Received: May 9, 2023. Revised: May 26, 2023.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Developing and Assessing a Human-Understandable Metric for Evaluating Local Interpretable Model-Agnostic Explanations

Rafael M. J. O. Silva[1*] Attilio Sbrana[1] Paulo A. L. de Castro[1] Nei Y. Soma[1]

[1]*Technological Institute of Aeronautics, Brazil*
* Corresponding author's Email: rafaelmj@ita.br

Abstract: Deep learning models, despite their potential, often function as "black boxes", posing significant challenges to interpretability, particularly in sensitive fields such as healthcare and finance. Addressing this issue, we introduce a novel, human-understandable metric aimed at enhancing the interpretability of local interpretable model-agnostic explanations (LIME). Distinct from previous methodologies, this metric is designed to assess the shift in classification probability upon the removal of features (words), thereby providing a unique insight into interpretability. We deploy a convolutional neural network (CNN) for sentiment analysis, interpret predictions utilizing LIME, and evaluate these explanations using three distinct metrics: our proposed metric, a conventional model-based metric, and human evaluations. Through rigorous validation, our metric demonstrated high recall performance, a key indicator of relevant instance retrieval. Results showed worst-case and best-case recalls of 80.29% and 98.19% respectively, against a logistic regression metric for "good" and "excellent" classifications. Comparisons with human evaluations using single-word explanations revealed worst-case and best-case recalls of 82.03% and 94.37%, respectively. These high recall values highlight our metric's effectiveness in aligning with both human judgments and model-based metrics, emphasizing its capacity to capture essential explainability aspects. Furthermore, our study also outlines certain LIME limitations, setting the stage for future interpretability-focused AI research.

Keywords: Explainable AI, LIME, Convolutional neural networks, Sentiment analysis.

1. Introduction

Artificial intelligence (AI) systems have pervaded numerous fields, ranging from facial recognition [1] and gaming [2, 3], to text analysis [4–6] and natural language processing in consumer devices [7]. They have also become invaluable tools in scientific research for prediction, simulation, and exploration [8–10]. The widespread success of AI systems can be attributed to advancements in deep learning methodologies [11, 12], the availability of diverse and large datasets [13, 14], and computational gains from powerful graphics processing units (GPUs) [15]. However, several challenges still hinder the adoption of AI in some applications, including the complexity and high energy demands of deep learning models in resource-limited environments [16], vulnerability to adversarial attacks [17], and the lack of explainability [18–20].

Indeed, the issue of explainability in AI, specifically in deep learning models, is a prevalent challenge that this paper aims to address. We present a novel approach to enhancing the interpretability of black-box models, introducing a distinct metric for evaluating local interpretable model-agnostic explanations (LIME) [21], a widely recognized explainable AI (XAI) method. XAI techniques, such as LIME, strive to elucidate the internal mechanisms of AI algorithms [22], providing a means for end users and stakeholders to justify and verify system outputs [23].

Our proposed metric diverges from traditional approaches by specifically assessing the change in classification probability when certain features (words) are removed, thereby offering a fresh perspective on interpretability. In this comprehensive study, we apply our proposed metric to the interpretation of results from convolutional neural

networks (CNNs) for sentence classification tasks, utilizing the CNN architecture proposed in [24].

To further evaluate the effectiveness of our proposed metric, we contrast it with two distinct evaluations: one derived from a more interpretable model, specifically logistic regression, and the other based on user assessments obtained via Amazon mechanical turk (MTurk), a crowdsourcing platform. We implement this methodology on two publicly available sentiment analysis datasets in English and Portuguese, thereby ensuring the wide applicability of our findings. Furthermore, we enrich these datasets with human-generated explainability labels via the MTurk platform, which not only serves to enhance the utility of the datasets for our study but also underscores the significance of interpretability in deep learning models. Our results demonstrate the effectiveness of our novel approach, highlighting its potential in advancing the field of explainable AI.

The remainder of this paper is organized as follows: section 2 details the motivation for this research. Section 3 reviews previous work related to this study. In section 4, we present the methods employed and the proposed metric. Section 5 describes the evaluation process used to validate our research. Section 6 discusses our results, and finally, in section 7, we conclude and suggest future work.

2. Motivation and research goal

The pervasive use of machine learning (ML) systems in vital sectors such as healthcare, finance, and criminal justice has magnified the urgency for transparency and explainability. Often portrayed as “black box” systems due to their non-linear and complex mechanisms, ML models pose significant challenges for domain experts striving to comprehend their decision-making processes, particularly in high-stakes applications.

This demand for comprehensibility has sparked the evolution of explainable artificial intelligence (XAI), a subfield dedicated to developing interpretable algorithms to satisfy the growing need for human interaction with ML systems and to build trust in their predictions. Regulatory frameworks, such as the general data protection regulation (GDPR) [25], further amplify this demand, underscoring the right to explanation and thus enabling individuals to question and challenge decisions made by automated ML models.

However, a crucial gap persists in the effective evaluation of XAI methods' explainability components. An evaluation mechanism is pivotal in verifying the credibility of the explanations, comparing different XAI methods, and selecting the best-suited technique for a specific domain. Current XAI research primarily concentrates on balancing explainability and predictive performance [26]. But the mere provision of explanations without robust evaluation can instill a false sense of security [27]. Furthermore, inherent human bias towards explanations may lead to a preference for more persuasive, rather than transparent, XAI systems [28].

The lack of a specific evaluation approach for individual XAI methods adds another layer to this complexity. While some studies propose general metrics for evaluating interpretability methods, these do not offer a distinct evaluation approach tailored for a single XAI technique. Existing evaluation methods often rely heavily on subjective human opinions or are excessively complex, complicating interpretation.

This research aims to bridge these gaps by proposing a specific, human-understandable metric to evaluate the performance of a single XAI method - local interpretable model-agnostic explanations (LIME). Our proposed metric, based on a scoring system that assesses the importance of words in the explanation provided by LIME and ranks them accordingly, provides an objective, tailored approach to evaluate LIME. We also introduce two grading strategies for sentences and classify the quality of explanation into three levels: “insufficient”, “good”, and “excellent”, thereby providing an objective measure of the explanation's quality.

Addressing potential cognitive biases, our metric aligns with human perception and interpretation of explanations. By comparing this proposed metric with one derived from the interpretable method of logistic regression and with human evaluations, we aim for a more comprehensive and objective evaluation.

Our research objectives include:

1. Developing a novel, human-understandable metric for evaluating the performance of LIME.
2. Comparing the proposed metric with one derived from the interpretable method of logistic regression and with human evaluations.
3. Providing a unique dataset with normalized word scores in the context of sentiment analysis.

