Explainable AI for text classification: lessons from a comprehensive evaluation of post hoc methods

Cognitive Computation (2024) 16:3077–3095
https://doi.org/10.1007/s12559-024-10325-w

REVIEW

Explainable AI for Text Classification: Lessons from a Comprehensive Evaluation of Post Hoc Methods

Mirko Cesarini[1] · Lorenzo Malandri[1] · Filippo Pallucchini[1] · Andrea Seveso[1] · Frank Xing[2]

Received: 1 April 2024 / Accepted: 3 July 2024 / Published online: 6 August 2024

Abstract

This paper addresses the notable gap in evaluating eXplainable Artificial Intelligence (XAI) methods for text classification. While existing frameworks focus on assessing XAI in areas such as recommender systems and visual analytics, a comprehensive evaluation is missing. Our study surveys and categorises recent post hoc XAI methods according to their scope of explanation and output format. We then conduct a systematic evaluation, assessing the effectiveness of these methods across varying scopes and levels of output granularity using a combination of objective metrics and user studies. Key findings reveal that feature-based explanations exhibit higher fidelity than rule-based ones. While global explanations are perceived as more satisfying and trustworthy, they are less practical than local explanations. These insights enhance understanding of XAI in text classification and offer valuable guidance for developing effective XAI systems, enabling users to evaluate each explainer's pros and cons and select the most suitable one for their needs.

Keywords Explainable AI · XAI evaluation · Text classification · Interpretability · Human-computer interaction

Introduction and Motivation

In several Machine Learning (ML) applications, the ability to explain a model's predictions and provide the rationale behind the output for any particular data point is just as important as the accuracy of those predictions in various applications [1–3]. Achieving peak accuracy on extensive modern datasets frequently entails employing complex models, like ensemble or deep learning models, and interpretation in this scenario is challenging and, in some cases, outright impossible. The trade-off between accuracy and interpretability has spurred the development of diverse methods to facilitate user comprehension of complex model predictions. Yet, how these methods address this trade-off is still the subject of ongoing research [4]. This study aims to provide a comprehensive overview of existing eXplainable Artificial Intelligence (XAI) methods documented in the literature and their suitability for text classification. Diverse data types are approached in a fundamentally distinct manner in XAI. For instance, tabular classifiers must deal with mixtures of continuous and categorical features, finding the right discretisation of the former, which can result in accurate and interpretable results at the same time [5]. XAI for images, to give another example, does not usually explain classification outputs at a single feature level (i.e. pixel level) but focuses on higher level features, often presented in the form of heatmaps or saliency maps [6]. Even the XAI methods for textual data differ regarding input features, underlying models, and output. Furthermore, Textual data need to be computationally feasible for many features, which typically comprise the dictionaries of textual corpora [7]. In this article, we narrow down the scope of the paper on XAI for text classification, which has gained great importance in academia and industry in the last years [8, 9].

✉ Lorenzo Malandri
lorenzo.malandri@unimib.it

Mirko Cesarini
cesarini.mirko@unimib.it

Filippo Pallucchini
filippo.pallucchini@unimib.it

Andrea Seveso
andrea.seveso@unimib.it

Frank Xing
xing@nus.edu.sg

[1] Department of Statistics and Quantitative Methods, University of Milan-Bicocca, Milan, Italy

[2] School of Computing, National University of Singapore, Singapore, Singapore

As the research field of XAI rapidly grows, some previous surveys [7, 10, 11] have attempted to identify the most suitable XAI methods for specific user needs. Still, as we will discuss in the "Related Works" section, none has succeeded in offering a truly exhaustive perspective, leaving users with limited guidance based on the available literature. To address this problem, we describe how to use several XAI methods in real-case scenarios, evaluating each algorithm's performance and the insights it provides.

The rationale behind preferring one XAI method over another varies depending on the specific requirements; for example, explainer *A*'s transparency might be a deciding factor, while explainer *B*'s ability to cover a broader range of data could be advantageous in others.

As previously introduced, many XAI methods are available in the literature. In the next section, we will present the decision-making process to identify the XAI methods included in this study and provide an overview of the selected techniques. Our XAI methods comparison is based on a real-world dataset, incorporating user evaluations and using existing metrics. The rationale for the approach used in this paper is that, despite that there are theoretical assurances regarding the selected XAI methods, certain properties may be compromised in specific application domains or datasets. Thus, a real-world application is essential for a comprehensive overview and to assist users in identifying and deploying the most suitable XAI method for their particular use case.

The big challenge is evaluating XAI methods since scholars from different disciplines focus on different objectives, which poses challenges for identifying appropriate design and evaluation methodology [12].

While numerous well-established works, such as Sokol and Flach [10], theoretically introduced several metrics, unfortunately, it is still unclear how to practically utilise them for comparing explanation methods [13]. Therefore, we have re-investigated the existing metrics, we chose the most suitable ones for the proposed benchmark and we developed a method for measuring them.

The contributions of this work can be summarised in three points:

(i) Gathering all these XAI methods into an Evaluation tasks will empower users of explainability systems to comprehensively assess the pros and cons of each explainer, facilitating informed decisions to select the most suitable one for their specific application.
(ii) Additionally, the proposed benchmark can serve as a valuable tool for both the development and deployment phases of explainable approaches, providing a structured checklist to ensure a thorough evaluation and to support a successful integration into various systems.
(iii) All explainers have been deployed in notebooks and are accessible through a GitHub[1] repository, promoting transparency, reproducibility, and easy adoption by the community.

All acronyms employed in this manuscript can be found in the Appendix in Table 5.

Machine Learning Explanations

We will now outline the XAI methods for classification and elucidate the process by which we selected the ones deemed suitable for our analysis. Figure 1 shows a concise and straightforward diagram that serves as a roadmap throughout the paper's literature, facilitating the understanding of the key features that comprise any XAI method, as presented in [10]. The feature descriptions in Fig. 1 are based on the framework proposed in [10]. The primary distinction between XAI methods lies in the contrast between *ante hoc* and *post hoc* methodologies. *Ante hoc* approaches employ the same model for prediction and explanation, from elucidating a linear regression through its feature weights to explaining sentiment analysis results with neural network attention weights [14]. It's crucial to acknowledge that certain techniques within this approach may be accompanied by caveats and assumptions regarding the training data or process, which must also be fulfilled for the explanation task. Still, the latter may not always be feasible. In *post hoc* approaches, predictions are generated using one model, while explanations are generated using a separate one. Post hoc approaches are this article's main focus and allow for tailoring explanations to specific information needs while keeping the prediction model untouched. Both methodologies can be further categorised into two distinct groups: *model-agnostic*, which can operate independently of any model family, and *model-specific*, which solely applies to a particular model, such as decision trees. Finally, focusing on the generalisability property, each of the previously identified (sub)groups could be divided into three stages as follows: *local*, which pertains to a single data point or prediction; *cohort*, which involves analysing a subgroup within a dataset or a subspace within the model's decision space; and *global*, which offers a comprehensive explanation of the model. We conducted a comprehensive literature review to identify the most commonly used XAI methods, and we framed the works in the categories just introduced, as reported in Table 1.

As advised by [15], we comprehensively searched electronic databases. The databases utilised for this search were as follows:

[1] https://github.com/Crisp-Unimib/XAI_Benchmark

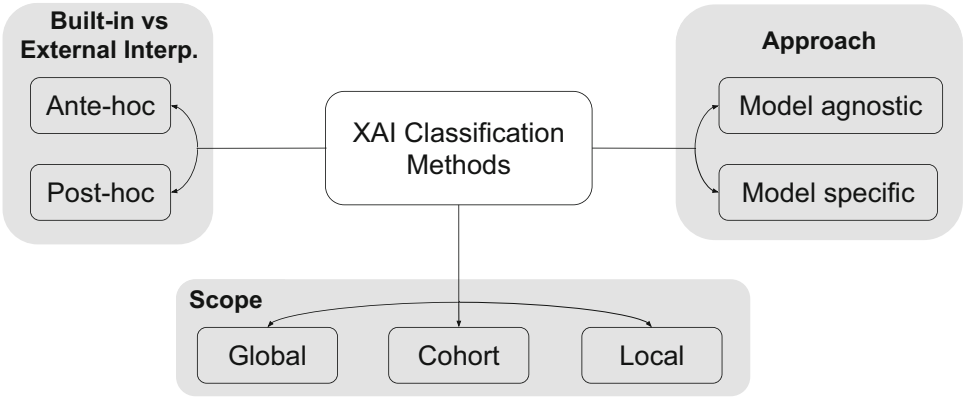


Fig. 1 Concise taxonomy of XAI methods for classification

- ACL (https://aclanthology.org/)
- Springer (www.springerlink.com)
- ACM Digital Library (www.acm.org/dl)
- ScienceDirect (www.sciencedirect.com)
- Wiley Interscience (www.Interscience.wiley.com)
- Google Scholar (www.scholar.google.co.in)
- IEEE eXplore (www.ieeexplore.ieee.org)
- Taylor Francis Online (https://www.tandfonline.com/)
- PubMed (https://pubmed.ncbi.nlm.nih.gov/)
- SemEval (https://semeval.github.io/)

