A Comprehensive Investigation of the Impact of Locality on the Explainability of Automatic Text Classification

Liziane Santos Soares[1,2,*]

[1]Federal University of Minas Gerais, Belo Horizonte, Brazil
[2]Federal University of Viçosa, Rio Paranaíba, Brazil

Abstract

Interpretability is a key machine learning (ML) requirement. Our focus here is to investigate how locality influences interpretability in *Automatic Text Classification* (ATC) tasks. Locality refers to how close similar data points are in the feature space, quantified by their average *similarity* to neighboring instances and the *entropy* of their class distribution. We hypothesize that instances in more *homogeneous localities*, with *high similarity* and *low entropy*, tend to yield classifications that are easier to explain. To investigate this, we developed a methodology to characterize locality and identify those that may contribute more to explaining instance classification. We then select a diverse set of instances and generate explanations using word clouds based on locality information and attention weights. We conducted a user survey to evaluate the quality of these explanations, comparing locality-based insights with those from attention mechanisms. Our preliminary results indicate that explanations for instances located in regions with more *homogeneous localities* (*high similarity* and *low entropy)* are more effective (better understood by participants) in the users' opinions. This highlights the importance of locality in ATC explainability, providing valuable insights for research in explainable AI and ATC.

Building on the preliminary findings, several methodological questions remain open. In particular, important trade-offs arise when comparing explanation formats. The extension toward LLM-based summarization introduces additional challenges related to prompt design, model selection, and the balance between conciseness, coverage, and faithfulness. It remains unclear how different explanation strategies, modeling choices, and representation forms affect perceived explanation quality, as well as how such explanations should be evaluated from a user-centered perspective. Furthermore, it is still an open question whether summarization-based explanations can complement or outperform existing formats across different locality configurations.

Keywords

Explainable Artificial Intelligence, Instance Locality, Human-grounded Evaluation, Automatic Text Classification, Large Language Models, Summarization.

1. Context and Motivation

Machine learning (ML) models are increasingly applied in critical domains such as medical diagnostics, finance, and fraud detection, where reliability, accountability, and transparency are essential. Despite their high predictive performance, many models remain "*black boxes*" due to complex and opaque decision processes, hindering user trust, bias detection, and error analysis, and motivating the growing field of explainable AI (XAI) [1].

Regulatory frameworks such as the *General Data Protection Regulation* (GDPR) further reinforce this demand by granting users the right to understand automated decisions [2]. While performance is typically assessed through accuracy, real-world data variability can affect model behavior [3]. In fact, high performing (and usually more complex) models often yield predictions that are difficult to interpret [4]. Improving interpretability can calibrate user trust, support the identification of biases, and facilitate error analysis, thereby increasing confidence in model outcomes [5, 6]. Moreover, interpretability can serve as a feedback mechanism to refine models by clarifying the factors that influence predictions.

Extensive efforts have addressed interpretability through two main directions: (i) *post-hoc explanations*, which interpret model outputs without internal access, and (ii) *intrinsically interpretable models*, designed

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.
liziane.soares@dcc.ufmg.br (L. S. Soares)
0000-0003-2387-2357 (L. S. Soares)

to be transparent by construction [1]. A common assumption in post-hoc methods, including LIME and SHAP [3, 7, 8, 9, 10], is that locality facilitates interpretability. Locality refers to neighborhoods of similar instances in the representation space [1, 11]. However, this assumption remains underexplored, particularly regarding *why* and *under which conditions* locality improves explanation quality.

Although these methods rely on locality, defining meaningful neighborhoods is challenging [12], and different construction choices may constrain interpretability outcomes. Moreover, factors such as feature correlation, data dispersion, high dimensionality, and bias further influence the effectiveness of locality [13, 11, 14]. This suggests that locality exhibits structural variations that impact its role in explainability.

In this context, this work conducts a *comprehensive investigation of the impact of locality on the explainability of automatic text classification*, proposing a methodology to characterize local structures and identify configurations that support more faithful and informative explanations.