This study adds to the growing body of XAI research, potentially influencing societal implications by enhancing ML systems' transparency and trustworthiness across various applications. Our objective is not only to make ML systems more understandable and reliable but also to ensure the

effectiveness and appropriateness of the explanations provided by these systems.

3. Related work

This section offers an overview of recent studies focusing on evaluating explainable AI (XAI) techniques, underscoring the lacunae in the literature that this research seeks to fill. The existing evaluation methodologies range widely, from automated quantitative metrics to innovative counterfactual-based methodologies, and yet none provides a fully satisfactory approach.

Research by [29] centered on assessing the quality of textual explanations generated by XAI techniques. They employed three automated quantitative metrics—BLEU, METEOR, and CIDEr, concentrating on evaluating sentence similarity and semantic similarity between words.

Another study [30] applied XAI methods such as SHAP and LIME to deep learning and random forest models to detect credit card fraud. Their performance was evaluated based on accuracy, recall, sufficiency, and the F1 score.

A counterfactual-based methodology was introduced by [31] to assess the faithfulness of explanations from a counterfactual reasoning perspective. This approach developed algorithms to identify counterfactuals in both discrete and continuous scenarios. Additionally, XAI methods were utilized in remote sensing multi-label classification tasks [32], with assessments based on quantitative metrics.

A novel trust metric for evaluating interpretability methods was proposed in [33]. It argued that an interpretability approach should be independent of the task and the machine learning method and should facilitate more intuitive, rapid, and accurate decisions.

While these studies have made significant strides in evaluating XAI methods, several gaps persist. The research by [30] employs questionnaires to evaluate explanations; this approach, although valuable, is intrinsically subjective due to its dependence on human opinions. Furthermore, the explanations presented by [29], [31] and [32] offer a more objective perspective, but their complexity and technical language can make them difficult for non-experts to interpret.

Finally, while [33] offers a promising metric for interpretability, it does not provide a distinct evaluation method for individual XAI techniques. It is those gaps that our research aims to fill.

We propose a unique, human-understandable metric explicitly tailored for evaluating a single XAI method—local interpretable model-agnostic explanations (LIME) in natural language processing tasks. This approach not only evaluates the effectiveness of LIME but also aligns with human cognition to address potential cognitive biases. By comparing our metric with explainable methods and user classifications, we provide a more nuanced, objective, and comprehensible assessment tool for XAI techniques, thus addressing the limitations identified in existing literature.

4. Methods

In this section, we provide an in-depth understanding of the datasets employed in our study, the rationale for using convolutional neural networks (CNNs) and logistic regression as machine learning techniques, and the details of their implementation. Moreover, we discuss the LIME interpretable method and the metric employed to evaluate our approach and machine learning methods' performance. Additionally, we describe how we utilized the MTurk tool to collect explanatory labels for our metrics and present the primary algorithm used to generate our metric.

4.1 Datasets

To evaluate our metric, we use two sentiment analysis datasets, one in English and one in Portuguese, aiming to investigate the metric's behaviour in different languages and expand research on the less-studied Portuguese language. We employ the following datasets in our experiments:

4.1.1. Brazilian E-commerce public dataset by olist

This anonymized dataset [34] comprises customer feedback from satisfaction surveys pertaining to purchases made on various Brazilian marketplaces between 2016 and 2018. The dataset includes a diverse range of features such as order status, price, freight performance, customer location, product attributes, and customer reviews. The customer reviews, which are textual data, were used for our sentiment analysis. They have been anonymized by replacing company references with Game of Thrones great houses' names. The dataset was split randomly into a training set and a test set, with 2/3 of the data used for training and 1/3 for testing (Table 1).

4.1.2. IMDB movie ratings sentiment analysis

This dataset, obtained from Kaggle [35], consists

Table 1. Train/test division of Brazilian e-commerce public dataset by Olist

Train	Test
26573	13287

Table 2. Train/test division of IMDB movie ratings sentiment analysis dataset

Train	Test
12103	6051

of tab-separated files containing movie review phrases from the rotten tomatoes dataset. Each phrase is paired with a sentiment label ranging from 0 (negative sentiment) to 4 (positive sentiment). In our analysis, we combined sentiment labels 0 and 1 into a 'negative' class and labels 3 and 4 into a 'positive' class. This dataset was also randomly split into training and test sets, with a 2:1 ratio (Table 2).

These datasets were chosen because they offer a rich source of real-world textual data for sentiment analysis. The diversity of the datasets, in terms of language and domain (e-commerce and movie reviews), helps to ensure the generalizability of our findings. Furthermore, both datasets have been preprocessed and cleaned, ensuring high-quality data for our analysis.

4.1.3. Data preprocessing and class distribution

Our study employed distinct data preprocessing methods tailored to the Brazilian E-commerce public dataset and the IMDB movie ratings sentiment analysis dataset for both the convolutional neural network (CNN) and logistic regression models.

For the CNN model, we integrated word vectors extracted from a publicly available[1], unsupervised neural language model. As for the logistic regression model, we utilized the grid search method [36] to pinpoint the optimal combination of preprocessing techniques and hyperparameters. The tested preprocessing techniques encompassed TF-IDF (term frequency-inverse document frequency), mean word Embeddings, Stemming, and Lemmatization. Of these, TF-IDF emerged as the most effective technique for both datasets.

We maintained a balanced distribution of the subsets used from the Brazilian E-commerce public dataset and the IMDB movie ratings sentiment analysis dataset. For the former, out of a total of 39,860 sentences, there were 21,213 negative sentences (representing 53.21%) and 18,647 positive sentences (representing 46.79%).

For the IMDB movie ratings sentiment analysis dataset, we aggregated sentiment labels 0 and 1 to form a 'negative' class, and labels 3 and 4 to form a 'positive' class. This yielded a distribution of 9,512 positive sentences (52.39%) and 8,642 negative sentences (47.61%) from the total of 18,154 sentences employed in our study.

4.2 Convolutional neural network

We chose the convolutional neural network (CNN) as our representative black-box model due to its intricate architecture, robust performance across diverse machine learning tasks, and widespread use in the field. This selection ensures that our study's findings are both relevant and applicable to a broader context.

Our study canters on the architecture proposed by [24]. The authors introduced a simplified CNN-based model specifically for text classification, which has since become a benchmark for contemporary models. The model employs a single convolutional layer on top of input word vectors derived from an unsupervised neural language model (word2vec). Several efforts have been made to enhance CNN-based model architectures [37-41]. This model utilizes one-dimensional convolution and max-over-time pooling, with individual pretrained token representations serving as input and facilitating the transformation of sequence representations for downstream applications.

