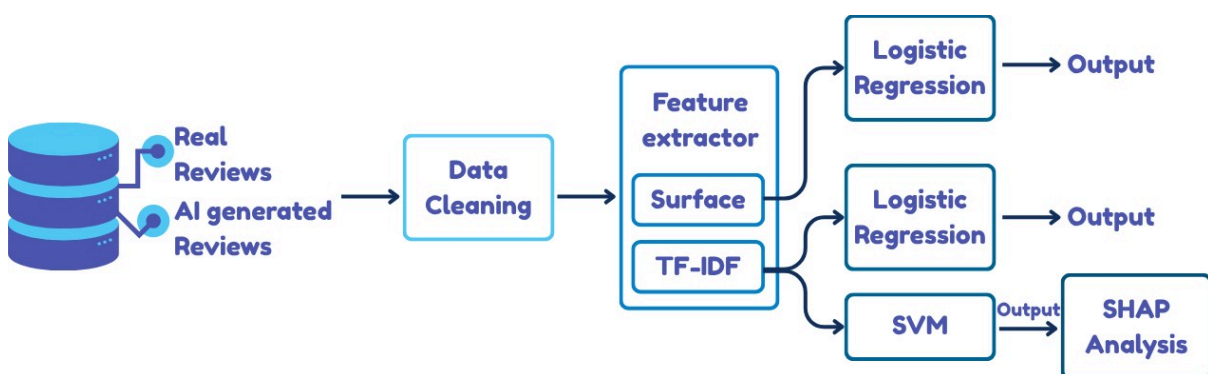
Feature	Description
<i>num_chars</i>	Total number of characters in the review
<i>num_words</i>	Total number of words
<i>num_sentences</i>	Total number of sentences
<i>avg_word_length</i>	Average number of characters per word
<i>avg_sentence_length</i>	Average number of words per sentence
<i>unique_words</i>	Number of distinct words
<i>type_token_ratio</i>	Ratio between unique words and total words
<i>num_exclamations</i>	Number of exclamation marks
<i>num_questions</i>	Number of question marks
<i>num_digits</i>	Number of numeric characters

The second strategy was based on the *Term Frequency–Inverse Document Frequency (TF-IDF)* approach [22], [23]. In this representation, each review is encoded as a vector in a high-dimensional feature space, where each dimension corresponds to a term or a sequence of terms extracted from the corpus. The *TF-IDF* weight of a term t in a document d is defined as:

$$TF-IDF(t, d) = tf(t, d) \cdot \log \left(\frac{N}{df(t)} \right)$$

where $tf(t, d)$ represents the term frequency, $df(t)$ denotes the document frequency, and N is the total number of documents in the dataset. The *TF-IDF* representation reduces the influence of frequently occurring terms that carry limited information, while emphasizing

terms that are more specific to individual reviews [22], [23]. In this study, both unigrams and bigrams were considered in order to capture not only individual word usage, but also short contextual patterns that may differentiate generated and authentic content. The inclusion of bigrams allows the model to capture short patterns that may reflect common phrasing used in generated content. The *TF-IDF* representation was selected due to its simplicity and effectiveness in text classification tasks [23], [24]. In addition, a supplementary sensitivity analysis was conducted in order to compare the unigram–bigram representation with a unigram-only *TF-IDF* configuration for the lexical classification models.

**Fig. 1.** The general architecture of the project

4.2 Classification Models

The classification task was addressed through three experimental configurations derived from the combination of two feature representations and two classifiers (Fig. 1). The baseline configuration used *Logistic Regression* applied to the surface features described

above. This setting was intended to provide a reference point for evaluating the contribution of richer lexical representations.

A second configuration employed *Logistic Regression* on top of the *TF-IDF* representation. This experiment was introduced to assess the effect of lexical information

independently of the final classifier and to measure the improvement obtained when moving from surface-level descriptors to sparse term-based features. The third configuration corresponds to the main model of the proposed framework. In this case, a *linear Support Vector Machine (SVM)* classifier was applied to the same *TF-IDF* feature space. This model was selected due to its ability to effectively handle high-dimensional sparse data, which is characteristic of text-based representations, as well as its strong performance in similar classification tasks [24].