We conducted an extensive literature review to identify the most pertinent XAI methods. Our search encompassed research studies from diverse sources, including conferences, journals, and arXiv. Specifically, we focused on papers published in conferences ranked as A, and journals ranked as Q1 or at least Q2. However, despite not being published in a conference or journal, we made an exception for a notable work that garnered over 500 citations since 2017 [16]. A summary of the most representative methods under consideration is reported here. We have taken into consideration 29 methods: (a) We will begin with XAI methods employing a *post hoc* approach, offering *global* explanations, and maintaining *model agnosticism*. Among these, TREPAN. Craven and Shavlik [17] stands out as one of the earliest explainers we examined. This algorithm induces a decision tree that approximates the outcome of a classifier. SAGE [18] uses shapley values to quantify the predictive power of individual input features at a global level while considering feature interactions. TREPAN constructs its tree using a hill-climbing search process and a gain ratio criterion to identify the best M-of-N splits for each node. ProfWeight [19] utilises linear probes to generate confidence scores via flattened intermediate representations, while GLRM [20] employs rule-based features for regression and probabilistic classification. These rules aid in model interpretation by capturing nonlinear dependencies and interactions. GLRM utilises column generation techniques to optimise over an exponentially large space of rules without the need to pre-generate a large subset of candidates or boost rules greedily one by one. (b) Concerning *model-specific* approaches, we find the work of Sushil et al. [21] to be particularly relevant for our purposes. Their research identifies if-then-else rules between various input features and the class labels that a trained network captures. (c) When examining *post hoc* methods with a *local* scope and *model-agnostic* nature, it's crucial to consider arguably the two most renowned XAI methods: LIME [22] and SHAP [23]. LIME elucidates the predictions of any classifier in an interpretable and accurate manner by constructing explanations locally around the prediction. On the other hand, SHAP assigns an importance value to each feature, based on the concept of Shapley values from the cooperative game theory, indicating its contribution to the model's prediction. Other methodologies within this category include the one proposed by van der Waa et al. [24], which utilises locally trained one-versus-all decision trees to identify the disjoint set of rules responsible for classifying data points as the foil rather than the fact. Another notable approach is the one introduced by Elenberg et al. [25], called STREAK, wherein the authors frame the interpretability of black box classifiers as a combinatorial maximisation problem and present an efficient streaming algorithm to solve it, subject to cardinality constraints. In addition to the previously mentioned explainers, various other approaches to explainability have been explored. These include techniques such as the one proposed by Lei et al. [26], which extracts concise and coherent pieces of input text as justifications; TCAV [27], utilising directional derivatives to measure the importance of user-defined concepts; CheckList [28], which evaluates explainers' capabilities through distinct test types; QII [29], breaking input correlations for causal reasoning and marginal influence computation; TED [30], providing explanations coherent with consumer mental models; Staniak and Biecek [31] alternative implementation of LIME; LORE [16], which employs a genetic algorithm to train local interpretable predictors for meaningful explanations; and lastly, we encountered CASME [32], an approach that involves the simultaneous training of a

Table 1 Mapping selected papers to our roadmap

Paper	Reproducibility Code	Scope Global	Cohort	Local	Output Features	Rules	Input Model Model Agnostic	Model Specific
[35] Singh et al. 2018 — ACD	git	○	○	●	○	●	○	●
[41] Wang et al. 2019 — BRS	git	○	○	●	○	●	○	●
[36] Dhurandhar et al. 2018 — CEM	git	○	○	●	●	○	○	●
[37] Shrikumar et al. 2017 — DeepLIFT	git	○	○	●	●	○	○	●
[24] van der Waa et al. 2018 — Foil Trees	git	○	○	●	○	●	●	○
[22] Ribeiro et al. 2016 — LIME	git	○	○	●	●	○	●	○
[38] Lapuschkin et al. 2016 — LRP	git	○	○	●	●	○	○	●
[39] Hu et al. 2018 — MLAM	git	○	○	●	●	○	○	●
[21] Sushil et al. 2018	git	●	○	○	○	●	○	●
[25] Elenberg et al. 2017 — STREAK	git	○	○	●	●	○	●	○
[33] Dash et al. 2018 — BRCG	git	●	○	○	○	●	○	●
[19] Dhurandhar et al. 2018 — ProfWeight	git	●	○	○	●	○	●	○
[26] Lei et al. 2016	git	○	○	●	●	○	●	○
[27] Kim et al. 2018 — TCAV	git	○	○	●	●	○	●	○
[17] Craven and Shavlik 1995 — TREPAN	git	●	○	○	○	●	●	○
[23] Lundberg and Lee 2017 — SHAP	git	○	○	●	●	○	●	○
[46] Wang and Rudin 2015 — FRL	git	●	○	○	○	●	●	○
[42] Ribeiro et al. 2018 — Anchors	git	○	●	○	○	●	●	○
[32] Zolna et al. 2020 — CASME	git	○	○	●	●	○	●	○
[28] Ribeiro et al. 2020 — CheckList	git	○	○	●	●	○	●	○
[47] Mothilal et al. 2020 — DiCE	git	○	○	●	●	○	●	○
[20] Wei et al. 2019 — GLRM	git	●	○	○	○	●	●	○
[29] Datta et al. 2016 — QII	git	○	○	●	●	○	●	○
[40] Petsiuk et al. 2018 — RISE	git	○	○	●	●	○	○	●
[30] Hind et al. 2019 — TED	git	○	○	●	●	○	●	○
[34] Selvaraju et al. 2017 — Grad-cam	git	○	○	●	○	●	○	●
[31] Staniak and Biecek 2018 — Live and breakDown	git	○	○	●	●	○	●	○
[16] Guidotti et al. 2018 — LORE	git	○	○	●	○	●	●	○
[18] Covert et al. 2020 — SAGE	git	●	○	○	●	○	●	○

(Code) → Not provided: 🔑, Provided no documentation: **git**, Provided with documentation: git; *(Rest of features)* → Not mentioned: ○, Applied: ●

classifier and a saliency mapping utilising stochastic gradient descent. (d) In the category of *ante hoc*, *model-specific*, and *global* XAI methods, we came across BRCG [33], a study focused on learning Boolean rules. These rules are presented in either disjunctive normal form (DNF, OR-of-ANDs, equivalent to decision rule sets) or conjunctive normal form (CNF, AND-of-ORs), serving as an interpretable method for classification. (e) We encountered eight notable works in the category encompassing *post hoc*, *model-specific*, and *local* approaches. The most famous one is Grad-CAM [34], which leverages the gradients associated with a specific target concept, propagating through the final convolutional layer to generate a rough localisation map. This map accentuates significant areas within the image that contribute to predicting the concept. ACD [35] utilises hierarchical clustering optimised to discern clusters of features learned by a Deep Neural Network as predictive. The CEM [36] algorithm identifies minimally and sufficiently present elements required to justify classification and those minimally and necessarily absent. DeepLIFT [37] decomposes a neural network's output prediction for a specific input by backpropagating the contributions of all network neurons to each input feature. LRP [38] explains a classifier's prediction for a given data point by attributing relevance scores to important input components using the model's learned topology. MLAM [39] focuses on identifying and interpreting attractive points in available content, explaining the user's choices. RISE [40] estimates importance empirically by probing the model with randomly masked versions of the input image, obtaining corresponding outputs. Finally, the work by Wang et al. [41] introduces an approximate inference method utilising association rule mining and a randomised search algorithm. In the final category (f), which includes *post hoc*, *model-agnostic*, and *cohort* explanation methods, we discovered only Anchors [42]. Anchors is a systematic method designed to elucidate the behaviour of complex models by establishing high-precision rules known as anchors. These anchors represent local, "sufficient" conditions for prediction.

ProtoryNet [43] is an approach for interpretable text classification based on prototypical learning [44, 45]. In computer vision, part-prototype XAI methods are deep neural networks explainable by design (since they identify key parts of the image and use them to perform both classification and explanation). ProtoryNet [43] is a notable work carrying on the prototype approach in text classification using neural networks. Unfortunately, ProtoryNet doesn't fit the aim of this paper since it is a model-specific and ante hoc approach.

We focus on supervised models, as a significant portion of the literature is dedicated to them [10]. Explanations for these methods provide a rationale behind the output for any particular data point, serving as justifications for the provided predictions [10].

In this work, we focused on *model-agnostic* models. As previously mentioned, they can work with any model family, as they focus on revealing certain properties of the black box model by requiring only input values and predictions [11]. It is worth recalling that a *post hoc* approach is required to make an explainability technique model-agnostic. Concerning the third characteristic considered for models, the examined XAI methods encompass both *local*, *cohort* and *global* aspects. The carefully selected features in our analysis enable us to conduct a comprehensive benchmark study, providing valuable insights for user decision-making across a wide array of applications.