For a text sequence represented by d-dimensional vectors comprising n tokens, the width, height, and number of channels of the input tensor are n, 1, and d, respectively. The model processes the input to generate the output through the following stages:

1. Multiple one-dimensional convolution kernels are defined to perform convolution operations on the inputs. Convolution kernels of varying widths can detect local features among diverse quantities of adjacent tokens.
2. Max-over-time pooling is performed on all output channels, followed by the concatenation of all scalar pooling outputs into a vector.
3. The fully connected layer transforms the concatenated vector into output categories.

[1] code.google.com/archive/p/word2vec/

Table 3. Hyperparameters used by CNN network

Filter windows	3,4,5
Feature maps per window	100
Dropout rate	0.5
L2 constraint	3
mini- batch size	50
optimization algorithm	Adadelta algorithm

Table 4. Hyperparameters used by logistic regression

Preprocessing technique	TF-IDF
Regularization technique	L2 penalties
optimization algorithm	saga solver
regularization strength (C)	5

For our model, we set hyperparameters according to those in [24]. These hyperparameters were determined using a grid search conducted on the datasets used in the paper. Table 3 shows parameters used by our CNN:

The code is available via our GitHub repository [42].

4.3 Logistic regression

We selected logistic regression as the comparative method due to its simplicity, interpretability, versatility, and widespread use. Additionally, its well-established nature facilitates meaningful comparisons with the novel metric proposed in our study.

As mentioned in section 4.1.3 we utilized grid search to identify the best combination of preprocessing techniques and hyperparameters. The preprocessing techniques tested include TF-IDF (term frequency-inverse document frequency), mean word Embeddings, Stemming, and Lemmatization. The best technique for preprocessing was TF-IDF for both datasets. The hyperparameters tuned are Penalty (we employ L1 and L2 penalties), C (we use a value of 5), and Solver (we utilize liblinear, newton-cg, saga solver and lbfgs to handle L2 penalties).

Table 4 shows the best preprocessing technique and hyperparameters for logistic regression after applied grid search in the datasets of this research.

4.4 Local interpretable model-agnostic explanations (LIME)

We chose LIME as the primary focus of our research due to its widespread use and well-established status in the field of explainable AI. Studying LIME enables us to provide insights directly applicable to numerous researchers and practitioners working with this technique. Furthermore, LIME serves as a representative example of a class of model agnostic XAI techniques, making it an ideal candidate for our study.

Local interpretable model-agnostic explanations (LIME) [21] is a technique designed to elucidate the predictions of any machine learning model, irrespective of its architecture or training methodology. LIME aims to provide an interpretable explanation of a model's behaviour within a specific context, such as a particular input or a distinct region of the feature space.

The primary concept of LIME involves approximating a model's decision boundary locally, surrounding the point of interest, by training an interpretable model (e.g., a linear model or decision tree) on a small, perturbed sample of the data. This interpretable model is subsequently employed to clarify the predictions of the original, more complex model. LIME consists of three primary steps:

1. Perturb the input by sampling new, synthetic instances around the point of interest.
2. Train a simple, interpretable model on the perturbed instances and their corresponding model outputs.
3. Explain the predictions of the original model by analysing the weights and feature importance of the interpretable model.

4.5 Performance evaluation metrics for machine learning algorithms

For our experiments to succeed, it is essential to ensure the satisfactory performance of the machine learning algorithms (logistic regression and CNN) when classifying negative and positive sentences in the datasets utilized in this research. We will use ⅔ of the datasets for training and ⅓ for testing. We employ accuracy as a performance measure, defined as follows:

1. TP (true positive): When the model correctly classifies a sentence with a positive connotation.

2. FP (false positive): When the model predicts a sentence with a positive connotation, but it has a negative connotation.
3. TN (true negative): When the model correctly predicts a sentence with a negative connotation.
4. FN (false negative): When the model predicts a sentence with a negative connotation, but it has a positive connotation.

Accuracy is calculated using Eq. (1):

$$accuracy = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (1)$$

Given that our datasets are balanced, and our intention is not to delve into model comparisons, we will use accuracy to compare our model with the best results found in the literature and machine learning competitions. A competitive model is crucial for the validity of this research, even though comparing and improving models is not our primary goal.

4.6 Investigating the recall of the proposed method

The objective of this work is to determine the number of words and sentences that our metric will classify at the same level ("insufficient", "good", and "excellent") as the classification generated by the logistic regression's metric or user evaluation. To achieve this goal, we will use the recall metric, which is appropriate for this purpose. Considering:

1. TP (true positive): The number of true positives, or the cases where our metric agrees in classification level ("insufficient", "good", and "excellent") with the logistic regression metric or human classification.
2. FN (false negative): The number of false negatives, which are cases when our metric classifies a word or sentence differently from the logistic regression metric or human classification.

Recall is calculated using Eq. (2):

$$recall = \frac{TP}{TP+FN} \quad (2)$$

4.7 Amazon mechanical turk (MTurk)

One approach to evaluate our metric involves comparing our classification with human classification. To facilitate this comparison, we recruited human subjects through MTurk, a crowdsourcing platform that enables individuals and businesses to delegate tasks to a virtual, distributed workforce.

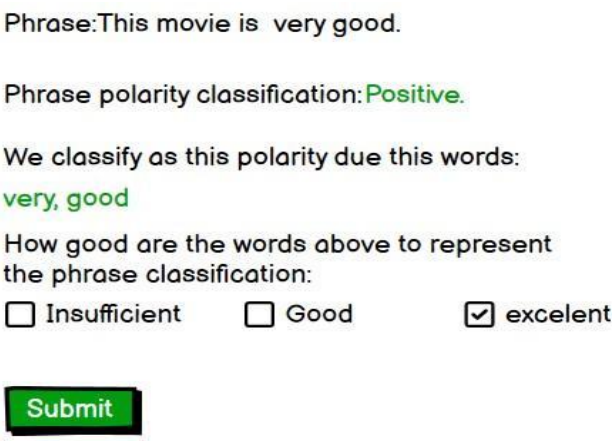

Figure. 1 Example of screen used to collect user's opinion of classification returned by LIME

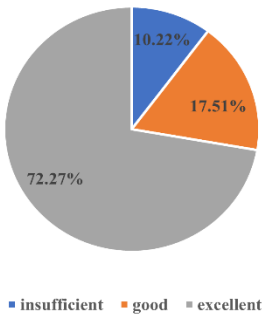

Figure. 2 Distribution of users' opinions on LIME ratings in sentences in Portuguese and English

We enlisted 200 participants who had prior experience in labelling sentiment analysis datasets in Portuguese and English but lacked expertise in machine learning. Participants were asked to assess the representativeness of the words returned by LIME (using one-word and two-word strategies) for classifying a given word as having a negative or positive sentiment. Each user was presented with a randomized screen similar to the one depicted in Fig. 1 and was instructed to select a classification ("insufficient", "good" or "excellent"). Fig. 1 provides an example of the screen displayed to each participant.

