A Grading Methodology for E-commerce Reviews based on a Mixed Lexical-semantic Approach

Giuseppe Scarpi[1], Amir Khorrami Chokami[1], Diego Reforgiato Recupero[1,*]

[1] *Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy*

Abstract

We introduce a methodology for grading sentiment in e-commerce reviews based on a mixed lexical and semantic approach, with a strong focus on interpretability. The methodology employs a simple, replicable, and operational procedure to identify sentiment-bearing substantives in textual reviews and uses them to assign granular sentiment scores on a 1–10 scale. Experimental results show that our method achieves performance comparable to zero-shot large language models (LLMs) when benchmarked against human-assigned grades. Unlike black-box LLM approaches, we offer enhanced transparency by explicitly highlighting the linguistic elements that drive its grading decisions.

Keywords

Sentiment Analysis, E-commerce, Interpretability, LLMs, Machine Learning

1. Introduction

Explainable Artificial Intelligence (xAI) has become a critical area of research as AI systems are increasingly entrusted with complex, high-stakes decision-making tasks [1, 2]. Even before the rise of large language models (LLMs), AI was already being deployed in sensitive domains such as medical diagnosis [3] and industrial quality assurance, where transparency is essential to ensure safety and accountability. While e-commerce—the domain in which we evaluate our methodology—does not involve life-or-death consequences, it is a highly monetized sector where biases and inaccuracies can erode consumer trust and distort markets. Research has shown that numerical review scores (e.g., "star ratings") are frequently unreliable due to commercial manipulation (e.g., sellers incentivizing positive reviews through refunds or discounts) [3] and psychological biases (e.g., users defaulting to 5-star ratings unless strongly dissatisfied). As a result, experienced consumers often discount aggregated scores and instead analyze textual reviews to assess product quality [4]. This manual process can be automated through machine learning (ML), but most methods and tools available operate as opaque black-box systems, offering limited transparency into their decision-making processes.

The purpose of this research is to introduce a novel methodology designed to advance the explainability of textual sentiment analysis while maintaining robustness and operational

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

* Corresponding author.

giuseppe.scarpi@unica.it (G. Scarpi), amir.khorramichokami@unica.it (A. Chokami), diego.reforgiato@unica.it (D. Reforgiato Recupero)

simplicity. Unlike traditional black-box approaches—such as deep learning models or LLM-based evaluators—we provide transparent, auditable, and linguistically grounded sentiment grading, making it particularly suitable for high-stakes applications where interpretability is critical.

Our methodology brings four core innovations:

1. Explainability by Design: unlike neural models that obscure decision pathways, it explicitly links sentiment scores to specific textual features (e.g., sentiment-laden nouns and phrases).
2. Robustness in Complex Scenarios: we tested the methodology with e-commerce reviews, a challenging domain due to its wide and diverse spectrum of topics.
3. Reduced Tuning and Operational Simplicity, because we have no explicit hyperparameters and a minimal set of rules-based parameters, eliminating the need for labor-intensive fine-tuning.
4. Flexibility and Scalability: while optimized for e-commerce, the methodology can be extended to other text-grading tasks (e.g., customer support transcripts, social media posts).

2. Materials and methods

In this section we will discuss the dataset used, and the machine learning and LLM that we have employed within our pipeline. For NLP operations like extraction of substantive-adjective pairs, we relied on Python library Spacy 3.8.2[2] with model *en_core_web_sm*. Sentiment grading of pairs exploited the popular VADER[3] algorithm found in NLTK 3.9.1[4]. The same library was also used to clean and lemmatize the pairs. For the execution of this research, we developed some Python scripts that are publicly available at https://github.com/kalbun/GENIS.

2.1. Data

For the benchmark, we used a selection of 500 e-commerce Amazon reviews belonging to five different item categories, 100 reviews for each. All the reviews were found in [5], part of [6]. For the extraction of the reviews, we used a random generator initialized with a controlled seed for reproducibility.

To assess performance, we involved a panel of human reviewers, assigning them the task to grade the sentiment of each review from 1 (worst) to 10 (best), with half grades allowed. After filtering outliers with Interquartile Range (IQR) [12], we obtained a golden standard of 500 human-graded reviews.