The dataset was divided into training and testing subsets in order to evaluate the generalization capability of the models. Standard classification metrics were used to assess performance. An explainability analysis was further conducted using *SHAP* [10], applied to the *SVM* model trained on *TF-IDF* features, to investigate the contribution of individual linguistic features to the prediction process.

5 Experimental Results and Evaluation

5.1 Experimental Setup

The experimental evaluation was conducted on the dataset described in Section 3, consisting of 5,893 labeled reviews, including 3,000 real samples and 2,893 AI-generated samples. Since the class distribution is approximately balanced, no additional resampling techniques were applied. The dataset was divided into training and testing subsets using a stratified 80/20 split, to preserve the class proportions in both partitions and to ensure a fair evaluation of the classification models. All experiments were implemented in Python using *scikit-learn*. The *TF-IDF* vectorizer used lowercase normalization, unigram or unigram-bigram features depending on the experiment, and a fixed maximum vocabulary size of 5,000 features. The Logistic Regression and Linear SVM classifiers were trained using the same training partition for all configurations.

Three experimental settings were considered. The first experiment corresponds to a baseline model based on simple surface features extracted from the review text. These features capture general structural properties, such as text length, punctuation-related information, and other basic stylistic indicators. The baseline classifier was implemented using *Logistic Regression* and was intended to provide a reference point for evaluating the contribution of richer lexical representations.

The second experiment uses a *TF-IDF* representation computed over unigrams and bigrams, combined with a *Logistic Regression* classifier. This experiment was introduced in order to evaluate the effect of lexical information independently of the final model choice and to measure the improvement obtained when moving from surface-level features to a sparse term-based representation.

The third experiment corresponds to the main configuration of the proposed framework. In this case, the same *TF-IDF* representation was used together with a *linear Support Vector Machine* classifier. This model was selected as the main classifier due to its suitability for high-dimensional sparse data and its strong performance in text classification tasks.

The evaluated models were assessed using standard classification metrics, namely accuracy, precision, recall, and F1-score [25]. In addition, the confusion matrix of the best-performing model was analyzed to examine the distribution of correct and incorrect predictions across the two classes. In addition to the three main experimental settings, a supplementary sensitivity analysis was performed in order to examine the effect of the *TF-IDF* n-gram configuration, by comparing unigram-only and unigram-bigram representations for the lexical models.

5.2 Results

The classification results obtained for the three experiments are summarized in Table 4.

Table 4. Classification performance of the evaluated models

Configuration	Accuracy	Precision	Recall	F1-score
LR + surface features	0.811	0.805	0.807	0.805
LR + TF-IDF	0.953	0.948	0.964	0.955
SVM + TF-IDF	0.979	0.985	0.973	0.978

The baseline model reached an accuracy of 0.811, indicating that simple structural features already contain discriminative information for separating real and AI-generated reviews. A clear increase in performance was observed after introducing lexical information through the *TF-IDF* representation. In this setting, *Logistic Regression* achieved an accuracy of 0.953, showing that term-based features provide a substantially stronger representation than surface-level descriptors alone. The highest performance was obtained with the *linear SVM* classifier applied to the same *TF-IDF* feature space. In this configuration, the model reached an accuracy of 0.979, with similarly high values for precision, recall, and F1-score. The limited difference between the second and third experiments suggests that the main contribution to performance is provided by the lexical representation, while the choice of classifier yields a smaller gain.

The confusion matrix associated with the best-performing model is presented in Fig. 2. It indicates that most reviews were assigned to the correct class, while the number of misclassifications remained low for both real and AI-generated samples. A more detailed interpretation of the linguistic patterns captured by the model is provided in Section 6.