We labelled 5519 sentences in Portuguese and 4812 sentences in English. Fig. 2 below shows an initial division of the user's opinion about LIME evaluation.

4.8 Proposed algorithm for evaluating LIME

Our research aims to propose an algorithm for evaluating the LIME technique by assigning scores to features (words). The fundamental premise of our

algorithm is that the more important a word is for classification, the greater the likelihood that the probability of correct classification decreases when the word is removed. Our metric employs this concept to assign scores to words and generate two lists of scores: one for words with a positive impact and another for words with a negative impact.

The algorithm, formally represented in Fig. 3, accepts three variable inputs: a dataset, a dictionary of negative words, and a dictionary of positive words. Each sentence in the dataset is processed by the algorithm, which interprets them using LIME. It then computes the word scores according to the change in classification probability when a word is removed. These word scores are then added to the appropriate dictionary. Finally, the algorithm processes the dictionaries to compute average scores, which are then normalized using min-max scaling within a range of 0 to 1.

To clarify the functions and variables, here are their definitions:

1. **dataset:** The input dataset with sentences and its classification (in our case negative and positive sentences)
2. **neg_words:** Dictionary that will host negative words and scores found for those words.
3. **pos_words:** Dictionary that will host positive words and scores found for those words.
4. **classify_sentence(model, sent):** This function classifies a sentence using the given model.
5. **interpret_using_lime(model, sent):** This function generates LIME interpretations for a given sentence and model.
6. **calculate_prob_class(sent):** This function calculates the probability of a sentence belonging to a specific class.
7. **remove_word(sent, word):** This function removes a specified word from a sentence.
8. **words_conotation(word):** This function checks the connotation (positive or negative) of a given word.
9. **add_word(dictionary, word, word_grade):** This function adds a word and its associated score to an existing key in the dictionary.
10. **create_key_add_word(dictionary, word, word_grade):** This function creates a new key in the dictionary for a word and adds its associated score.
11. **avg_grades(dictionary):** This function calculates the average score of all the words in a dictionary.
12. **normalize_min_max_0_1(dictionary):** This function applies min-max scaling to the scores in a dictionary, normalizing them to a range of 0 to 1.

For each word in a sentence, we compute a **word_grade** which is the difference in classification probability before and after the word is removed. In mathematical terms, if **P(sent)** represents the probability of correct classification of the sentence **sent**, and **P(sent - word)** represents the probability of correct classification after the word is removed from the sentence, then the **word_grade** can be computed as:

$$word\_grade = P(sent) - P(sent - word) \quad (3)$$

This **word_grade** serves as a measure of the importance of the word to the classification of the sentence.

For the min-max normalization, if we have a dictionary of word grades **D**, and **min(D)** and **max(D)** represent the minimum and maximum word grades in **D** respectively, then the normalized word grade **word_grade_norm** for a word with grade **word_grade** can be computed as:

$$word\_grade\_norm = \frac{word\_grade - \min(D)}{\max(D) - \min(D)} \quad (4)$$

This normalization process scales the word grades to a range of 0 to 1, allowing for better comparison and interpretation of word grades across different sentences and datasets.

5. Evaluation methodology

By comparing the performance of our metric with logistic regression-based scoring and human-generated classifications, we seek to validate the metric's consistency, reliability, and ability to capture meaningful insights.

The experiments are designed to evaluate the performance of our metric at both word and sentence levels, considering different scoring strategies (i.e., using the two most important words and the most important word separately). This multifaceted evaluation approach allows us to understand the impact of various scoring methods on assessing LIME-generated explanations and identify potential areas of improvement for our proposed metric.

Algorithm 1 Create lists with word's grades using proposed metric

Input: dataset, neg_words, pos_words;

```
for (sent in dataset) do
   classify_sentence(model, sent)
   words_from_lime ← interpret_using_lime(model, sent);
   for (word in words_from_lime) do
      before_remove_prob ← calculate_prob_class(sent);
      sent_with_word_removed ← remove_word(sent, word);
      after_remove_prob ← calculate_prob_class
  (sent_with_word_removed);
      word_grade ← before_remove_prob-after_remove_prob;
      if (words_conotation(word) == 'negative') then
         if (word in neg_words.keys()) then
            add_word(neg_words, word, word_grade);
         else
            create_key_add_word(neg_words, word, word_grade);
         end if
      else
         if (word in pos_words.keys()) then
            add_word(pos_words, word, word_grade);
         else
            create_key_add_word(pos_words, word, word_grade);
         end if
      end if
   end for
end for
avg_grades(neg_words);
avg_grades(pos_words);
normalize_min_max_0_1(neg_words);
normalize_min_max_0_1(pos_words);
```

Figure. 3 Algorithm that create lists with word's grades using proposed metric

Additionally, by incorporating human evaluation, we ensure that our metric aligns with human intuition and provides meaningful insights that can be easily understood and utilized by practitioners in the field of explainable AI.

5.1 Logistic regression-based feature grading

We compare the proposed grades with those generated by a logistic regression. A logistic regression model is trained on a given dataset, and the coefficients of the model are then utilized to assign positive or negative weights to the words based on their sign. The weights of these words are normalized using the min-max scaling method between 0 and 1, creating two separate dictionaries for positive and negative words.

5.2 Grade categorization

We will divide the grades of the features into "insufficient", "good", and "excellent". The sorted list of grades is partitioned into four equally sized batches. The first two batches are assigned the label "insufficient", the third batch is labelled "good", and the final batch is designated "excellent".

This step creates a clear separation of word grades based on their performance. This approach will be used to separate the grades generated by the logistic regression model and our algorithm. We will use the lists generated here to attribute levels ("insufficient", "good", and "excellent") to words and sentences.

5.3 LIME Interpretation and sentence grading

We will incorporate LIME interpretation to grade sentences. This process is performed in two distinct ways, as described below. In the first case, we use the two most important words returned by LIME, and in the second case, we return only the most important word. By exploring both approaches, we aim to better understand the impact of different grading strategies on the evaluation of LIME-generated explanations.