Selected Tools

The evaluation focuses on model-agnostic tools, meaning they should not depend on internal model components like weights or structural information, ensuring applicability to any black box model. This choice is dictated by the fact that these explainers are more generally applicable and make it easier to compare several classification models. Again, for comparability, we discard papers that require handcrafted inputs, such as checklists [28] or input explanations to be validated [30]. Moreover, we consider only tools that have public, updated, and working Python code. Finally, we added transparent machine learning models to the above list that can be used as surrogates, i.e. decision trees (DT), logistic regression (LR), and naive Bayes (NB). Given these criteria, the selection falls on the following methods: LIME[2] [22], SHAP[3] [23], SAGE[4] [18], BRCG[5] [33], Anchors[6] [42], QII[7] [29], DT classifier,[8] LR,[9] and NB.[10] Moreover, we add a rule-based random explainer, which generates random rules, and a random feature importance generator, built similarly as a baseline.

Related Works

This work delves into the direction of evaluating XAI methods and explanations to facilitate *human evaluation*, according to the open challenge outlined in the XAI Manifesto [48]. In preceding literature, several researchers have

[2] https://github.com/marcotcr/lime

[3] https://github.com/slundberg/shap

[4] https://github.com/iancovert/sage

[5] https://github.com/IBM/AIX360

[6] https://github.com/marcotcr/anchor

[7] https://github.com/hovinh/QII

[8] https://scikit-learn.org/stable/modules/tree.html

[9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[10] https://scikit-learn.org/stable/modules/naive_bayes.html

proposed a comparison of XAI methods. Some works [49, 50] focus on global, post hoc, rule-based explainers, comparing the rules generated by different decision trees. Other works evaluate a larger plethora of methods. In [51], the authors propose a scoring system that uses various functional tests from existing research, categorising the tests into four groups: fidelity, fragility, stability, and stress tests. They display results for 13 XAI methods using 11 functional tests. In [52], the authors propose EXPLAN, an algorithm that produces interpretable logical rules, ideal for qualitative analysis of the model's behaviour, comparing it with LIME, LORE, and Anchor. However, EXPLAN is limited to local and cohort methods. A lot of works [4, 53–57] survey and discuss several XAI techniques to understand their capabilities and limitations and to categorise the investigated methods. Unfortunately, many taxonomies for XAI methods of varying levels of detail and depth can be found in the literature. While they often have a different focus, they also exhibit many points of overlap [54]. The works above perform a rigorously structured and theoretically grounded analysis, but none compares the XAI methods to a common dataset.

Finally, in [58], the authors propose a framework to benchmark XAI methods for time series. In all these methods, the comparison lacks differentiation between local and global methods and between rule-based and feature-based explanations, making it challenging. Moreover, none of the previous approaches compares XAI methods both through metrics and user studies, which is crucial for incorporating the user perspective into XAI method evaluation. Therefore, as far as we know, this is the first paper that (1) builds a comprehensive evaluation of different types of XAI methods, comparing local, global, and cohort methods and rule-based and feature-based explanations, also highlighting their differences (2) brings together metrics and user evaluation for a joint comparison and (3) makes the comparison reproducible by providing a repository with the implementation of the benchmarked methods and the metrics.

Evaluation of ML Explanations

The recent proliferation of XAI methods requires to rigorously evaluate their efficacy and interpretability. Previous researches tend to agree that the main distinction is between objective versus human-centred metrics [13, 59]. In contrast, the former is more functionality-oriented and objective, while the latter is more human-centred and subjective. To summarise,

1. **Objective Evaluation (OE)** contains objective metrics and automated approaches to evaluate XAI methods.
2. **Human-Centred Evaluations (HCE)** encompass methods utilising a human-in-the-loop approach, where end-users are engaged, and their feedback or informed judgment is used.

Within this main partition, previous studies have yielded a plethora of metrics and evaluation frameworks tailored to assess the efficacy and quality of XAI explanations. However, the vast majority focus either on the objective evaluation or on the human-centred one. In [10], On the other hand, the authors propose a framework that groups 34 metrics into 5 dimensions: (1) functional requirements, ensuring the method's core capabilities are met; (2) operational requirements, detailing practical implementation needs; (3) usability criteria, evaluating user experience and effectiveness; (4) security and privacy considerations, identifying potential vulnerabilities; and (5) validation methods, confirming the method's reliability through testing. These dimensions provide a comprehensive framework for evaluating and comparing explainability methods, ensuring thorough understanding and standardisation in the field. We selected this paper as a reference because the proposed framework offers a comprehensive yet synthetic comparison of capabilities and limitations of XAI methods that (1) covers both the objective and human-centred evaluation and (2), in addition to the evaluation, paves the way for framing the methods chosen through their characteristics.

Among the 5 dimensions proposed by [10], four refer to objective evaluations and one to human-centred ones. Below, we describe the five dimensions, specifying if they belong to OE or HCE:

1. **Functional Requirements — OE** include the algorithmic requirements, e.g. the problem type (regression, classification, or clustering), the explanation scope (global, local, or cohort), the explainer's computational complexity, etc.
2. **Operational Requirements — OE** focus on the user and explainer interaction, e.g. the explanatory medium (summarisation, visualisation, etc.), the trade-off between performances and explainability, and the type of interaction with the system (static or dynamic).
3. **Usability Requirements — OE** are objective metrics centred on the user's perspective. They focus on making the explanation more natural and easily comprehensible. Some examples are the soundness, completeness, and interactiveness of the explanation.
4. **Safety Requirements — OE** cover the impact of XAI systems on the robustness, security, and privacy aspects of the underlying predictive models.
5. **Validation Requirements — HCE** encompass user studies and synthetic experiments. Because XAI aims to make algorithmic decisions more comprehensible to humans, their final efficacy needs to be evaluated by users.

The remainder of this section will assess the chosen XAI methods through objective and human-centred evaluations. Specifically, the "Objective Evaluation" section is dedicated to objective evaluation. The metrics utilised will be examined in the "Objective Metrics" section, while their implementation and experimental results will be shown in the "Objective Evaluation" section. In the "Human Evaluation" section, we will introduce the human-centred evaluation, defining its measures and practices in the "Evaluation Design and Experiments" section and presenting its results in the "Human Evaluation Results" section.

Objective Evaluation

This section will delve into the four dimensions from [10] that refer to OE. The authors present 34 criteria, of which 9 belong to *functional requirements* (denoted with codes from F1 to F9), 10 to *operational ones* (O1–O10), 11 to *usability* (U1–U11) and 4 to *safety requirements* (S1–S4).

Of these 34 criteria, some are characteristics or desiderata of the explainer, while others are evaluation metrics. For instance, F1, the problem supervision level, is a characteristic of the XAI method, expressing whether it works with unsupervised, supervised, or semisupervised ML algorithms. On the contrary, U1 is a metric that measures the Soundness of the XAI methods with respect to the prediction of the underlying ML model. In the "Characteristics of the Selected Methods" section, we present all the characteristics of the selected XAI methods, while in the "Objective Metrics" section, we present the metrics that are used to evaluate the selected XAI methods, presenting the results of their measurement in the "Objective Evaluation" section. Both characteristics and metrics are taken from [10] and abbreviated as in the paper, e.g. the Functional Requirement 1 — Problem Supervision Level, which is abbreviated as F1.

Characteristics of the Selected Methods

In this article, we have selected several methods for comparison based on specific characteristics. This section discusses the selected characteristics according to [10] formalisation.

Functional Requirements In XAI, the vast majority of the literature is about **supervised learning** (F1 — Problem Supervision Level), in the context of **classification** (F2 — Problem Type) where explanations serve as a justification of **predictions** (F3 — Explanation Target) [10], and this will also be the approach of this research. Regarding the Explanations Scope (F4), we will consider explanations at all levels: **local, global, and cohort**. For the sake of comparability, and because of their greater adoption, we will test only **model agnostic** (F6 — Applicable Model Class), **post hoc** (F7 — Relation to the predictive system) explainers. Regarding the Compatible Feature Types (F8), in this article, we focus on **textual data**.

Operational Requirements The family of explanations (O1) that we target is the **associations between antecedent and consequent**, while counterfactual and contrastive explanations are evaluated according to different methodologies, paradigms, and measures [60, 61]. Regarding the explanatory medium (O2) and the system interaction (O3), all the explainers tested present **statistic summarisation** and **static interaction**, respectively. Researchers have found that in current XAI methods, the presentation layer is usually distinctly delineated and less curated than the core algorithm [6]. Some works use relevant word highlighting as a visualisation technique (e.g. in [22]), and the present work goes in that direction. The explanation domain (O4) in this work focuses on text classification. Considering the transparency of the data model (O5), we opt to concentrate on post hoc, model-agnostic XAI methods, enabling the utilisation of **any opaque underlying model**. Concerning the explanation audience (O6) and the purpose of the explanation (O7), the user study includes a **broad audience, both experts in XAI and non-experts**. The study encompasses **various functions, explained in the** "Evaluation Design and Experiments" section, such as, but not limited to, understandability, trust, and satisfaction. All the explanations provided by the considered methods are of **causal nature** (O8). The rest of the paper discusses the requirements of trust vs performance (O9) and provenance (O10).