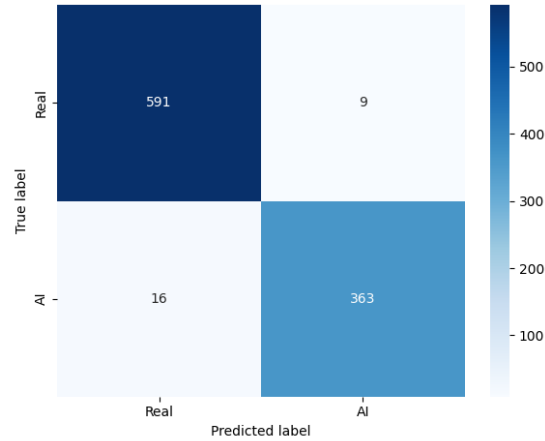


Fig. 2. Confusion matrix for the TF-IDF + Linear SVM model

5.3 Error Analysis

The error patterns summarized in Table 5 indicate that the remaining classification errors are not random, but are concentrated in cases where the lexical profile of the two classes partially overlaps. In particular, false positives tend to occur when authentic reviews contain generic positive language, broad evaluative expressions, or balanced formulations similar to those observed in generated content. Such cases suggest that some real reviews naturally exhibit stylistic patterns that resemble the regularized phrasing learned by the model from the AI-generated class.

Table 5. Error patterns observed in misclassified reviews

Review type	True label	Predicted label	Possible explanation
Real review with generic positive tone	Real	AI	resembles templated evaluative language
AI-generated review with natural phrasing	AI	Real	lacks strong lexical regularities
Real review with balanced style	Real	AI	contains mild contrast markers similar to generated text

False negatives appear when AI-generated reviews are expressed in a more concise and less formulaic manner, without the stronger lexical regularities that typically characterize generated text. This behavior indicates that the proposed approach is highly effective, but remains sensitive to borderline examples in which the stylistic distinction between authentic and generated reviews is attenuated. Overall, the error pattern is consistent with the

explainability analysis presented in Section 6, which showed that the classification is driven primarily by lexical and stylistic regularities rather than by product-specific semantic content.

5.4 Ablation Study on N-gram Configuration

In order to further examine the contribution of short contextual information to the detection

task, an additional sensitivity analysis was conducted on the lexical models. More specifically, the *TF-IDF* representation based on unigrams and bigrams was compared against a unigram-only configuration for both

Logistic Regression and *linear SVM*. The purpose of this analysis was to evaluate whether the inclusion of bigrams provides a measurable improvement beyond single-term lexical cues.

Table 6. Effect of n-gram configuration on lexical model performance

Model	TF-IDF features	Accuracy	Precision	Recall	F1-score
Logistic Regression	Unigrams	0.935	0.925	0.945	0.935
Logistic Regression	Unigrams + Bigrams	0.953	0.948	0.964	0.955
Linear SVM	Unigrams	0.958	0.943	0.964	0.953
Linear SVM	Unigrams + Bigrams	0.979	0.985	0.973	0.978

The results presented in Table 6 show that the inclusion of bigrams leads to a consistent improvement for both lexical classifiers. For *Logistic Regression*, the unigram–bigram configuration improves the accuracy from 0.935 to 0.953 and the F1-score from 0.935 to 0.955. A similar effect can be observed for the *linear SVM* model, where the accuracy increases from 0.958 to 0.979 and the F1-score from 0.953 to 0.978. These results indicate that short multi-word patterns provide additional discriminative information beyond isolated lexical terms.

The observed gain is consistent with the nature of the task. AI-generated reviews often contain recurrent short expressions and standardized evaluative phrases that are only partially captured at unigram level. By incorporating bigrams, the *TF-IDF* representation becomes more sensitive to such local contextual patterns, which contributes to the improved separation between authentic and generated reviews. This finding further supports the use of the unigram–bigram configuration in the final experimental setting.

6 Explainability Analysis

The classification results indicate that the proposed framework is effective in distinguishing between real and AI-generated reviews. However, predictive performance alone does not explain which linguistic patterns drive the model decisions. This aspect is particularly relevant in the context of AI-generated fake

review detection, where the distinction between manipulated and authentic content may depend on subtle lexical regularities rather than on obvious semantic cues [8], [17]. For this reason, an additional explainability analysis was conducted in order to identify the features that contribute most strongly to the classification process. The analysis was performed using *SHAP*, a model-agnostic explanation framework that quantifies the contribution of individual features to a given prediction [10]. The relevance of such an approach is also supported by recent work on explainable fake review detection and explainable e-commerce review analytics, which emphasizes that transparent model behavior is essential for understanding the mechanisms underlying manipulated textual content [9], [18]. In the present study, *SHAP* was applied to the *linear SVM* model trained on *TF-IDF* features, corresponding to the best-performing configuration of the proposed framework [10].