5.3.1. Two most important words

In the first approach, sentences in the dataset are graded based on the two most important words identified by LIME. For each sentence, its classification is determined, and the two most relevant words are obtained via LIME interpretation. The grades of these words are retrieved from the corresponding dictionaries of negative or positive words presented previously. The final sentence grade is then calculated as the average of the two-word grades. This approach considers the combined contribution of the top two features, which may provide a more comprehensive evaluation of the sentence. However, it may also dilute the impact of the most critical word on the sentence classification, leading to less accurate grading.

5.3.2. Most important word

In the second approach, sentences are graded by considering the most important word returned by LIME. Like the first approach, each sentence in the dataset is classified, and the most relevant word is identified through LIME interpretation. The grade of this word is acquired from the appropriate negative or positive words dictionary, and the sentence grade is subsequently recorded.

Focusing on the most important word simplifies the grading process and emphasizes the significance of the primary feature in the classification. However, it may overlook the contribution of other important features, leading to an incomplete evaluation of the sentence.

5.4 Experiment description

In this section, we provide a comprehensive overview of the experiments designed to assess our proposed metric. The experiments concentrate on evaluating the performance of our metric, using logistic regression-based metrics and human-generated explanations as references. We begin by

executing the algorithms required to generate word grades and levels for each dataset. The generated lists are as follows:

1. Word grades from our metric (positive and negative)
2. Word grades from logistic regression (positive and negative)
3. Levels generated from our metric (positive and negative)
4. Levels generated from logistic regression (positive and negative)

We assess the level of agreement between our metric and the logistic regression metric or user evaluation using recall, as explained in Section 4.6.

For the experiments, we calculate the individual recall of each list (English/negative, English/ positive, Portuguese/ negative, Portuguese/positive) for each case (word or sentence) relative to the logistic regression metric or human classification and analyse the average of those recalls. This analysis will serve as the foundation of our evaluation.

5.4.1. Experiment 1: Word-level agreement

This experiment compares the number of words classified at the same level ("insufficient," "good," and "excellent") in the lists generated by our metric and logistic regression. We train a logistic regression model with robust performance for sentence classification. Such a model accurately scores the importance of features based on their coefficients. Consequently, comparing the scores generated by our metric and the logistic regression model provides evidence of the validity and reliability of our proposed metric.

5.4.2. Experiment 2: Sentence-level agreement with two most relevant words

In this experiment, we compare the classification of sentences by our metric and the logistic regression metric when considering the two most important features returned by LIME. The steps are as follows:

1. Grade sentences using our metric and the two most relevant words and take the average.
2. Assign levels to sentence grades based on our metric.
3. Grade sentences using logistic regression and the two most relevant words and take the average.
4. Assign levels to sentence grades based on logistic regression.
5. Compare the number of sentences classified at the same level.

The goal of the experiment is to determine the frequency of disagreement between the two metrics when classifying explanations.

5.4.3. Experiment 3: Sentence-level agreement with most relevant word

This experiment complements Experiment 2. We investigate the recall of our metric using the logistic regression metric as a reference when considering only the most relevant word returned by LIME. The steps are similar to Experiment 2, with the primary difference being the focus on the most relevant word.

5.4.4. Experiment 4: Human evaluation with two most relevant words

This experiment investigates the agreement between our metric and human-generated classifications when considering the two most important features returned by LIME. The steps are as follows:

1. Grade sentences using our metric and the two most relevant words and take the average.
2. Assign levels to sentence grades based on our metric.
3. Obtain human-generated classifications of LIME explanations.
4. Compare the number of classifications that coincide with human evaluations.

A high agreement with human evaluations indicates the potential usefulness of our metric in quantifying LIME-generated explanations.

5.5 Experiment 5: Human evaluation with most relevant word

This experiment explores the agreement between our metric and human-generated classifications when considering only the most relevant word returned by LIME. The steps are similar to Experiment 4, with the primary difference being the focus on the most relevant word. This experiment helps to determine if a single feature is sufficient to explain the classification and understand how often such a situation occurs.

Table 5. Accuracy comparison

	CNN	Logistic Regression	Best Result found
Brazilian E-Commerce Public Dataset	**88.10%**	85.20%	86.99%
IMDB Movie Ratings Sentiment Analysis	88.70%	**90.45%**	85.20%

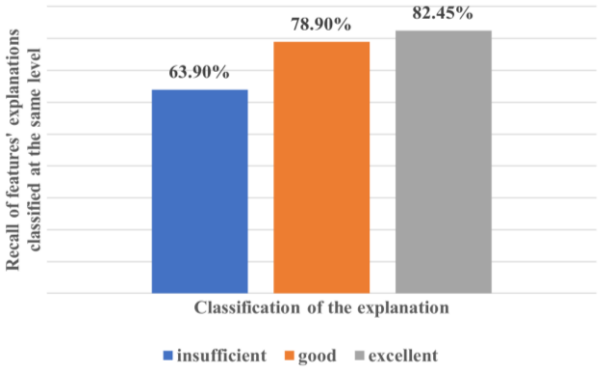

Figure. 4 Average recall of features' classified at same level

6. Results and discussion

6.1 Performance of CNN and logistic regression models

To assess the performance of the CNN and logistic regression models, we compared their accuracy to the best results available in the literature. Establishing a competitive model is crucial for the validity of this research. Table 5 presents the performance of the CNN and logistic regression models in comparison to the best results found in the literature and Kaggle competitions [35], utilizing accuracy as the metric.

For the Brazilian E-commerce public dataset by olist, the best result in the literature was obtained using a random forest model [43]. The IMDB movie ratings sentiment analysis dataset's best result was achieved with a CNN model from a Kaggle competition [44]. As shown in Table 1, our models' results are competitive with those found in recent literature and machine learning competitions.

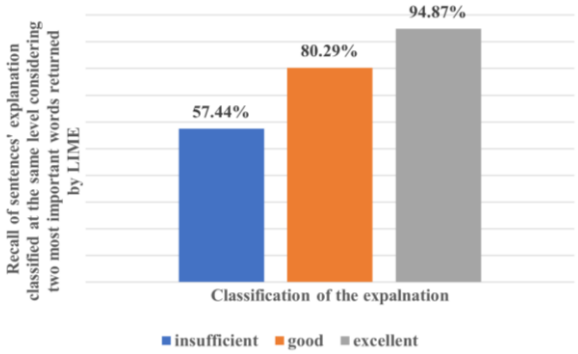

Figure. 5 Average recall of sentences' explanations classified at the same level considering the two most important words returned by LIME

6.2 Results of experiment 1

In experiment 1, we compared the number of words classified at the same level ("insufficient", "good", and "excellent").