2. Related Work

Explainable Artificial Intelligence (XAI) has emerged in response to the growing adoption of complex ML models, particularly in NLP, where transparency, accountability, and user trust are critical concerns [15]. In the literature, explainability and interpretability are often used interchangeably to denote the ability to make model decisions understandable to humans. However, some works distinguish interpretability as an intrinsic property of transparent models and explainability as post-hoc reasoning applied to opaque systems [12]. Within this landscape, post-hoc approaches such as LIME [3] and SHAP [10] have become prominent, relying on local approximations or feature attributions to explain individual predictions.

A central yet often implicit concept in these methods is locality. Many explanation techniques assume that local neighborhoods provide meaningful insights into model behavior, either through synthetic perturbations or local sampling [11]. These approaches include local surrogate models, perturbation-based methods, and counterfactual explanations [16]. Despite their differences, they share the assumption that local regions exhibit coherent and interpretable behavior. However, prior work has highlighted that defining meaningful neighborhoods is challenging and that factors such as data sparsity, feature correlation, and high dimensionality can significantly affect explanation reliability [17, 18].

In parallel, entropy and similarity have been widely used as analytical tools in Machine Learning. Entropy quantifies uncertainty and has been applied to measure class distribution and feature relevance [19, 17], while similarity metrics underpin neighborhood-based methods such as kNN and clustering. Recent studies have explored neighborhood-based entropy measures to better capture local structure in high-dimensional spaces [18]. However, these metrics are typically used as auxiliary measures rather than as primary descriptors of explainability.

More recently, Large Language Models (LLMs) have been applied to generate textual explanations and summaries, offering a complementary perspective to traditional visual or feature-based explanations. Advances in instruction tuning and prompt engineering have improved the quality and controllability of generated summaries. Summarization techniques have evolved from extractive methods to neural and LLM-based approaches, being applied across diverse domains. Despite these advances, ensuring faithfulness and avoiding hallucinations remain key challenges [20, 21].

Overall, while prior work extensively leverages locality, entropy, similarity, and diverse explanation formats, there is still limited understanding of how the structure of local neighborhoods in the empirical data distribution influences the perceived quality of explanations. In particular, recent advances in LLM-based summarization introduce new forms of explanations whose effectiveness is not yet well understood in relation to underlying data characteristics. Existing studies often evaluate explanations in isolation, focusing on either model behavior or explanation techniques, with limited attention to how different explanation modalities (e.g., feature-based, visual, or textual summaries) interact with locality properties. As a result, there is a lack of systematic investigation into how locality affects the

user-perceived quality of different types of explanations. This gap motivates our work, which explicitly characterizes locality and investigates its impact on the perceived quality of explanations in automatic text classification.

3. Research Questions and Objectives

The main goal of this research is to investigate how locality influences the explainability of automatic text classification by proposing a methodology to characterize local neighborhoods, identify configurations that better support explanations, and generate explanations grounded in these structures.

To achieve this goal, the study is guided by four research questions:

RQ1 investigates how locality can be formally analyzed and characterized from an interpretability perspective. We define locality using Transformer-based representations [22] and kNN neighborhoods, and characterize it through two key factors: average similarity and class-distribution entropy. We hypothesize that more homogeneous localities (high similarity, low entropy) are associated with more explainable instances.

RQ2 examines whether locality-based explanations effectively improve explainability. Through a user study on AGNews and ACM datasets, we evaluate explanations using *Satisfaction, Completeness*, and *Trust*. Results indicate that explanations derived from highly homogeneous localities are consistently better perceived, highlighting the relevance of locality for interpretability.

RQ3 explores whether Large Language Models (LLMs) can enhance locality-based explanations through summarization. By generating textual summaries of neighborhoods, we aim to produce more fluent and human-readable explanations and evaluate their impact on user perception compared to other formats.

RQ4 investigates whether dataset characterization can provide a complementary, global perspective on explainability. By analyzing the distribution of locality structures, we aim to identify patterns and thresholds that may position datasets along an "*explainability scale,*" indicating varying levels of difficulty in generating meaningful explanations.