Figure 3 presents the *SHAP* summary plot obtained for the *linear SVM* model trained on *TF-IDF* features. Positive *SHAP* values indicate features that increase the probability of the AI-generated class, whereas negative values correspond to features more strongly associated with authentic reviews. The plot shows that the classification is influenced mainly by a limited subset of lexical features, including little, really, product, wish, bit, recommend, great, and definitely. These results

suggest that the model captures recurring lexical and stylistic regularities rather than product-specific semantic content. The presence of general evaluative expressions and mild

balancing markers further indicates that generated reviews tend to follow more uniform and predictable patterns.

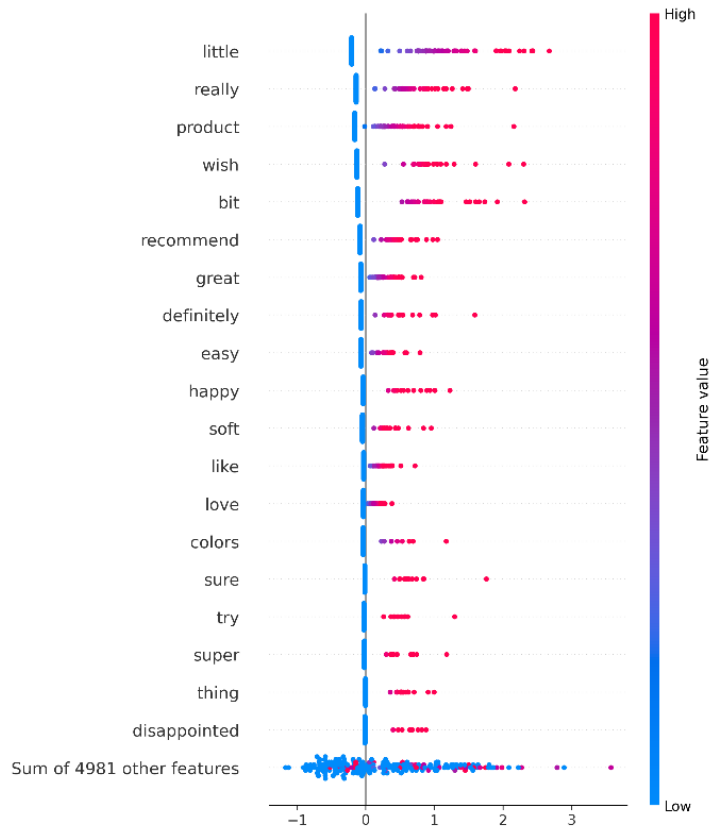


Fig. 3. SHAP summary plot for the TF-IDF + Linear SVM model

The feature distribution shown in Figure 3 suggests that the classification is influenced mainly by recurrent lexical regularities rather than by product-specific semantic content. Positive evaluative terms such as *recommend*, *great*, *happy*, and *easy* are among the most relevant indicators of the AI-generated class, while terms such as *little*, *bit*, and *wish* contribute positively when they occur, pointing to the presence of mild balancing or attenuating constructions. These findings indicate that generated reviews tend to follow more uniform and predictable phrasing patterns. In this sense, the explainability analysis complements the quantitative results by showing that the learned separation between the two classes is associated primarily with differences in stylistic consistency and lexical formulation. The *SHAP* distribution also provides a more detailed view of how individual features influence the decision function across different

samples [10], [26]. Features placed higher in the plot have a stronger average impact on the model output, while the horizontal spread reflects the variation of their contribution across the dataset. In addition, the color gradient indicates the feature value, making it possible to distinguish between the effect of feature presence and the effect of feature absence. This aspect is particularly relevant in the present study, since some terms become strongly discriminative only when they occur in the review, whereas their absence has little or no contribution to the final prediction.

Figure 4 provides a local explanation of the decision for a single review classified as AI-generated. The waterfall plot shows that the final prediction is determined by the cumulative contribution of several positively weighted lexical features, among which *bit*, *little*, *product*, *sure*, and *recommend* have the strongest effect.