Fig. 4 shows the average of the recall in the four lists when using the metric generated by logistic regression as reference. The highest recall occurs at the highest grades (82.45% for "excellent" and 78.90% for "good"), while the lowest recall is observed for grades classified as "insufficient" (63.90%). We attribute this result to the concentration of features with strong negative/positive connotations at higher grades, which both our metric and logistic regression metric capture. The limited use of features and the high number of features with weak negative/positive connotations contribute to the lower recall for "insufficient" grades.

In conclusion, the grades generated by our metric can be useful for evaluating the explainability of sentences using the LIME technique. This is demonstrated by comparing the behaviour of our metric's grades with those generated by logistic regression.

6.3 Results of experiment 2

In experiment 2, we aimed to compare the grades our metric assigned to each sentence with the grades assigned by logistic regression, considering the two most relevant words returned to classify a level by LIME. Fig. 5 shows the recall of our metric when using logistic regression metric as reference.

Our metric demonstrated a recall of approximately 94.87% and 80.29%, on average, with the metric generated by the explainable method for explanation grades classified as "excellent" and "good" respectively. This is considered a satisfactory recall. An initial conclusion suggests that, in sentences with higher grades, the features do not

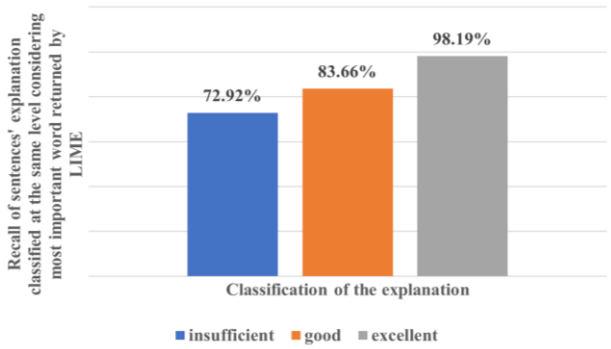

Figure.6 Average recall of sentence's explanations classified at the same level considering the most important word

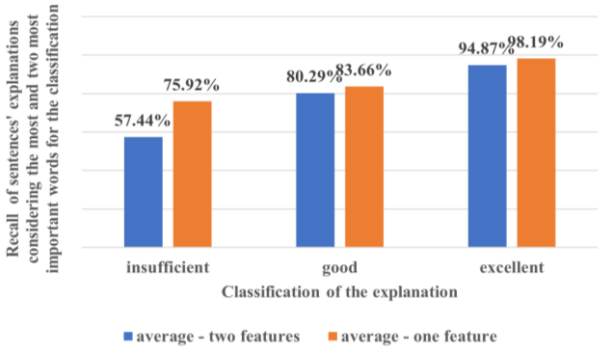

Figure. 7 Comparison of recall considering the most and two most important words for the classification

significantly interfere with each other. In other words, the inclusion of two features leads to most cases agreeing on both the first and second features. However, for the "insufficient" level, a lower recall rate (57.44% on average) is observed. This can be attributed to the higher variability in grades when they are too low. For example, consider a situation where LIME attributes a positive classification based on the words "actor" and "movie" which do not have strong positive/negative connotations. Consequently, the grades of both metrics are low, leading to an "insufficient" classification for the explanation.

However, slight differences in the grades assigned by logistic regression and our metric could result in different classifications (e.g., "insufficient" versus "good") despite both being close to the boundary. This behaviour is more common in the lowest level due to the higher number of features with low grades, a characteristic of the dataset.

6.4 Results of experiment 3

In experiment 3, we aimed to determine whether our metric recall increases or decreases when considering only the most relevant feature, as opposed to the two most relevant features.

Fig. 6 displays the average recall between all lists of our metric for each level. The pattern observed in experiment 2 persists, with a higher percentage of metric agreement at the highest levels ("good" and "excellent") and lower at the lowest level ("insufficient"). The same explanation provided in experiment 2 applies in this experiment. However, when using only the most important feature, the recall increases across all levels compared to when just one feature is used.

Fig. 7 compares the recall when using the most relevant feature and the two most relevant features returned by LIME.

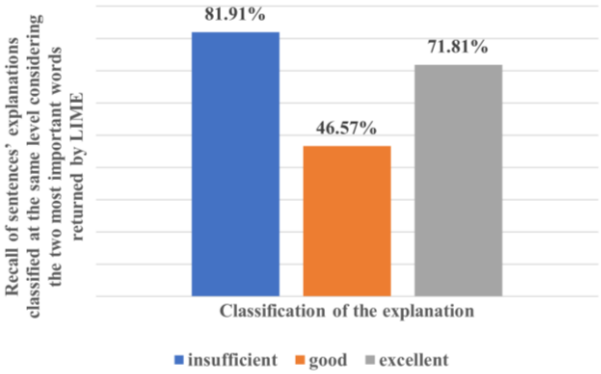

Figure. 8 Average recall of sentences' explanations classified at the same level considering the two most important words returned by LIME

We observe that the recall increases across all levels. Notably, the increase in the "insufficient" level is more significant. While the "excellent" and "good" levels experienced increases of 3.32 and 3.37 percentage points, respectively, the "insufficient" level saw an increase of 18.48 percentage points (nearly six times the increment in other levels). We can conclude that in datasets with characteristics like the one used in this study, where few words are decisive for classification, using more features can adversely impact the outcome. This is especially true when the features used for classification carry low weight in the final classification (in our case, low positive/negative connotation)

6.5 Results of experiment 4

In experiment 4, we compared the classifications derived from our metric with those provided by human evaluators. We asked participants to classify the explanations generated by LIME into "insufficient", "good" and "excellent". We then calculated the recall. Subsequently, we calculated the recall. It is crucial to note that our metric, in this experiment, employs the average grade of the two

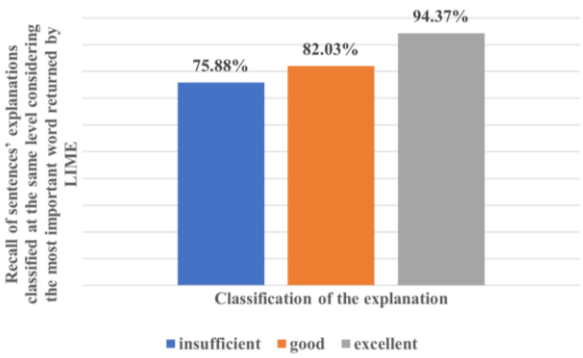

Figure. 9 Average recall of sentences' explanations classified at the same level considering the most important word returned by LIME

most relevant words returned by LIME when explaining a classification.