Together, these research questions provide a structured framework to analyze locality not only as a methodological component but as a fundamental factor influencing explainability, combining instance-level analysis, user-centered evaluation, and dataset-level characterization.

4. Research Approach, Methods, and Rationale

Our research comprises a sequence of activities designed to represent, analyze, and evaluate the role of locality in automatic text classification, as illustrated in Figure 1. The study methodology integrates both completed and ongoing components. The initial stages of the methodology, corresponding to **RQ1** and **RQ2**, have already been implemented and empirically validated [23]. These include dataset preparation, locality characterization, instance sampling, explanation generation using word clouds, and user-centered evaluation.

Building on these results, the next stages of the research, associated with **RQ3** and **RQ4**, are currently in progress and represent key open directions. These include the generation and evaluation of LLM-based summarization explanations, as well as dataset-level characterization aimed at understanding explainability patterns at a global level. These components introduce important methodological decisions and are central to the remaining contributions of this work.

The following subsections describe the complete methodology, explicitly distinguishing between validated components [23] and ongoing extensions that will benefit from further discussion and feedback.

4.1. Experimental Datasets

To instantiate and evaluate the proposed framework, we used two text classification datasets: *AGNews* and *ACM*. Dataset selection was guided by methodological and cognitive constraints, namely: (i) a mod-

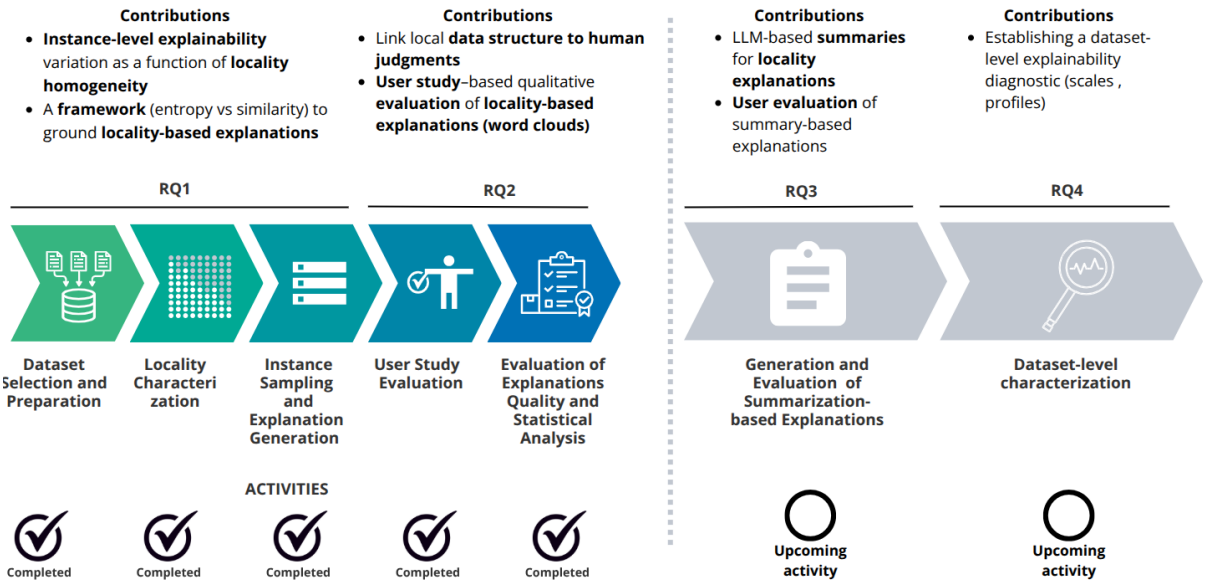

Figure 1: Study Methodology

erate number of classes, allowing meaningful variation in neighborhood entropy without overwhelming participants; (ii) texts of medium length, reducing cognitive load during explanation inspection; and (iii) domain diversity, enabling the analysis of locality effects across distinct semantic settings.