2.2. VADER

For grading sentiment-loaded pairs of terms we adopted VADER (Valence Aware Dictionary and sEntiment Reasoner) [13], a popular algorithm for sentiment analysis specially developed

[2] https://spacy.io/
[3] https://vadersentiment.readthedocs.io/en/latest/
[4] https://www.nltk.org/

for sentiments expressed in social media. Unlike ML-based algorithms, VADER works in the lexical domain. We use VADER to grade the sentiment of adjective-substantive pairs in the context-unaware part of the toolchain. We have used so far only the compound value, but in the future may examine the possibility of using the other values as well.

2.3. MISTRAL

Mistral AI, a French company specializing in AI products, was established in April 2023 by former employees of Meta Platforms and Google DeepMind[5]. In this work we have adopted Mistral Small 3.1, which is a compact, yet efficient language model developed by Mistral AI, designed to balance performance with computational efficiency. Its architecture reflects Mistral's focus on optimizing smaller models without sacrificing robustness, making it a practical choice for developers seeking a lightweight but capable alternative to larger LLMs. Being interested also in assessing our methodology against other "competitors", we involved Mistral to assign a grade to each review, using the zero-shot prompt illustrated in Figure 1. The used model is mistral-small-latest, corresponding to mistral-small-2503.

3. Methodology

The idea behind this methodology is straightforward: rather than assigning a sentiment-based score to the whole text, it identifies and individually grades sentiment-laden substantives. This information is then used to train a ML algorithm that eventually makes predictions on new data. The final sentiment grade is directly influenced by the original score in stars and the sentiment of each substantive in the review. This makes it easier to trace back why a particular sentiment score was assigned to a text.
For the identification of substantives with an emotional load, we adopted a classical approach based on part-of-speech (POS) tagging. Works like [7] or [8] confirm that POS can convey sentiments more effectively when words are syntactically combined in specific ways, with the most relevant combination being substantive-adjective pairs (although other combinations like substantives-verbs or substantives-adverbs can express sentiment).
The most important phases of our methodology are then:

1. Select sentiment-laden noun-adjective pairs in the lexical domain (no context).
2. Use LLMs to grade the sentiment of selected nouns, this time within context.

This method may seem cumbersome and raises the legit question of why not using directly an LLM. The immediate answer is that it performs better than a zero-shot LLM in predicting human-generated grades (according to Wasserstein distance). Furthermore, the nouns identified in the first phase may objectively explain the grade assigned, even in long texts.

3.1. Core algorithm

Figure 1 shows the blocks that form the core of the algorithm and that calculate the number of positive, negative and neutral aspects from an input review. The entire processing algorithm is

[5] https://en.wikipedia.org/wiki/Mistral_AI

written in Python language and is available in the public repository https://github.com/kalbun/GENIS.

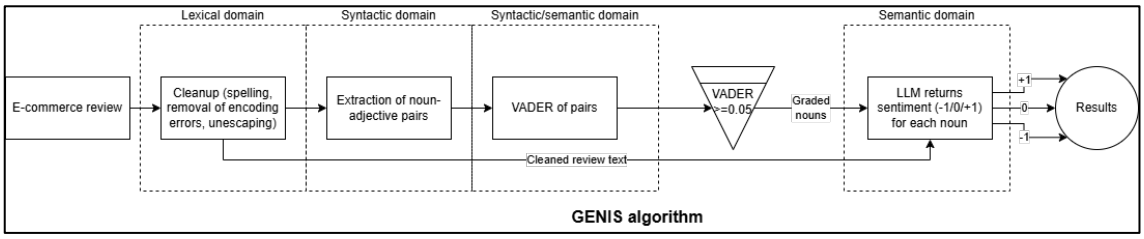

Figure 1: The core of the methodology for extraction of sentiment-laden nouns.

After cleaning up the reviews from typos and unescaped sequences we extract the pairs using Spacy's NLP processor "en_core_web_sm" model. We look for noun-adjective pairs in the two cases of adjectival modifier, like in "good music" and adjectival complement, like in "the music is good". Afterwards, VADER filters out pairs not expressing a sentiment, that is, exhibiting a VADER compound sentiment score of less than 0.05. For the second and final check in the semantic domain, we invoke Mistral AI LLM with a simple, carefully designed prompt, passing the review and the list of nouns, asking to assign a score between -1 (negative), 0 (neutral) or +1 (positive) to each of them.

The remaining, trivial steps consist in integrating the algorithm into a complete pipeline where a machine learning algorithm is trained to correlate input data (coming from our methodology) and output data (human-assigned grades) and then used for inferences and performance measurements.