As illustrated in Fig. 8, our metric's recall is high for explanations classified as "insufficient" but low for other classifications, which is not ideal. We attribute this discrepancy to LIME's dependence on a limited number of words to interpret results, coupled with the scarce presence of words with strong positive or negative connotations.

For instance, consider the phrase "I love this movie" with LIME's output being "love" and "movie," and their corresponding grades being 0.1 for "movie" and 0.8 for "love." Suppose that grades below 0.5 are classified as "insufficient", those between 0.5 and 0.75 as "good" and those above 0.75 as "excellent". A human evaluator may reasonably classify the sentence as "excellent" due to the presence of the word "love." However, our metric would classify it as "insufficient" due to the average grade of 0.45. We will assess the validity of this hypothesis in the subsequent experiment.

6.6 Results of experiment 5

In experiment 5, we once again compared our metric-based classifications with those provided by human evaluators. Participants were asked to classify explanations as "insufficient", "good" and "excellent." We then calculated the percentage of sentences that our metric classified at the same level as those classified by human evaluators. In this experiment, our metric considered the grade of the most relevant word returned by LIME when explaining a classification.

As depicted in Fig. 9, the recall is high across all levels. This result supports the hypothesis proposed in the previous experiment, suggesting that LIME's reliance on a limited number of words to interpret results, in conjunction with the restricted presence of words with strong positive or negative connotations, leads to suboptimal outcomes when multiple words are employed for interpretation.

Our findings indicate that it is essential to consider individual word grades before providing a final evaluation. For instance, when using more than one word, we should assess the average grade and exclude additional features from the final evaluation if they lower the overall score. Future research could explore the potential benefits of presenting these additional features as part of the explanation for a given classification.

In conclusion, the results of our experiments demonstrate that the choice of words used in explanations can significantly impact classification accuracy. Experiment 4, which employed the average grade of the two most relevant words returned by LIME, showed lower recall for "good" and "excellent" classifications. This indicates that using multiple words for interpretation can hinder classification accuracy. Conversely, Experiment 5, which considered the grade of the most relevant word returned by LIME, displayed high recall across all levels, suggesting that focusing on the most relevant word can improve classification accuracy. These findings emphasize the importance of carefully selecting the words used in explanations and their impact on the effectiveness of classification models.

7. Conclusion and future work

The increasing demand for explainability in black-box models, such as deep learning models, has spurred the development of explainable artificial intelligence (XAI) techniques. Assessing the explainability components of these methods is crucial for ensuring their utility and trustworthiness for end-users. Our research is a novel attempt in this direction; we introduce a human-understandable metric tailored specifically for evaluating local interpretable model-agnostic explanations (LIME), one of the popular XAI methods. This metric is juxtaposed with another interpretable method (logistic regression) and human evaluations on two sentiment analysis datasets, one in English and another in Portuguese. Our metric demonstrated strong recall performance, with the worst recall being 80.29% and the best 98.19% for "good" and "excellent" classifications using logistic regression-generated metrics, and 82.03% worst and 94.37% best for "good" and "excellent" classifications using single-word explanations based on human evaluations.

We found that generating explanations with a limited number of words was more effective. However, we identified that LIME does not capture

phrase structures, relying exclusively on isolated features to explain deep learning model classifications. This suggests a potential area for improvement by incorporating more complex structures, such as bigrams or trigrams, as we observed that combinations of words with low negative or positive connotations often determine sentence polarity.

While our study makes notable strides in evaluating LIME, it is not without limitations that offer future research opportunities. Expanding the scope of investigation to include a wider variety of black-box models can enhance the generalizability of our findings. Additionally, considering the subjectivity inherent in human evaluations, future studies could involve larger, more diverse participant groups and incorporate alternative evaluation methods. Furthermore, our study paves the way for developing a theoretical foundation for our proposed metric, which can bolster its validity and contribute to a more profound understanding of its performance.

Despite these limitations, our research serves as a pioneering effort in developing evaluation methods for LIME-generated explanations. By addressing these limitations in future work, we can refine our approach and better understand the nuances of explainable AI, thereby enhancing its relevance and applicability. Numerous other future research directions remain. We suggest testing this metric in various domains to confirm the consistency of the observed behaviour and using different datasets within the same domain to investigate and measure potential benefits when attributing grades to features based on more extensive data.

In conclusion, this study makes a novel contribution by developing a unique, human-understandable metric for evaluating XAI techniques. We tested this proposed method with two different approaches and successfully compiled a list of graded words valuable for future research. We identified gaps in LIME and proposed enhancements, thus paving the way for future research directions. Ultimately, our work aims to increase the transparency and trustworthiness of XAI, further bridging the gap between machine learning algorithms and human cognition.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Rafael M. J. O. Silva and Nei Y. Soma; methodology, software, Rafael M. J. O. Silva; validation, data curation, writing—original draft preparation, Rafael M. J. O. Silva, and Attilio Sbrana; supervision and writing—review and editing, Paulo A. L. de Castro, and Nei Y. Soma; funding acquisition, Paulo A. L. de Castro, and Nei Y. Soma.

References

[1] Z. Lei, X. Zhang, Y. Shuangyuan, and O. Akindipe, "RFR-DLVT: A Hybrid Method for Real-Time Face Recognition Using Deep Learning and Visual Tracking", *Enterprise Information Systems*, Vol. 14, pp. 1379-1393, 2018.

[2] D. Soemers, V. Mella, C. Browne, and O. Teytaud, "Deep Learning for General Game Playing with Ludii and Polygames", *ICGA Journal*, Vol. 43, No. 3, pp. 146-161, 2021.

[3] I. Oh, S. Rho, S. Moon, S. Son, H. Lee, and J. Chung, "Creating Pro-Level AI for a Real-Time Fighting Game Using Deep Reinforcement Learning", *IEEE Transactions on Games*, Vol. 14, No. 2, pp. 212-220, 2022.

[4] B. Jlifi, C. Sakrani, and C. Duvallet, "Towards a Soft Three-Level Voting Model (Soft T-LVM) for Fake News Detection", *Journal of Intelligent Information Systems*, 2022.

[5] P. Yu, W. Tan, W. Niu, and B. Shi, "Aspect-Location Attention Networks for Aspect-Category Sentiment Analysis in Social Media", *Journal of Intelligent Information Systems*, 2022.

[6] M. Imani and S. Noferesti, "Aspect Extraction and Classification for Sentiment Analysis in Drug Reviews", *Journal of Intelligent Information Systems*, Vol. 59, pp. 613-633, 2022.