Usability Requirements This category includes five metrics, discussed in the "Objective Evaluation" and "Objective Evaluation" sections, respectively: U1, U2, U3, U9, and U11. Moreover, since neither system provides interactive nor actionable outputs, as specified in Par. *Operational Requirements*, neither U4 nor U5 is discussed. Chronology (U6), coherence (U7), novelty (U8), and personalisation (U10) assume previous knowledge and expectations regarding the output of the system and its interaction with the user, which in turn implicates a continuous use of the system over time, which is not this case.

Safety Requirements Of the explanation requirements, S3 will be discussed and measured in the "Objective Evaluation" and "Objective Evaluation" sections, respectively. The other safety requirements (S1, S2, and S4) are very specific to the application domain where XAI is used. Therefore, they are not part of the scope of this paper.

Objective Metrics

Below, we present the metrics used for the objective evaluation. The metrics used differ depending on the type of explanation, e.g. local vs. global explanations and rules vs. feature-returning explanations. However, all metrics adopted are viable for the classification task.

Computational Complexity (F5) and Caveats (F9) The choice of an XAI method should consider its time, memory, and computational complexity constraints. In text classification, the number of features is typically very high. Not all the XAI methods presented scale well beyond certain amount of features regarding computational times and memory requirements. In the "Objective Evaluation" section, we will present the results of four different runs, in which we consider respectively the 10, 100, 1000 and 10,000 most common features of the corpus. In this way, we can identify the explainers' computational limits. From the perspective of memory, we consider only algorithms that do not exceed the 64GB of RAM requirement, which is the RAM size on which we are conducting experiments, while from the perspective of computational times, an algorithm will be considered intractable if it takes longer than one day for global explainers and more than 3 h for a single explanation of local ones.

Explanation Fidelity Measures: Soundness (U1) and Completeness (U2) Those two dimensions measure how well the explainer agrees with the underlying model developed by the classifier. **Soundness** is usually measured for global surrogate models through *fidelity* [11], i.e. the concordance S of the predictions of the XAI method w; taken from a set of possible white box models I approximated on the training data $X = \{x_i, \ldots, x_n\}$; with the predictions of the underlying black box one b, as in Eq. (1).

$$\arg\max_{w \in I} \frac{1}{|X|} \sum_{x \in X} S(w(x), b(x)) \tag{1}$$

On the other hand, **completeness** assesses the extent to which an explanation can generalise. It can be evaluated by verifying the accuracy of an explanation across comparable data points (individuals) within various groups across a dataset [10]. For rule-based explainers, it can be measured with their *correctness*, i.e. the number of correctly predicted instances explained by the output rules r over total instances X [50], following Eq. (2).

$$\frac{r}{X} \tag{2}$$

For feature-based ones, we refer to the measure of *faithfulness* [62, 63] shown in Eq. (3). For an explainer to be faithful, the important features of the model should correspond to the important ones of the explainer. It is measured by perturbing the explainer's features. For an explainer to be faithful, given a subset size $|S|$, the change in the predictor b's output between the perturbed explanation and the unchanged one should be proportional to the sum of attribution scores. The proportionality is computed using Pearson correlation.

$$\operatorname*{corr}_{S \subseteq \binom{|d|}{|S|}} \left(\sum_{i \in S} w(x)_i, b(x) - b(x_{x_s = x_i}) \right) \tag{3}$$

Contextfulness (U3) It is important to frame the single explanation in a context for cohort explanations. The context can be used in several ways, e.g. to check for safe generalisation. This measure applies only to rule-based explanations. Each instance is classified by a rule to which a class is associated. To measure the contexfulness, we select the widely known rule **coverage** metric [49] shown in Eq. (4), computing the ratio of covered input instances c over total input instances X.

$$\frac{c}{X} \tag{4}$$

Parsimony (U11) Explanations should be selective and concise to prevent users from being overwhelmed with unnecessary details. In other words, parsimonious methods should aim to address the most significant (explanation) gaps using the fewest arguments possible. For rule-based explainers, the **selectiveness** of rules is measured through their features fraction overlap [59], the degree of overlap between every pair of rules r_i, r_k in a ruleset R, according to Eq. (5). **Conciseness** is measured by the ruleset's cardinality $|R|$ and the rules' average length.

$$\frac{2}{R(R-1)} \sum_{i,j:i \leq j} \frac{\text{overlap}(r_i, r_j)}{X} \tag{5}$$

For feature-based explainers, a complex explanation is one in which all the features have equal attribution, while the simplest explanation would be concentrated on one feature [10]. Consequently, in [63], the authors measure the complexity of an explanation as the entropy of its features attributions. Using their metric, in Eq. (6) we measure the parsimony of feature base explainers as 1 minus the complexity, where $P_w(i)$ is the fractional contribution of feature x_i to the total magnitude of the attribution.

$$1 - \sum_{i=1}^{d} P_w(i) \ln(P_w(i)) \tag{6}$$

Explanation Invariance (S3) The ideal explainer should represent the underlying model and its changes in behaviour without introducing variability of their own. For this reason, explanations must be:

Consistent, i.e. given a fixed ML model, explanations of similar data points should be similar. If we define the sensitivity [62] as the variation of the features/rules of the explanation function concerning a change in the input, we can measure the consistency as 1-sensitivity. Given the black box model b, the explanation function w, the distance metric D, an instance x and its perturbation z, we follow Eq. (7) [63].

$$1 - \max D(w(b, x), w(b, z)) \tag{7}$$

In the case of textual data, variations are usually of small magnitude and typically do not alter the sentence's meaning, such as introducing typos and substituting some words with synonyms. The changes were made using the nlpaug library [64]. **Stable**, i.e. different runs of the same XAI method, should provide the same output. Like consistency, stability will be measured using a binary variable equal to 1 if the rules generated after a rerun with a different randomisation seed are the same.

Objective Evaluation

Dataset and Preprocessing The experiments were conducted on the International Movie Reviews dataset (IMDB) [65]. The dataset contains 50,000 movie reviews collected from IMDB with relative ratings and consists of an even number of positive and negative reviews. The IMDB dataset is widely known and frequently used in research, [66]. It is a large dataset containing several reviews on diverse movies and is well suited for text classification.

Following previous literature, the authors considered a negative review with a score ≤ 4 out of 10 and a positive review with a score ≥ 7 out of 10. Therefore, only highly polarised reviews are considered. For comparability, the preprocessing has been kept as simple as possible. The dataset undergoes standard preprocessing transformations for feature normalisation and noise reduction: converting all the words to lowercase and stopwords [67] and punctuation removal.

Training of the Underlying Classification Model As in the case of preprocessing, the process was made as simple and repeatable as possible. Three classic yet powerful ML models were chosen: Random Forest Classifier (RF), Gradient Boosting Classifier (GB), and Support Vector Classifier (SVC) with a linear kernel. Several factors drove the selection.

- The paper results can be easily reproduced not using special hardware (e.g. GPU)
- Although Deep Learning models perform better on text classification w.r.t. Non-Deep Learning models [68], the difference is not so big and other factors might have a more significant impact, e.g. according to [69] text preprocessing (e.g. slang and abbreviation replacement, repeated punctuation removal, …) and simple classification methods can achieve state-of-the-art results, sometimes outperforming complex and recent pre-trained architectures (i.e. Transformer-based models)
- According to [70], developing a text classifier is a trial-and-error process. Therefore, Non-Deep Learning models can be an effective solution in the early stages of the process thanks to the reduced training time and the low computational effort required

The chosen models (i.e. RF, GB, and SVC) are popular algorithms in text classification [68]. The models used are implemented using scikit-learn [71] with default parameters, splitting the IMDB dataset into 80% training instances and 20% (10,000) testing instances. The dataset is transformed via Bag Of Word (BOW), and only the 10,000 most common features are kept to have a less sparse matrix devoid of spurious features. The classification results are reported in Table 2. Given its highest Accuracy and F1 values, we will test the XAI methods using the predictions and model weights of the SVC classifier.

Objective Evaluation Results In Table 3, we present the results of the experiments for the rule-based explainers while the feature-based ones are presented in Table 4. The execution time is reported in average seconds. Following [63], we computed all the measures as the average on a sample of 50 instances for the cohort and the local explainers. In the experimentation phase, each dimension was evaluated across varying BOW feature counts: 10, 100, 1000, and 10000. However, BRCG and QII measurements were unattainable for 10,000 features due to excessively long computational times, and SAGE for 1000 and 10,000 features. Similarly, the required RAM exceeded 64GB for Shap, rendering the evaluation unfeasible. Another caveat concerns the fidelity of SAGE, which cannot be computed since it does not implement predictive functions but returns features' importance based on how much predictive power they contribute.