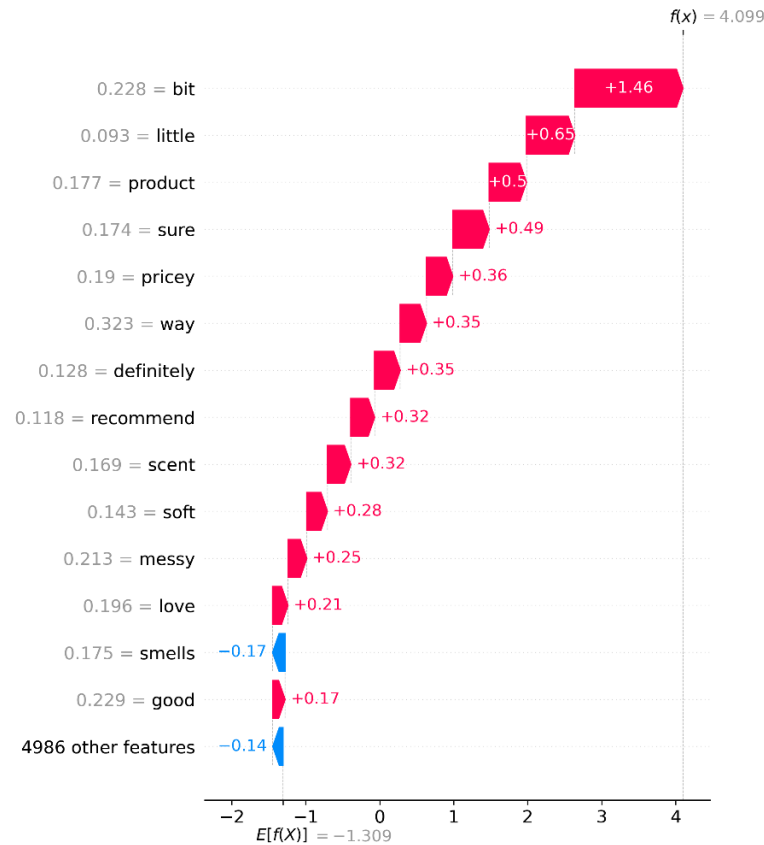


Fig. 4. SHAP waterfall plot for a representative review classified as AI-generated

A smaller number of terms, such as *smells*, contribute in the opposite direction, but their influence is insufficient to alter the final decision. This example is consistent with the global *SHAP* analysis and further suggests that the classifier relies mainly on short lexical and stylistic cues characteristic of generated reviews.

From a practical perspective, the explainability analysis is useful not only for interpreting the current classifier, but also for assessing the robustness of the overall detection framework. By identifying the lexical cues that contribute most strongly to the AI-generated class, the analysis helps determine whether the model relies on plausible stylistic regularities or on accidental artifacts of the dataset. In the present case, the most influential features are consistent with the observed properties of generated reviews, namely generic positive evaluations, mild balancing constructions, and relatively standardized phrasing. At the same time, these explanations remain specific to the current dataset and feature representation and therefore should not be interpreted as

universal markers of AI-generated reviews in all e-commerce settings.

These observations provide a useful basis for discussing both the strengths and the limitations of the proposed framework, as presented in the next section.

7 Discussion and Limitations

The results obtained in this study indicate that the proposed framework is effective for distinguishing between authentic and AI-generated product reviews. A first relevant observation is that the largest performance increase is associated with the transition from surface-level descriptors to lexical *TF-IDF* features, rather than with the choice of classifier alone. While the baseline model based on surface features already achieved a reasonable level of performance, the introduction of *TF-IDF* led to a substantial improvement, and the *linear SVM* model further refined this result. This pattern suggests that the main source of discriminative information lies in the lexical representation of the reviews, whereas the classifier contributes an additional but

comparatively smaller gain.

The supplementary analyses provide further support for this interpretation. The error analysis showed that the remaining misclassifications are concentrated in borderline cases where the stylistic profile of the two classes overlaps, rather than in cases of complete model failure. Authentic reviews may be classified as AI-generated when they contain broad evaluative language or balanced formulations similar to those found in generated content, while generated reviews may be classified as authentic when they are shorter, less formulaic, or stylistically closer to natural user feedback. In the same direction, the ablation study on the n-gram configuration showed that the inclusion of bigrams improves both lexical models, indicating that the detection task benefits not only from isolated terms, but also from short contextual patterns and recurrent multi-word expressions.