AGNews is a benchmark dataset for topic classification in which each instance contains a title and a short description, labeled into four classes: *World*, *Sports*, *Business*, and *Science/Technology* [24]. **ACM** is a curated subset of Computer Science research articles, from which we selected four representative categories: *Hardware*, *Software*, *Theory of Computation*, and *Information Systems*. Together, these datasets provide different vocabularies, styles, and domains, offering an initial basis to assess the robustness and generalizability of the proposed locality-based explainability framework.

4.2. Framework for Locality-Based Explanations

Our framework comprises a sequence of steps to characterize locality and generate explanations grounded in neighborhood structure. Each document is encoded using BERT contextual embeddings [22], and locality is estimated through k-nearest neighbors (kNN). For each instance, we compute two locality descriptors: the Shannon entropy of the neighborhood class distribution [19] and the average cosine-based similarity to neighboring instances. These descriptors support the identification of regions with different levels of locality homogeneity, directly addressing **RQ1**.

Based on this characterization, we sample representative instances from high- and low-homogeneity regions, including both correctly classified and misclassified cases. For each instance, we generate two explanation types: a *locality-based* explanation, derived from the lexical content of neighboring texts, and an *attention-based* explanation, derived from Transformer attention weights. Both are presented as word clouds, allowing a controlled comparison under the same visual metaphor and minimizing presentation bias. This component has been fully implemented and evaluated, supporting the investigation posed in **RQ2**.

The use of word clouds is a deliberate methodological choice. As compact visual summaries, they make lexical concentration, topical coherence, and dispersion patterns directly observable, which is central to our goal of relating locality structure to explanation quality [25, 26].

4.3. User Study Design

To evaluate the generated explanations from a user perspective, we conducted an online survey-based study. Participants were asked to inspect text instances together with two word clouds, one locality-based and one attention-based, and then evaluate each explanation separately and comparatively. The presentation order of the explanation types was alternated across stages to reduce positional bias.

The study targeted Computer Science researchers and practitioners with Machine Learning experience, as representatives of the intended audience. To mitigate order effects in the presentation of instances, we adopted a counterbalancing strategy based on a *Graeco–Latin Square* design [27], controlling for class and correctness–homogeneity configuration. Because participants could leave the survey at any point, two survey versions with reversed presentation order were distributed in balanced form.

4.4. Evaluation of Explanation Quality

The evaluation follows a human-centered XAI perspective, focusing on users' subjective judgments rather than causal faithfulness or task-performance gains [28, 29, 30]. We evaluate explanations along three complementary dimensions: *Satisfaction*, *Completeness*, and *Trust*.

To address **RQ3**, the framework is currently being extended to incorporate summarization-based explanations generated from neighborhood texts using instruction-tuned LLMs (e.g., Gemma, DeepSeek, Qwen, LLaMA). This stage represents an ongoing component of the research and introduces open challenges related to prompt design, faithfulness, and the balance between conciseness and informativeness.

The generated summaries may be evaluated using the same human-centered protocol, enabling direct comparison with word cloud explanations. In addition, automatic metrics such as ROUGE and BERTScore may be employed to assess textual quality and semantic fidelity. This stage is currently under development and will benefit from further methodological refinement.

Finally, as part of **RQ4**, we plan to extend the analysis to a dataset-level perspective, investigating how distributions of locality structures can inform explainability patterns. This component remains exploratory and represents a key open direction for future work.

5. Preliminary results and contributions to date

Preliminary results from a user study conducted on two datasets (AGNews and ACM), with 74 participants in total, provide consistent evidence that the structure of data locality significantly impacts perceived explanation quality [23]. Explanations derived from *high-homogeneity* neighborhoods, characterized by low entropy and high similarity, were systematically rated higher in terms of *Satisfaction*, *Completeness*, and *Trust* compared to those from low-homogeneity regions, with statistically significant differences across both datasets.