4. Results and performance evaluation

To evaluate the performance of our methodology, we compared its results against two competitors: the LLM and the VADER algorithm (converted from [-1…1] to [1…10] as explained above). As a benchmark, we employed the Wasserstein distance, see [9] and [10]. Intuitively, it measures the minimal effort required to transform a distribution into another one. Formally, the Wasserstein distance of order 1 between two probability distributions P and Q over the same metric space is defined as:

$$W_1(P,Q) = \frac{1}{n}\sum_{i=1}^{n}\left|P_{(i)} - Q_{(i)}\right|$$

Where $P_{(i)}$ and $Q_{(i)}$ are the order statistics of the samples extracted from P and Q, respectively. Table 1 reports the pairwise Wasserstein distances between the benchmark scores and those produced by each computational method:

Table 1
Wasserstein distance from benchmark for the three methods

Comparison w. benchmark	Distance
Our methodology	0.665
LLM	0.720
VADER-converted	1.660

Both our methodology and LLM exhibit very similar Wasserstein distances from the benchmark, indicating a comparable level of divergence from the expert human evaluations. In contrast, VADER shows a higher distance, suggesting a substantially greater discrepancy from the benchmark distribution. These experimental results thus demonstrate that we perform comparably to a zero-shot LLM approach in terms of predicting human-generated grades but offer a significant advantage in terms of explainability.

5. Conclusions

In this paper, we presented a novel methodology for grading sentiment in e-commerce reviews that focuses on explainability. Our experimental results demonstrate performances comparable to zero-shot LLM approaches in predicting human-generated grades but with better interpretability. This is achieved by identifying sentiment-laden substantives and using them to assign grades, providing clear and interpretable explanations for the scores.

This methodology can offer significant advantages, specifically where transparency and auditability are relevant. In e-commerce platforms, it can provide clear explanations for review scores, helping to build trust among users.

6. Declaration of Generative AI

During the preparation of this work, the authors did not use any AI tool.

7. References

[1] H. Hagras, «Toward Human-Understandable, Explainable AI,» *Computer,* vol. 51, n. 9, pp. 28-36, 9 2018.

[2] S. Mohseni, N. Zarei e E. D. Ragan, «A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems,» *ACM Transactions on Interactive Intelligent Systems,* vol. 11, n. 3-4, pp. 1-45, 3 9 2021.

[3] L. Chen, W. Li, H. Chen e S. Geng, «Detection of Fake Reviews: Analysis of Sellers' Manipulation Behavior,» *Sustainability,* 2019.

[4] C. Gallagher, E. Furey e K. Curran, «The Application of Sentiment Analysis and Text Analytics to Customer Experience Reviews to Understand What Customers Are Really Saying,» *International Journal of Data Warehousing and Mining,* 2019.

[5] «UCSD Amazon Reviews,» 2023. [Online]. Available: https://amazon-reviews-2023.github.io/.

[6] Y. Hou, J. Li, Z. He, A. Yan, X. Chen e J. McAuley, «Bridging Language and Items for Retrieval and Recommendation,» *arXiv preprint arXiv:2403.03952,* 2024.

[7] P. D. Turney, «Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews,» in *Proceedings of the 40th annual meeting on association for computational linguistics,* 2002.

[8] P. Bo e L. Lee, «Opinion mining and sentiment analysis,» *Foundations and trends in information retrieval,* vol. 2, n. 1-2, pp. 1-135, 2008.

[9] C. Villani, Optimal transport: old and new, vol. 338, Berlin: Springer, 2008.

[10] V. M. Panaretos e Y. Zemel, «Statistical aspects of Wasserstein distances,» vol. 6, pp. 405-431, 2019.

[11] S. Kaur, J. Singla, L. Nkenyereye, S. Jha, D. Prashar, G. P. Joshi, S. El-Sappagh, M. S. Islam e S. M. R. Islam, «Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives,» *IEEE Access,* vol. 8, pp. 228049-228069, 2020.

[12] Dekking, Frederik Michel; Kraaikamp, Cornelis; Lopuhaä, Hen Paul; Meester, Ludolf Erwin (2005). A Modern Introduction to Probability and Statistics. Springer

Texts in Statistics. London: Springer London. Doi: 10.1007/1-84628-168-7. ISBN 978-1-85233-896-1.

[13] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.