The two tables show high consistency and sensitivity for rule-and feature-based explainers, except (as expected) for

Table 2 Performance of classifiers

Classifier	Accuracy	Precision	Recall	F1
Random Forest	0.822	0.827	0.818	0.822
SVC	0.863	0.857	0.875	0.866
Gradient Boosting	0.800	0.776	0.849	0.811

Table 3 Objective evaluation results for global and cohort rule-based explainers

		Metric Category	**Completeness** U2	**Contextfulness** U3		**Parsimony** U11			**Soundness** U1	**Stability** S3	**Consistency** S3
		Metric	Avg Correct Rule	Rule Coverage	Rule Overlap	Avg Rule Length	Time	Number Rules	Fidelity	Change Seed	1-Sensitivity
Scope	Explainer	N Features									
Any	Random Rule	10	-	0.26	0.09	-	0.00	100	0.53	0.00	0.33
		100	-	0.97	0.09	-	0.00	100	0.51	0.00	0.00
		1000	-	0.99	0.08	-	0.00	100	0.51	0.00	0.00
		10,000	-	0.99	0.07	-	0.00	100	0.50	0.00	0.00
Global	Decision Tree[a]	10	-	0.27	0.01	4.83	0.06	20.00	0.88	1.00	1.00
		100	-	0.29	0.00	8.33	0.23	20.00	0.79	1.00	1.00
		1000	-	0.48	0.00	8.33	0.78	20.00	0.78	1.00	1.00
		10,000	-	0.48	0.00	8.33	1.47	20.00	0.76	1.00	1.00
	BRCG [33]	10	-	0.67	0.00	1.00	0.10	1.00	0.55	1.00	1.00
		100	-	0.63	0.02	1.5	2.25	2.00	0.63	1.00	1.00
		1000	-	0.76	0.00	1.5	72.3	2.00	0.62	1.00	1.00
Cohort	Anchors [42]	10	0.54	0.32	-	1.58	0.04	-	-	1.00	1.00
		100	0.61	0.23	-	1.8	0.03	-	-	1.00	1.00
		1000	0.63	0.11	-	1.26	0.02	-	-	1.00	1.00
		10,000	0.65	0.07	-	1.40	0.03	-	-	1.00	1.00

[a]https://scikit-learn.org/stable/modules/tree.html

Table 4 Objective evaluation results for global and local feature-based explainers

		Metric Category Metric	**Completeness** U2 Faithfulness Correlation	**Parsimony** U11 Complexity	Execution Time (s)	**Soundness** U1 Fidelity	**Stability** S3 Change Seed	**Consistency** S3 1-Sensitivity
Scope	Explainer	N Features						
Any	Random Feature Importance	10	−0.82	0.98	0.00	0.53	0.29	0.67
		100	0.11	0.97	0.00	0.51	0.46	0.73
		1000	0.01	0.88	0.00	0.51	0.46	0.85
		10,000	0.02	0.98	0.00	0.50	0.29	0.67
Global	Logistic Regression[a]	10	-	0.08	0.05	1.00	1.00	1.00
		100	-	0.03	0.09	0.99	1.00	1.00
		1000	-	0.01	0.44	0.99	1.00	1.00
		10,000	-	0.00	1.30	0.99	1.00	1.00
	Multinomial NB[b]	10	-	0.00	0.01	0.86	1.00	1.00
		100	-	0.00	0.01	0.90	1.00	1.00
		1000	-	0.00	0.01	0.90	1.00	1.00
		10,000	-	0.00	0.02	0.87	1.00	1.00
	Sage [18]	10	-	0.19	3.3	-	1.00	1.00
		100	-	0.05	8.37	-	1.00	1.00
Local	Lime [22]	10	−0.02	0.99	1.06	-	1.00	1.00
		100	0.94	0.09	1.07	-	1.00	1.00
		1000	0.91	0.12	1.08	-	1.00	1.00
		10,000	0.92	0.10	1.14	-	1.00	1.00
	Shap [23]	10	-	0.11	0.02	-	0.98	1.00
		100	−0.10	0.01	0.21	-	1.00	1.00
		1000	0.78	0.06	40.18	-	1.00	1.00
	QII [29]	10	-	0.23	0.26	-	0.99	0.99
		100	−0.03	0.03	0.34	-	0.99	1.00
		1000	0.95	0.06	3.27	-	0.99	1.00

[a] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[b] https://scikit-learn.org/stable/modules/naive_bayes.html

the random-generated ones. This confirms two facts: (1) explainers do not create variability in the results, and (2) random explainers constitute an effective control method. This is confirmed by the accuracy values, which consistently linger around 0.5 for rule- and feature-based methods.

The DT surrogate exhibits the highest soundness among the global rule-based explainers. The number of rules generated by DT is a parameter chosen by design, and we used the default one of 20. The BRCG instead uses a smaller number of rules. Moreover, BRCG uses shorter rules with lower fidelity values but greater coverage. For instance, with 1000 features, the DT has a fidelity of 0.76 and coverage of 48% of the dataset with an average rule length of 8.33, while the BRCG has a fidelity of 0.62 on 76% of the dataset with an average rule length of 1.5 words. The execution time of the DT is lower, with the BRCG explainer being even intractable for 10,000 features. Anchors also exhibit short execution times, typically 1 to 5 s, even when dealing with 10,000 features. Being a cohort explainer, we measure the average rule coverage of a sample of 50 instances instead of the coverage on the whole dataset. Increasing the number of features, the coverage decreases while the average rule correctness on 50 rules increases. A potential interpretation of this could be that, by using more features, Anchors generate more precise rules with less common words, which consequently have lower coverage over the data.

In Table 4 we can observe the results for feature-based explainers. In this case, the fidelity of the surrogate logistic regression is close to one, whatever the number of features, and is higher than that of the Naive Bayes classifier. However, the entropy of the weight distributions of the explanation features in Naive Bayes is consistently less than 1%, rendering it simpler and more interpretable compared to logistic regression, albeit with lower fidelity. Local explainers exhibit high faithfulness only when employing many features, which is intuitive given their local nature. Indeed, despite their frequency, it is challenging for a small subset of features to manifest within the selected evaluation phrases consistently. As discussed above, due to time and memory issues, Shap becomes intractable after 1000 features. Usually, text classification involves more than this number of features.

Human Evaluation

The FOUSV (Functional, Operational, Usability, Safety, Validation) evaluation framework used above and summarised by Sokol and Flach [10] is based on requirements that can be quantitatively formulated, therefore, this framework does not fit user-centred metrics. Another recently proposed user-centred human evaluation framework [12] considers the "user types" and "design goals" (i.e. AI Novices, Data Experts, or AI Experts will use the explanations with different purposes, hence having different requirements). According to this 5Ms framework, there are also five dimensions to measure for human evaluation:

- M1: Mental Model
- M2: Usefulness and Satisfaction
- M3: User Trust and Reliance
- M4: Human-AI Task Performance
- M5: Computational Measures

The mental model (M1) evaluates how helpful the explanations are in conceptualising and understanding the mechanism of the target ML model. Therefore, they are very context-specific and difficult to compare across methods. The M2 and M3 dimensions are usually self-reported, and previous studies have used interviews, questionnaires, and case studies as subjective measures [12]. M4 has overlaps with functionality metrics. M5 is less commonly implemented according to a survey covering 42 recent papers [12] and overlaps a lot with objective metrics. Therefore, we concluded that only M2 and M3 are relevant in our study. In the context of text classification, the human evaluation will measure (1) Usefulness, (2) Satisfaction, and (3) Trustworthiness. We dropped "Reliance" because those text classification systems usually achieve high accuracy, hence they are considered very reliable already.

A reliable metric for Usefulness is the response time. In the translation application example, one complaint is that the explanations for recommended translations extend the time needed for the translator to complete the task. Given that processing the explanations always takes time, useful explanations only minimally increase or reduce the response time to the same task. So, the Usefulness question is designed to ask the user to classify a piece of text into given categories, with/without a hint (system-generated explanations).

To be consistent with previous studies, Satisfaction and Trustworthiness are measured by self-report:

- Rate from scale 1(worst)-5(best), how satisfied the user is with a given explanation?
- Compare explanations for the given classification task and choose the more trustworthy one that leads to the classification outcome.