The explainability analysis is consistent with these observations. The *SHAP* summary plot suggests that the model relies primarily on lexical and stylistic regularities rather than on product-specific semantic content. Features such as *recommend*, *great*, *happy*, and *easy* are associated with a generalized evaluative style, while terms such as *little*, *bit*, and *wish* become informative when they occur in attenuating or balancing constructions. This behavior is plausible in the context of AI-generated reviews, which are often designed to sound coherent, moderately nuanced, and broadly positive. At the same time, the explanations indicate that the classifier separates the two classes mainly through patterns of wording and phrasing, not through deeper contextual understanding.

Several limitations should nevertheless be acknowledged. First, the AI-generated portion of the dataset was produced using a single large language model, namely the *Gemma 2B*, under a relatively controlled prompting strategy. As a result, the learned patterns may partly reflect the stylistic tendencies of this specific generator and prompting setup, rather than universal properties of machine-generated reviews. Second, the authentic reviews were collected from a single large-scale e-

commerce source, which limits the diversity of writing styles, product categories, and platform-specific conventions represented in the real class. Third, the task was formulated as a binary classification problem, although practical review moderation systems may require more complex distinctions, including human-written deceptive reviews, post-edited AI content, or reviews generated by multiple model families.

The proposed methodology also has limitations at representation level. Although *TF-IDF* and *linear SVM* provide strong performance and remain attractive due to their simplicity, reproducibility, and low computational cost, they operate primarily in a sparse lexical space and do not explicitly model longer-range semantic dependencies. Consequently, the approach may be less robust in situations where generated reviews are heavily paraphrased, deliberately simplified, or manually revised to reduce their stylistic regularity. In addition, the *SHAP* explanations remain specific to the present dataset and modeling configuration. The highlighted terms should therefore be interpreted as informative cues for the current corpus, rather than as stable markers of AI-generated reviews across all e-commerce settings.

Despite these limitations, the proposed framework has practical value. Its performance, relative simplicity, and interpretability make it suitable for use as a first-stage screening mechanism in e-commerce environments, where suspicious reviews could be automatically flagged for further inspection. In such a setting, the model would not replace human moderation, but could provide a transparent and computationally efficient decision-support component. Future work may extend this study by incorporating multiple text generators, more diverse sources of authentic reviews, multilingual data, and additional model families, in order to evaluate the robustness of the observed patterns under broader and more realistic conditions.

8 Conclusion

This paper investigated the problem of detecting AI-generated product reviews in e-

commerce environments using classical machine learning techniques. A binary classification framework was developed by combining authentic user reviews with content generated using a large language model, followed by a preprocessing stage designed to reduce noise and standardize the textual input. Three experimental configurations were evaluated, including a baseline based on surface features and two lexical models based on *TF-IDF* representations. The results showed that the proposed approach achieves strong classification performance, with the best results obtained for the *linear SVM* model applied to unigram–bigram *TF-IDF* features.

Beyond predictive accuracy, the study also examined the linguistic patterns underlying the classification process through *SHAP*-based explainability analysis. The obtained explanations suggest that the distinction between authentic and generated reviews is driven mainly by lexical and stylistic regularities, including generic evaluative expressions, mild balancing constructions, and relatively standardized phrasing. These findings indicate that classical text classification models remain effective for this task when combined with suitable lexical representations, while also offering the advantage of transparency and reproducibility.

At the same time, the conclusions of the study should be interpreted in relation to the scope of the dataset and the specific generation setup used for the artificial reviews. Future work may extend the present framework by considering multiple text generators, more diverse sources of authentic reviews, multilingual corpora, and more challenging settings involving post-edited or mixed-origin content. Such extensions would provide a broader evaluation of the robustness and generalizability of the proposed approach.