This effect was observed not only for correctly classified instances but also, to a lesser extent, for misclassified ones, indicating that locality contributes to explanation clarity even when model predictions are incorrect. Importantly, the influence of prediction correctness itself was limited and largely mediated by locality structure, suggesting that explanation quality depends more on neighborhood coherence than on model accuracy.

Comparisons with attention-based explanations further reinforce these findings. While both approaches performed similarly in the AGNews dataset, locality-based explanations consistently outperformed attention-based ones in the ACM dataset, particularly in a more specialized domain with finer-grained semantic structure. This indicates that aggregating information from semantically coherent neighborhoods provides more informative and trustworthy explanations than instance-level signals in complex settings.

These results support the central hypothesis of this research: that locality is a key factor in explainability. They demonstrate that data-centric properties, particularly neighborhood homogeneity, play a crucial role in shaping how users perceive and understand model explanations, highlighting the importance of incorporating locality-aware strategies in explainable AI for text classification.

6. Expected next research steps and expected final contribution

Building on the preliminary findings, the next stage of this research will focus on consolidating and extending the locality-based framework for explainability in automatic text classification, while addressing current limitations and open methodological questions.

6.1. Open methodological decisions

Several methodological choices remain open and will be systematically investigated in the next stages of this research. These include: (i) the sensitivity of locality characterization to different embedding models (e.g., alternative Transformer architectures), distance metrics, and neighborhood sizes; (ii) the definition of thresholds for locality homogeneity and their role in establishing interpretable "explainability regimes" and dataset-level diagnostic profiles; and (iii) the comparison between explanation formats, particularly the trade-offs between transparency, expressiveness, and faithfulness when moving from visual (word clouds) to textual (LLM-based) explanations.

In particular, the extension toward LLM-based summarization introduces additional methodological challenges related to prompt design, model selection, and the balance between conciseness, coverage, and faithfulness. This includes decisions on how to structure prompts to capture multi-topic neighborhoods, how to ensure that summaries faithfully reflect neighborhood evidence, and how different models and prompting strategies impact explanation quality.

To address these questions, an important open challenge concerns how to design and evaluate summarization-based explanations generated by LLMs in a way that preserves faithfulness to the underlying neighborhood structure while remaining concise and interpretable. It remains unclear how different models, prompting strategies, and summary representations may influence explanation quality, or how these explanations should be evaluated in a way that meaningfully captures user perception.

In addition, it is not yet established whether summary-based explanations can effectively complement or outperform existing formats, such as word clouds, particularly across different locality configurations. Understanding how these explanation modalities interact with locality properties, and under which conditions each format is more effective, remains an open question.

Furthermore, the role of dataset properties, such as class imbalance, semantic overlap, and annotation noise, will be further analyzed as mediating factors of explainability. This analysis will support the development of a dataset-level diagnostic, enabling the identification of explainability patterns and the construction of an *explainability scale* grounded in entropy and similarity.

6.2. Expected final contributions

This research is expected to deliver a comprehensive, data-centric framework for explainability in text classification, positioning locality as a foundational concept linking data distribution, model behavior, and user perception. Concretely, the thesis will contribute: (i) a formal methodology for characterizing locality based on similarity and entropy; (ii) empirical evidence, through user studies, of how locality influences perceived explanation quality; (iii) guidelines for locality-aware evaluation and instance selection; and (iv) a comparative analysis of explanation formats, including visual and LLM-based approaches.

In summary, the expected contribution is to advance explainable AI beyond model-centric perspectives by demonstrating that explainability is not only a property of models or explanation techniques, but also fundamentally shaped by the structure of the data itself.

Acknowledgments

This work was supported by Fapemig, FAPESP, CNPq, CAPES. This work was also supported by Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (grant #408490/2024-1).

This study involving human participants was reviewed and approved by the Research Ethics Committee of the Federal University of Minas Gerais (CAAE ID: 44032821.0.0000.5149).

Declaration on Generative AI

During the preparation of this work, the author(s) used Generative AI tools only for translation and minor grammatical revision of some text passages; all scientific content was produced and verified by the authors.