Evaluation Design and Experiments

To cover the different XAI methods with a resource constraint, we sample four representative methods, i.e. LIME for local features [22], Logistic Regression for global features,[11] Decision Tree for global rules [72], and Anchors [42]

[11] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

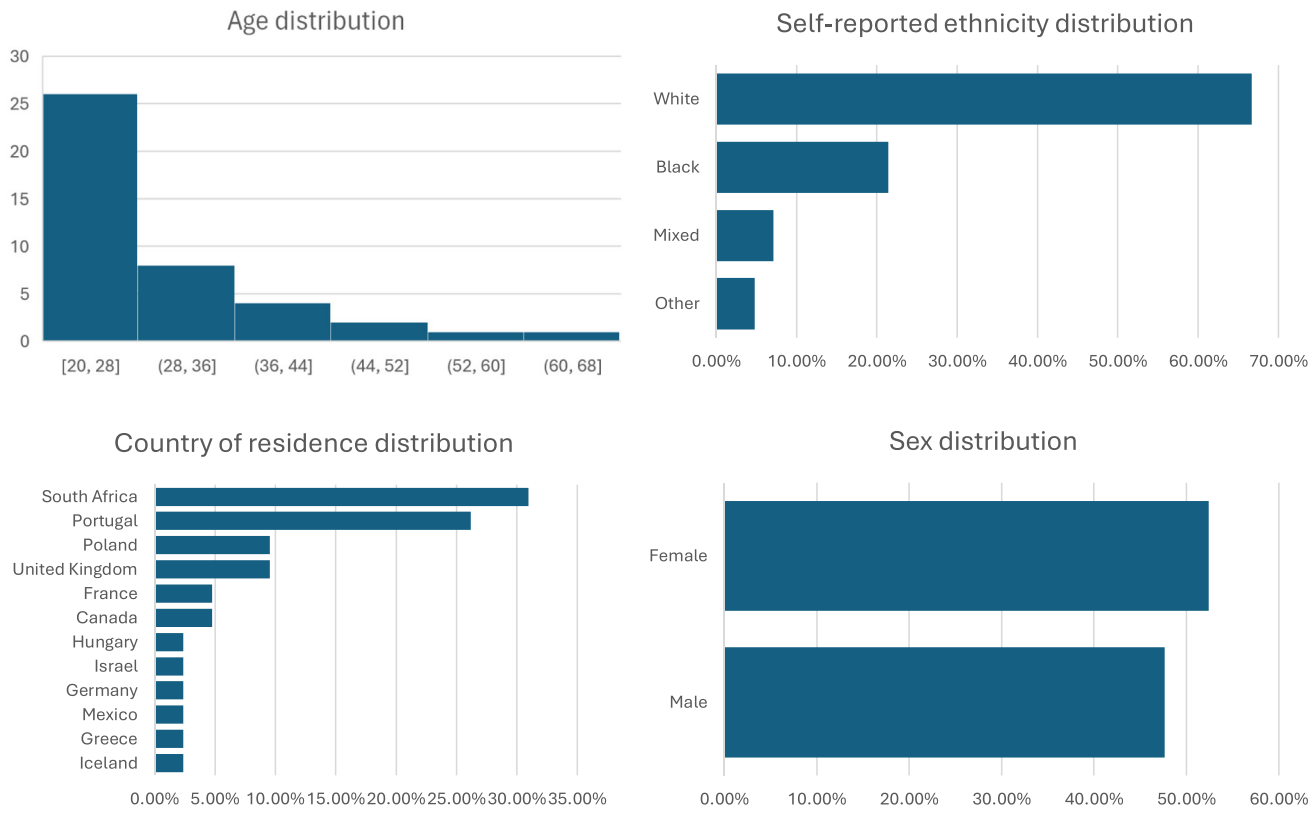


Fig. 2 Demographic information of the user study participants

for cohort rules. Each user is presented with 8 questions randomly selected from the explanation no-explanation pairs and 2 additional questions for satisfaction and trustworthiness. This ensures the interdependence of response times and ratings for the same text samples. An example of the survey questions can be found in the Appendix.

We recruited 42 users in two batches from Prolific[12] using the LimeSurvey[13] online tool. The participants (paid approx. £ 9 per hour and took 13 min to finish the survey in average) are filtered using two criteria: fluent in English, and have completed secondary education or above.

Figure 2 provides more information on the participants' demographics. It can be observed that the sample is diverse across gender, ethnicity, and country. Because the result distributions, e.g. in Fig. 3 are continuous, we believe our findings have good generalisability across these demographic features. Further, we noticed that the participants are skewed to young people under 40: that is probably also the demographics of online gig workers. The geographic distribution also mainly covers Eurafrica and overlooks Asia, probably due to the language criterion. Though we believe these potential biases are unlikely to have significant impacts on user perception of explanations, it may be more confident to limit our user study findings to relatively young English-speaking people.

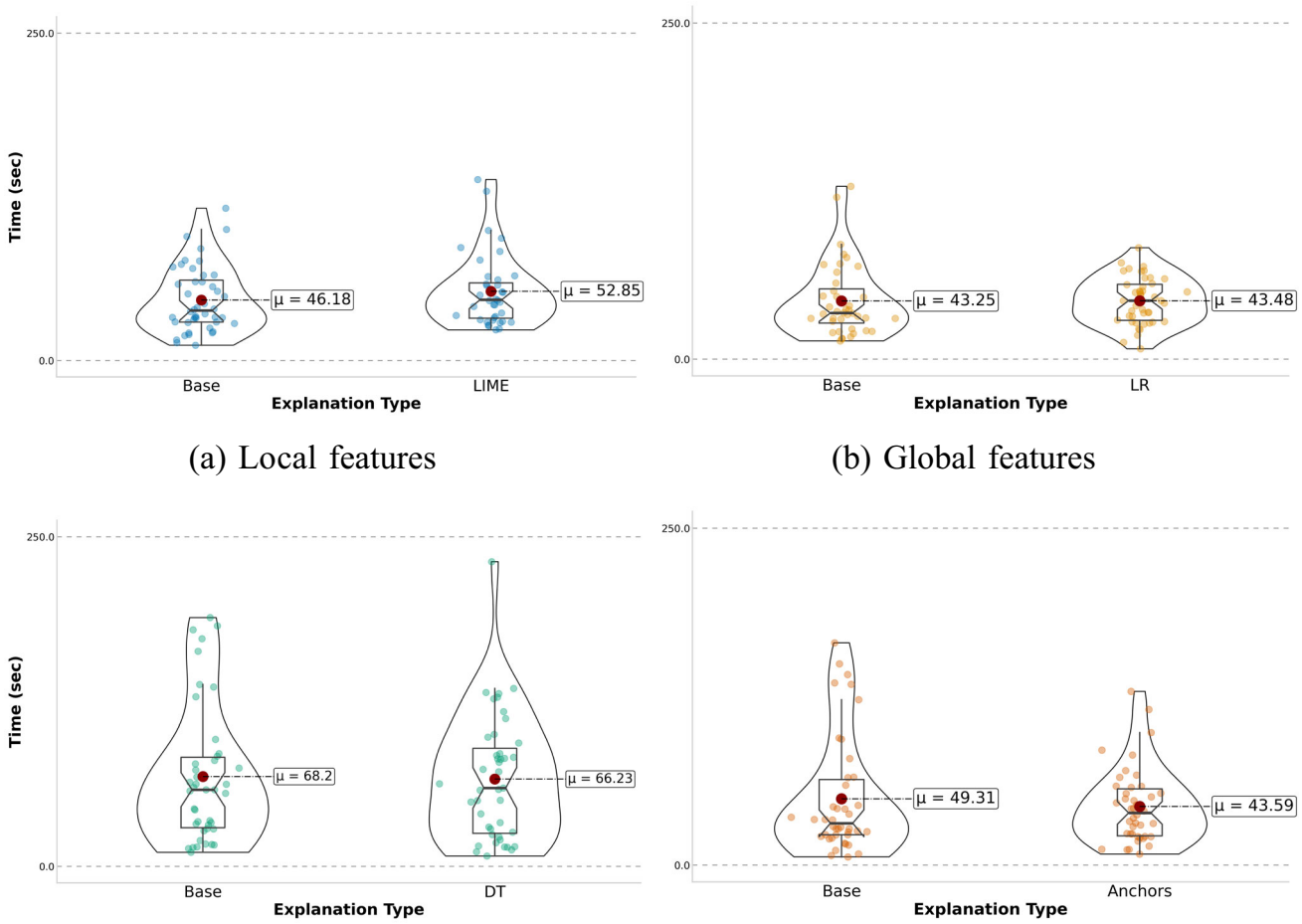
Human Evaluation Results

Figure 3 shows the violin plot [73] of users' response time with/without different explanations. An interesting observation across all explanations is that their presence makes the response time distributions more concentrated. The feature-based explanations (LIME and LR) generally increase the quantiles of response times. Local features seem to be more useful with the 3-quantile decreased. A median decrease is also observed for LIME before averaging. The quantile decreases for rule-based explanations are more pronounced. We conclude that Anchors produce the most useful explanations among the evaluated methods. The general trends are that rule-based explanations are more useful than feature-based ones, and localising explanations increases their usefulness.

Regarding satisfaction ratings, users are mostly satisfied with the LR (s = 3.04) and DT (s = 2.96) explanations. At the same time, they are less satisfied with the explanations for the Anchors (s = 2.62) and LIME (s = 2.55), although the explanations for the Anchors were the most useful. Users

[12] https://app.prolific.com/

[13] https://www.limesurvey.org/



(a) Local features (b) Global features

(c) Global rules (d) Cohort rules

Fig. 3 Average response time (seconds) with/without the explanations

seem to be more satisfied with global explanations than local ones. This may be because global explanations are usually well-aligned with common sense, while local ones, despite their discriminative power, feel unnatural with less frequent features/rules.