References

- [1] S. Senecal and J. Nantel, “The influence of online product recommendations on consumers’ online choices,” *Journal of Retailing*, vol. 80, no. 2, pp. 159–169, 2004.
- [2] J. A. Chevalier and D. Mayzlin, “The effect of word of mouth on sales: Online book reviews,” *Journal of Marketing Research*, vol. 43, no. 3, pp. 345–354, 2006.
- [3] D. H. Park, J. Lee, and I. Han, “The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement,” *International Journal of Electronic Commerce*, vol. 11, no. 4, pp. 125–148, 2007.
- [4] Y. Ma and F. Li, “Detecting review spam: Challenges and opportunities,” in *Proc. 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, Pittsburgh, PA, USA, 2012, pp. 651–654.
- [5] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, “Fake reviews detection: A survey,” *IEEE Access*, vol. 9, pp. 65771–65802, 2021.
- [6] L. Y. Wu and I. Segura-Bedmar, “AI-generated text detection with a GLTR-based approach,” *arXiv preprint arXiv:2502.12064*, 2025.
- [7] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-shot machine-generated text detection using probability curvature,” in *Proc. International Conference on Machine Learning (ICML)*, PMLR, 2023, pp. 24950–24962.
- [8] A. Bader, Y. Suhweil, D. A. Shaheen, B. Alhijawi, and S. Abu-Soud, “Detecting ChatGPT-generated fake reviews using supervised machine learning,” in *Proc. 2023 14th International Conference on Information and Communication Systems (ICICS)*, 2023, pp. 1–5.
- [9] M. Shajalal, M. Atabuzzaman, A. Boden, G. Stevens, and D. Du, “What matters in explanations: Towards explainable fake review detection focusing on transformers,” *arXiv preprint arXiv:2407.21056*, 2024.
- [10] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [11] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, and J. McAuley, “Bridging language and items for retrieval and recommendation,” *arXiv preprint arXiv:2403.03952*, 2024.

- [12] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. 2008 International Conference on Web Search and Data Mining (WSDM)*, 2008, pp. 219–230.
- [13] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 309–319.
- [14] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, vol. 2, no. 1, article 23, 2015.
- [15] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," in *Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 632–640.
- [16] P. Hajek, L. Hikkerova, and J. M. Sahut, "Fake review detection in e-commerce platforms using aspect-based sentiment analysis," *Journal of Business Research*, vol. 167, article 114143, 2023.
- [17] J. Luo, G. Nan, and D. Li, "AI-generated fake review detection," *Decision Support Systems*, vol. 203, article 114628, 2026, doi: 10.1016/j.dss.2026.114628.
- [18] R. Das, W. Ahmed, K. Sharma, M. Hardey, Y. K. Dwivedi, Z. Zhang, and R. Filieri, "Towards the development of an explainable e-commerce fake review index: An attribute analytics approach," *European Journal of Operational Research*, vol. 317, no. 2, pp. 382–400, 2024.
- [19] D. Teodorescu, K. A. Aivaz, D. P. C. Vancea, E. Condrea, C. Dragan, and A. C. Olteanu, "Consumer trust in AI algorithms used in e-commerce: A case study of college students at a Romanian public university," *Sustainability*, vol. 15, no. 15, article 11925, 2023.
- [20] A. Nistor and E. Zadobrischi, "The scale-up of e-commerce in Romania generated by the pandemic, automation, and artificial intelligence," *Telecom*, vol. 5, no. 3, pp. 680–705, 2024.
- [21] Gemma Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. Le Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, et al., "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.
- [22] M. Das and P. J. A. Alphonse, "A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset," *arXiv preprint arXiv:2308.04037*, 2023.
- [23] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge University Press, 2008.
- [24] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. European Conference on Machine Learning (ECML)*, Berlin, Heidelberg: Springer, 1998, pp. 137–142.
- [25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
- [26] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.



Sebastian TONU graduated from the Faculty of Automatic Control and Computer Engineering in Iași in 2025 and is currently pursuing a Master's degree in Artificial Intelligence at the same institution. He works as a research assistant in AI and serves as an associate teaching assistant for laboratory courses in Basics of Computer Functioning and Numerical Signal Processing. His research interests include natural language processing, neuro-symbolic AI, and embedded systems.



Ioana - Alexandra TONU graduated from the Faculty of Automatic Control and Computer Engineering in Iași in 2025 and is currently pursuing a Master's degree in Artificial Intelligence at the same institution. She works as a research assistant in AI and serves as an associate teaching assistant for laboratory courses in System Identification. Her research interests include applied computer vision, natural language processing, and explainable AI.