[7] B. J. Abbaschian, D. S. Sosa, A. Elmaghraby, "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models", *Sensors*, Vol. 21, No. 4, 2021.

[8] S. Chmiela, H. Sauceda, K. Müller, and A. Tkatchenko, "Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields", *Nature Communications*, Vol. 9, No. 3887, 2018.

[9] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-Chemical Insights from Deep Tensor Neural Networks", *Nature Communications*, Vol. 8, 2017.

[10] D. Wu, L. Wang, and P. Zhang, "Solving Statistical Mechanics Using Variational Autoregressive Networks", *Physical Review Letters*, Vol. 122, 2019.

[11] Y. Chen, Y. Hu, L. He, and H. Huang, "Multi-Stage Music Separation Network with Dual-Branch Attention and Hybrid Convolution", *Journal of Intelligent Information Systems*, Vol. 59, pp. 635-656, 2022.

[12] Z. Abbasiantaeb, and S. Momtazi, "Entity-Aware Answer Sentence Selection for Question Answering with Transformer-Based Language Models", *Journal of Intelligent Information Systems*, Vol. 59, pp. 755-777, 2022.

[13] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. F. Fei, "Imagenet: A Large-Scale Hierarchical Image Database", In: *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, pp. 248-255, 2009.

[14] J. G. Cañón, N. G. Páez, L. Porcaro, A. Porter, E. Cano, P. H. Boyer, A. Gkiokas, P. Santos, D. H. Leo, C. Karreman, and E. Gómez, "Trompa-Mer: An Open Dataset for Personalized Music Emotion Recognition", *Journal of Intelligent Information Systems*, 2022.

[15] E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym, "Nvidia Tesla: A Unified Graphics and Computing Architecture", *IEEE Micro*, Vol. 28, pp. 39-55, 2008.

[16] S. Han, J. Pool, J. Tran, and W. Dally "Learning Both Weights and Connections for Efficient Neural Network", In: *Advances in Neural Information Processing Systems*, Montreal, Canada, 2015.

[17] Y. Wang, S. Jha, and K. Chaudhuri, "Analyzing the Robustness of Nearest Neighbors to Adversarial Examples", In: *Proc. of the 35th International Conference on Machine Learning*, Chicago, USA, pp. 5133-5142, 2018.

[18] F. D. Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning", Preprint at https://arxiv.org/abs/1702.08608, 2017.

[19] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and Explainability of Artificial Intelligence in Medicine", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 9, 2019.

[20] W. Samek, T. Wiegand, and K. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models", Preprint at https://arxiv.org/abs/1708.08296, 2017.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier", In: *KDD '16: Proc. of the 22nd ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016.

[22] S. Mathews, "Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review", *In: K. Arai, R. Bhatia, and S. Kapoor (eds.), Intelligent Computing*, Springer, Cham, pp. 1269-1292, 2019.

[23] F. Hussain, R. Hussain, and E. Hossain, "Explainable Artificial Intelligence (XAI): An Engineering Perspective", Preprint at https://arxiv.org/abs/2101.03613, 2021.

[24] Y. Kim, "Convolutional neural networks for sentence classification", Preprint at https://arxiv.org/abs/1408.5882, 2014.

[25] "GDPR-info", accessed April 18, 2023, https://gdpr-info.eu/.

[26] G. Vilone and L. Longo, "Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence", *Information Fusion*, Vol. 76, pp. 89-106, 2021.

[27] M. Eiband, D. Buschek, A. Kremer, and H. Hussmann, "The Impact of Placebic Explanations on Trust in Intelligent Systems", In: *Proc. of ACM Conference*, New York, USA, pp. 1-6, 2019.

[28] B. Herman, "The Promise and Peril of Human Evaluation for Model Interpretability", Preprint at https://arxiv.org/abs/1711.07414, 2019.

[29] S. Barratt, "InterpNET: Neural Introspection for Interpretable Deep Learning", Preprint at https://arxiv.org/abs/1710.09511, 2017.

[30] Y. Ji, "Explainable AI Methods for Credit Card Fraud Detection", *Master's thesis, University of Skövde*, 2021.

[31] Y. Ge, S. Liu, Z. Li, S. Xu, S. Geng, Y. Li, J. Tan, F. Sun, and Y. ZHang "Counterfactual Evaluation for Explainable AI", Preprint at https://arxiv.org/abs/2109.01962, 2021.

[32] I. Kakogeorgiou, and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing", *International Journal of Applied Earth Observation and Geoinformation*, Vol. 103, 2021.

[33] P. Schmidt and F. Biessmann, "Quantifying interpretability and trust in machine learning systems", Preprint at https://arxiv.org/abs/1901.08558, 2019

[34] S. Olist and A. Sionek, "Brazilian e-commerce public dataset by offlist", 2018.

[35] "Kaggle website, accessed April 18, 2023, https://www.kaggle.com.

[36] P. Liashchynskyi and L. Liashchynskyi, "Grid search, random search, genetic algorithm: A big

comparison for NAS", Preprint at arXiv https://arxiv.org/abs/1912.06059, 2019.
[37] J. Liu, W. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification", In: *Proc. of ACM*, New York, USA, pp. 115-124, 2017.
[38] R. Johnson, and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks", Preprint at https://arxiv.org/abs/1412.1058, 2014.
[39] R. Johnson, and T. Zhang, "Deep pyramid convolutional neural networks for text categorization", In: *Proc. of ACL*, Vancouver, Canada, pp. 562-570, 2017.
[40] A. B. Duque, L. L. J. Santos, D. Macêdo, C. Zanchettin, "Squeezed very deep convolutional neural networks for text classification", *Artificial Neural Networks and Machine Learning – ICANN 2019: Theoretical Neural Computation*, pp. 193-207, 2019.
[41] Z. Liu, H. Huang, C. Lu, and S. Ly, "Multichannel CNN with attention for text classification", Preprint at https://arxiv.org/abs/2006.16174, 2020.
[42] "R. M. J. O. Silva's Github", accessed April 18, 2023, https://github.com/rafajacomel/phd-ita.
[43] F. R. Oliveira, "Modelos lineares e não lineares aplicados à análise de sentimentos de consumidores de e-commerce no Brasil", 2020. doi:https://doi.org/10.13140/RG.2.2.28178.27849.
[44] S. Shinde, "Multi-channel CNN model for sentiment analysis", accessed April 18, 2023, https://www.kaggle.com/code/shivamshinde123/multi-channel-cnn-model-for-sentiment-analysis.