The trustworthiness choices reveal that out of 39 valid responses, the LR explanation is the most trustworthy (n = 27) and is significantly better than DT (n = 9), LIME (n = 2) and Anchors (n = 1) explanations. Further investigation into the user responses suggests the importance of "semantic correctness" of the explanations for trustworthiness. Even those who chose DT mentioned that the words are "really related to the movie" or the narrative "sounds like a review". On this factor, global explanations usually do a better job of discovering semantically relevant words.

The human evaluation shows that the results can be quite diverse for different XAI methods, even for subjective metrics like Usefulness, Satisfaction, and Trustworthiness. The rule-format and localised explanations are useful for assisting human text classification tasks, while the global explanations are more satisfying and trustworthy.

Discussion

The objective evaluation of XAI methods highlights nuanced characteristics across different global explainers, as depicted in Table 3. At a global level, some methods consistently demonstrate high fidelity, irrespective of the number of features. On the other hand, their simplicity and interpretability are underscored by the consistently low entropy of its weight distributions. On the other hand, local explainers exhibit heightened faithfulness primarily with a substantial feature set, aligning with their inherent nature of capturing localised patterns. Notably, the computational constraints become apparent with several methods, rendering it intractable with many features, a common occurrence in text classification scenarios where feature counts tend to be high.

Transitioning to the human evaluation, the user study sheds light on complementary insights. Anchors emerge as the most useful explanation method, with rule-based approaches generally deemed more valuable than feature-based ones, particularly in enhancing response time distributions. Users express higher satisfaction with global explanations, potentially attributed to their alignment with common understanding. In contrast, despite their discriminative power, local explanations are perceived as less natural due to incorporating less frequent features or rules. The trustworthiness ratings underscore the significance of semantic correctness in explanations, with global explanations often excelling in uncovering semantically relevant words.

In synthesis, while the objective evaluation provides crucial insights into the technical capabilities of XAI methods, the human evaluation underscores the subjective nuances in user perception, emphasising the importance of considering technical efficacy and user satisfaction in selecting the appropriate explainers. The absence of a singular silver bullet in XAI necessitates a nuanced approach, where the choice of explainer should be tailored to the specific use case and user requirements.

Conclusions and Future Contributions

In this study, we conducted a comprehensive survey of XAI methods. Then, we evaluated different levels of granularity, i.e. their scope and output. In the evaluation, we employed both objective metrics and user evaluations and discussed the results in depth. Our findings underscore the complexity of choosing an appropriate XAI method, as there is no one-size-fits-all solution. Instead, the selection process should be approached on a case-by-case basis, considering the specific context and requirements of the task at hand. By thoroughly analysing the strengths and limitations of different explainers, this paper is a valuable resource for users navigating the landscape of XAI methods, aiding their decision-making process and fostering a deeper understanding of the trade-offs involved. Future research will explore papers focusing on XAI for embeddings, which also hold a particular significance in text classification.

Appendix

In Table 5, we present all acronyms used throughout the paper.

Below is the survey template presented to participants; the response time is recorded for all the questions, especially Q1–Q8.

Table 5 Full list of acronyms employed in this manuscript

Concept	Acronym
Machine Learning	ML
eXplainable Artificial Intelligence	XAI
International Movie Database	IMDB
Random Forest Classifier	RF
Gradient Boosting Classifier	GB
Support Vector Classifier	SVC
Bag Of Words	BOW
Shapley Additive exPlanations	SHAP
Local Interpretable Model-agnostic Explanations	LIME
SHAP Additive exPlanations	SAGE
Bayesian Rule Change Generators	BRCG
Quantitative Input Influence	QII
Generalised Low Rank Models	GLRM
Objective Evaluation	OE
Human-Centred Evaluations	HCE
Usability requirements indices	U1–U11
Soundness	U1
Completeness	U2
Contextfullness	U3
Interactiveness	U4
Actionability	U5
Chronology	U6
Coherence	U7
Novelty	U8
Complexity	U9
Personalisation	U10
Parsimony	U11
Functional requirements indices	F1–F9
Problem Supervision Level	F1
Problem Type	F2
Explanation Target	F3
Explanations Scope	F4
Computational Complexity	F5
Applicable Model Class	F6
Relation to the Predictive System	F7
Compatible Feature Types	F8
Caveats and Assumptions	F9
Operational requirements indices	O1–O10
Explanation Family	O1
Explanatory Medium	O2
System Interaction	O3
Explanation Domain	O4
Data and Model Transparency	O5
Explanation Audience	O6
Function of the Explanation	O7
Causality vs. Actionability	O8

Table 5 continued

Concept	Acronym
Trust vs. Performance	O9
Provenance	O10
Safety requirements indices	S1–S4
Information Leakage	S1
Explanation Misuse	S2
Explanation Invariance	S3
Explanation Quality	S4
Functional, Operational Usability, Safety, Validation	FOUSV

Q1: Read the following description,
[text sample 1]
[no LIME explanation]
Do you think of the comment as a positive or negative one?

Q2: Read the following description,
[text sample 2]
[with LIME explanations]
Do you think of the comment as a positive or negative one?

Q3: Read the following description,
[text sample 3]
[no LR explanation]
Do you think of the comment as a positive or negative one?

Q4: Read the following description,
[text sample 4]
[consider LR explanation words]
Do you think of the comment as a positive or negative one?

Q5: Read the following description,
[text sample 5]
[no DT explanation]
Do you think of the comment as a positive or negative one?

Q6: Read the following description,
[text sample 6]
[consider DT rules: if..., choose positive/negative]
Do you think of the comment as a positive or negative one?

Q7: Read the following description,
[text sample 7]
[no Anchors explanation]
Do you think of the comment as a positive or negative one?

Q8: Read the following description,
[text sample 8]
[consider Anchors explanation words]
Do you think of the comment as a positive or negative one?

Q9: You are told the following description is positive/negative,
[text sample 9]
because: /shuffled
[LIME explanation]
[LR explanation]
[DT explanation]
[Anchors explanation]
Rate the above four explanations from a scale of 1–5, how are you satisfied with them.

Q10: You are told the following description is positive/negative,
[text sample 10]
Choose the explanation that you trust most, and briefly explain why do you trust it?
/shuffled
[LIME explanation]
[LR explanation]
[DT explanation]
[Anchors explanation]

Author Contributions Lorenzo Malandri performed the conceptualisation, supervised the project; Mirko Cesarini performed the conceptualisation, supervised the project; Frank Xing defined the methodology, performed the formal analysis; Andrea Seveso defined the methodology, performed the formal analysis, performed the experimental part; Filippo Pallucchini defined the methodology, performed the formal analysis, performed the experimental part, validated the results. All authors were engaged in writing and reviewing the manuscript.

Funding Open access funding provided by Universitá degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. No funding was obtained for this study.

Data Availability The datasets generated and analysed during the current study and the code used to perform the evaluation will be available in a public GitHub (https://github.com/Crisp-Unimib/XAI_Benchmark) repository, and the URL will be provided upon acceptance.

Declarations

Research Involving Human Participants The user study conducted in this paper adhered to ethical guidelines by utilising the Prolific platform, which ensures voluntary participation, informed consent, anonymity, confidentiality, and transparent payment information. All participants were informed about the nature of the study, their rights, and the han-

dling of their data, ensuring their welfare and privacy throughout the research process.

Conflict of interest The authors declare no Conflict of interest.