References

[1] A. Abusitta, M. Q. Li, B. C. Fung, Survey on explainable ai: Techniques, challenges and open issues, Expert Syst. Appl. 255 (2024) 124710.

[2] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI Mag. 38 (2017) 50–57.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?: Explaining the predictions of any classifier, in: SIGKDD, 2016, pp. 1135–1144.

[4] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: ICML, JMLR. org, 2017, pp. 1885–1894.

[5] G. Tolomei, F. Silvestri, A. Haines, M. Lalmas, Interpretable predictions of tree-based ensembles via actionable feature tweaking, in: SIGKDD, 2017, pp. 465–474.

[6] S. Das, N. Agarwal, D. Venugopal, F. T. Sheldon, S. Shiva, Taxonomy and survey of interpretable machine learning method, in: SSCI 2020, IEEE, 2020, pp. 670–677.

[7] S. Saito, E. Chua, N. Capel, R. Hu, Improving lime robustness with smarter locality sampling, arXiv:2006.12302 (2020).

[8] S. Ghalebikesabi, L. Ter-Minassian, K. DiazOrdaz, C. C. Holmes, On locality of local explanation models, Adv. Neural Inf. Process. Syst. 34 (2021) 18395–18407.

[9] R. Gaudel, L. Galárraga, J. Delaunay, L. Rozé, V. Bhargava, s-lime: Reconciling locality and fidelity in linear explanations, in: Int. Symp. on Intel. Data Analysis, 2022, pp. 102–114.

[10] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: NeurIPS, 2017, pp. 4765–4774.

[11] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2020. Online book.

[12] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM TIIS 11 (2021) 1–45.

[13] D. Christopher, P. Raghavan, H. Schütze, Introduction to information retrieval, JASIST 43 (2008) 824–825.

[14] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box dss, arXiv:1805.10820 (2018).

[15] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Inf. Fusion 76 (2021) 89–106.

[16] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM CSUR 51 (2018) 93.

[17] C. M. Bishop, N. M. Nasrabadi, Pattern recognition and machine learning, Springer, 2006.

[18] Q.-H. Hu, D.-R. Yu, Neighborhood entropy, in: Int. Conf. on ML and Cybernetics, volume 3, IEEE, 2009, pp. 1776–1782.

[19] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423.

[20] S. Minaee, et al., Large language models: A survey, arXiv (2024). doi:10.48550/arXiv.2402.06196. arXiv:2402.06196.

[21] H. Zhang, P. S. Yu, J. Zhang, A systematic survey of text summarization: From statistical methods to large language models, ACM Computing Surveys (2025). doi:10.1145/3731445, just Accepted.

[22] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, volume 1, 2019, p. 2.
[23] L. S. Soares, B. G. Lopes, J. M. Almeida, R. O. Prates, M. A. Gonçalves, Are all instances equally explainable? a study on the impact of locality on the explainability of automatic text classification tasks, in: Proceedings of the 4th World Conference on Explainable Artificial Intelligence (XAI 2026), 2026. Accepted for publication.
[24] A. Gulli, The anatomy of a news search engine, in: WWW Conf., 2005.
[25] R. M. Hicke, M. Goenka, E. Alexander, Word clouds in the wild, in: Workshop VIS4DH, 2022, pp. 43–48.
[26] M. Skeppstedt, M. Ahltorp, K. Kucher, M. Lindström, From word clouds to word rain: Revisiting the classic word cloud to visualize climate change texts, Inf. Vis. 23 (2024) 217–238.
[27] A. D. Keedwell, J. Dénes, Latin Squares and Their Applications, 2nd ed., Elsevier / North-Holland, 2015.
[28] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[29] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance, Front. Comput. Sci 5 (2023) 1096257.
[30] J. Kim, H. Maathuis, D. Sent, Human-centered evaluation of explainable ai applications: a systematic review, Front. Artif. Intell. 7 (2024) 1456486.