References

1. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI—explainable artificial intelligence. Sci Robot. 2019;4(37):7120.
2. Gozzi N, Malandri L, Mercorio F, Pedrocchi A. XAI for myo-controlled prosthesis: Explaining EMG data for hand gesture classification. Knowl Based Syst. 2022;240:108053.
3. Xing F, Malandri L, Zhang Y, Cambria E. Financial sentiment analysis: An investigation into common mistakes and silver bullets. In: Proceedings of the 28th International Conference on Computational Linguistics. 2020. pp. 978–87.
4. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, Scardapane S, Spinelli I, Mahmud M, Hussain A. Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput. 2024;16(1):45–74.
5. Malandri L, Mercorio F, Mezzanzanica M, Seveso A. Model-contrastive explanations through symbolic reasoning. Decis Support Syst. 2024;176:114040.
6. Cambria E, Malandri L, Mercorio F, Mezzanzanica M, Nobani N. A survey on XAI and natural language explanations. Inf Process Manag. 2023;60(1):103111.
7. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82–115.
8. Li Q, Peng H, Li J, Xia C, Yang R, Sun L, Yu PS, He L. A survey on text classification: From traditional to deep learning. ACM Trans Intell Syst Technol (TIST). 2022;13(2):1–41.
9. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning-based text classification: a comprehensive review. ACM Comput Surv (CSUR). 2021;54(3):1–40.
10. Sokol K, Flach P. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020. pp. 56–67.
11. Burkart N, Huber MF. A survey on the explainability of supervised machine learning. J Artif Intell Res. 2021;70:245–317.
12. Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. ACM Trans Interact Intell Syst. 2021;11(3–4):1–45. https://doi.org/10.1145/3387166.
13. Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics. 2021;10(5):593.
14. Du K, Xing F, Cambria E. Incorporating multiple knowledge sources for targeted aspect-based financial sentiment analysis. ACM Trans Manag Inf Syst. 2023;14(3):23.
15. Keele S, et al. Guidelines for performing systematic literature reviews in software engineering. Technical Report, ver. 2.3 ebse technical report. ebse. 2007.
16. Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F. Local rule-based explanations of black box decision systems. arXiv:1805.10820 [Preprint]. 2018. Available from: http://arxiv.org/abs/1805.10820.
17. Craven M, Shavlik J. Extracting tree-structured representations of trained networks. Adv Neural Inf Process Syst. 1995;8.
18. Covert I, Lundberg SM, Lee S-I. Understanding global feature contributions with additive importance measures. Adv Neural Inf Process Syst. 2020;33:17212–23.
19. Dhurandhar A, Shanmugam K, Luss R, Olsen PA. Improving simple models with confidence profiles. Adv Neural Inf Process Syst. 2018;31.
20. Wei D, Dash S, Gao T, Gunluk O. Generalized linear rule models. In: International Conference on Machine Learning. PMLR; 2019. pp. 6687–96.
21. Sushil M, Šuster S, Daelemans W. Rule induction for global explanation of trained models. In: Analyzing and Interpreting Neural Networks for NLP (BlackBoxNLP), Workshop at EMNLP. 2018. pp. 82–97.
22. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. pp. 1135–44.
23. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
24. van der Waa J, Robeer M, van Diggelen J, Brinkhuis M. Neerincx M. Contrastive explanations with local foil trees. In: Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden. 2018. p. 37.
25. Elenberg E, Dimakis AG, Feldman M, Karbasi A. Streaming weak submodularity: Interpreting neural networks on the fly. Adv Neural Inf Process Syst. 2017;30.
26. Lei T, Barzilay R, Jaakkola T. Rationalizing neural predictions. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016. pp. 107–17.
27. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning. PMLR; 2018. pp. 2668–77.
28. Ribeiro MT, Wu T, Guestrin C, Singh S. Beyond accuracy: Behavioral testing of NLP models with checklist. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. pp. 4902–12.
29. Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: 2016 IEEE Symposium on Security and Privacy (SP). IEEE; 2016. pp. 598–617.
30. Hind M, Wei D, Campbell M, Codella NC, Dhurandhar A, Mojsilović A, Natesan Ramamurthy K, Varshney KR. Ted: Teaching AI to explain its decisions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019. pp. 123–9.
31. Staniak M, Biecek P. Explanations of model predictions with live and breakdown packages. R J. 2018;10(2).
32. Zolna K, Geras KJ, Cho K. Classifier-agnostic saliency map extraction. Comput Vis Image Underst. 2020;196:102969.

33. Dash S, Gunluk O, Wei D. Boolean decision rules via column generation. Adv Neural Inf Process Syst. 2018;31.
34. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. 2017. pp. 618–26.
35. Singh C, Murdoch WJ, Yu B. Hierarchical interpretations for neural network predictions. In: International Conference on Learning Representations. 2018.
36. Dhurandhar A, Chen P-Y, Luss R, Tu C-C, Ting P, Shanmugam K, Das P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. Adv Neural Inf Process Syst. 2018;31.
37. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International Conference on Machine Learning. PMLR; 2017. pp. 3145–53.
38. Lapuschkin S, Binder A, Montavon G, Müller K-R, Samek W. The LRP toolbox for artificial neural networks. J Mach Learn Res. 2016;17(114):1–5.
39. Hu L, Jian S, Cao L, Chen Q. Interpretable recommendation via attraction modeling: Learning multilevel attractiveness over multimodal movie contents. In: IJCAI International Joint Conference on Artificial Intelligence. 2018.
40. Petsiuk V, Das A. Saenko K. Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC). 2018.
41. Wang T, Rudin C, Doshi-Velez F, Liu Y, Klampfl E, MacNeille P. A bayesian framework for learning rule sets for interpretable classification. J Mach Learn Res. 2017;18(70):1–37.
42. Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence (vol. 32). 2018.
43. Hong D, Wang T, Baek S. Protorynet - interpretable text classification via prototype trajectories. J Mach Learn Res. 2023;24(264):1–39.
44. Nauta M, Seifert C. The co-12 recipe for evaluating interpretable part-prototype image classifiers. In: Longo L, editor. Explainable Artificial Intelligence. Cham: Springer; 2023. p. 397–420.
45. Datta P, Kibler D. Learning prototypical concept descriptions. In: Machine Learning Proceedings 1995. 1995. pp. 158–66.
46. Wang F, Rudin C. Falling rule lists. In: Artificial Intelligence and Statistics. PMLR; 2015. pp. 1013–22.
47. Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020. pp. 607–17.
48. Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Ser JD, Guidotti R, Hayashi Y, Herrera F, Holzinger A, Jiang R, Khosravi H, Lecue F, Malgieri G, Páez A, Samek W, Schneider J, Speith T, Stumpf S. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Inf Fusion. 2024;106:102301. https://doi.org/10.1016/j.inffus.2024.102301.
49. Vilone G, Rizzo L, Longo L. A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence. 2020.
50. Vilone G, Longo L. A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. Front Artif Intell. 2021;4:717899.
51. Belaid MK, Bornemann R, Rabus M, Krestel R, Hüllermeier E. Compare-XAI: Toward unifying functional testing methods for post-hoc XAI algorithms into a multi-dimensional benchmark. In: World Conference on Explainable Artificial Intelligence. Springer; 2023. pp. 88–109.
52. Rasouli P, Yu IC. Explan: Explaining black-box classifiers using adaptive neighborhood generation. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE; 2020. pp. 1–9.
53. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G, Ranjan R. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Comput Surv. 2023;55(9). https://doi.org/10.1145/3561048.
54. Schwalbe G, Finzel B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Min Knowl Disc. 2023;1:1–59.
55. Saeed W, Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowl Based Syst. 2023;263:110273. https://doi.org/10.1016/j.knosys.2023.110273.
56. Yang W, Wei Y, Wei H, Chen Y, Huang G, Li X, Li R, Yao N, Wang X, Gu X, et al. Survey on explainable AI: From approaches, limitations and applications aspects. Hum Centric Intell Syst. 2023;3(3):161–88.
57. Rong Y, Leemann T, Nguyen T-T, Fiedler L, Qian P, Unhelkar V, Seidel T, Kasneci G, Kasneci E. Towards human-centered explainable AI: A survey of user studies for model explanations. IEEE Trans Pattern Anal Mach Intell. 2024;46(4):2104–22. https://doi.org/10.1109/TPAMI.2023.3331846.
58. Fauvel K, Masson V, Fromont E. A performance-explainability framework to benchmark machine learning methods: Application to multivariate time series classifiers. In: Proceedings of the IJCAI-PRICAI 2020 Workshop on Explainable AI. 2021. pp. 1–8.
59. Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion. 2021;76:89–106.
60. Keane MT, Kenny EM, Delaney E, Smyth B. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In: IJCAI. 2021. pp. 4467–74.
61. Waa J, Nieuwburg E, Cremers A, Neerincx M. Evaluating XAI: a comparison of rule-based and example-based explanations. Artif Intell. 2021;291:103404.
62. Yeh C-K, Hsieh C-Y, Suggala A, Inouye DI, Ravikumar PK. On the (in) fidelity and sensitivity of explanations. Adv Neural Inf Process Syst. 2019;32.
63. Bhatt U, Weller A, Moura JM. Evaluating and aggregating feature-based model explanations. arXiv:2005.00631 [Preprint]. 2020. Available from: http://arxiv.org/abs/2005.00631.
64. Ma E. NLP Augmentation. 2019. https://github.com/makcedward/nlpaug.
65. Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011. pp. 142–50.
66. Kumar H, Harish B, Darshan H. Sentiment analysis on imdb movie reviews using hybrid feature extraction method. Int J Interact Multimed Artif Intell. 2019;5(5).
67. Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing text with the natural language toolkit. 2009.
68. Li Q, Peng H, Li J, Xia C, Yang R, Sun L, Yu PS, He L. A survey on text classification: From traditional to deep learning. ACM Trans Intell Syst Technol. 2022;13(2). https://doi.org/10.1145/3495162.
69. Siino M, Tinnirello I, La Cascia M. Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. Inf Syst. 2024;121:102342. https://doi.org/10.1016/j.is.2023.102342.
70. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning-based text classification: A comprehensive review. ACM Comput Surv. 2021;54(3). https://doi.org/10.1145/3439726.

71. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
72. Breiman L. Classification and regression trees. 2017.
73. Hintze JL, Nelson RD. Violin plots: a box plot-density trace synergism. Am Stat. 1998;52(2):181–4. https://doi.org/10.1080/00031305.1998.10480559.