Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Explainable AI: A case study on a Citizen's Complaint Text Classification Model

Stella Mendes Meireles Bonifácio

Dissertação apresentada como requisito parcial para conclusão do Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Guilherme Souza Rodrigues

Brasília
Junho 2024

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Mendes Meireles Bonifácio, Stella
MB715e Explainable AI: A case study on a Citizen’s Complaint Text
Classification Model / Stella Mendes Meireles Bonifácio;
orientador Guilherme Rodrigues. -- Brasília, 2024.
50 p.

Dissertação(Mestrado Profissional em Computação Aplicada)
-- Universidade de Brasília, 2024.

1. Machine Learning. 2. XAI. 3. Explainability. 4.
Artificial Intelligence. 5. LIME. I. Rodrigues, Guilherme,
orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Explainable AI: A case study on a Citizen's Complaint Text Classification Model

Stella Mendes Meireles Bonifácio

Dissertação apresentada como requisito parcial para conclusão do Mestrado Profissional em Computação Aplicada

Prof. Dr. Guilherme Souza Rodrigues (Orientador)
EST/UnB

Prof. Dr. Thiago Paulo Faleiros	Prof. Dr. Roberta Akemi Sinoara
CIC/UnB	IFSP

Prof. Dr. Gladston Luiz da Silva
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 14 de Junho de 2024

Dedicatória

Dedico esse trabalho aos meus pais e irmãos, ao meu esposo e ao meu filho.

Agradecimentos

Agradeço à minha família pelo apoio, ao meu orientador pela ajuda em buscar o melhor resultado neste trabalho e ao meu esposo pelo cuidado e carinho durante todo esse processo

Resumo Expandido

IA Explicável: Um Estudo de Caso sobre um Modelo de Classificação de Textos de Denúncias de Cidadãos

A sociedade atual é muito influenciada por sistemas de Inteligência Artificial (IA) em diversos contextos e, embora ela tenha proporcionado diversas contribuições, é importante construir uma abordagem transparente e responsável para os modelos de IA, de modo que as pessoas possam se beneficiar de suas vantagens porém sem deixar de prevenir eventuais danos que o uso dessa nova tecnologia possa causar à sociedade.

No âmbito da Controladoria-Geral da União, o benefício da IA pode ser observado pelo uso da ferramenta FARO que, além de outras funcionalidades, incorpora um modelo de classificação de textos em denúncias aptas e não-aptas, de modo a auxiliar os auditores na tarefa de tratamento das denúncias feitas pelos cidadãos.

Esse modelo recebe como entrada a denúncia de cidadãos sobre situações de corrupção, comportamentos inadequados de servidores públicos, assédio moral etc. A análise dos textos é feita e o modelo gera uma nota que varia entre 0 e 1 sobre a aptidão da denúncia analisada. Para que uma denúncia seja considerada apta, é preciso que ela colecione certos elementos como por exemplo estar relacionada a uma entidade que tenha vínculo com recursos públicos de origem federal, descrever uma irregularidade que reflita dano ao patrimônio público e que contenha uma justificativa mínima para viabilizar investigação da situação relatada.

No decorrer do processo de avaliação da denúncia em apta ou não-apta, outras informações relevantes são levantadas e podem ser utilizadas tanto pelo modelo de classificação textual, que faz parte do escopo deste estudo, quando por outras etapas da própria ferramenta FARO.

Atualmente, os auditores tem acesso ao resultado gerado pelo modelo mas sem mecanismos de explicabilidade que poderiam aumentar a transparência do processo, aumentar o entendimento do resultado gerado e melhorar o processo decisório da classificação de denúncias.

Apesar de não existir uma metodologia ou estrutura ideal para interpretar ou explicar modelos de aprendizado de máquina, é possível encontrar estudos sobre os diferentes métodos de explicabilidade, inclusive sobre as suas limitações na tarefa de explicar totalmente o modelo.

Estudos recentes apontam que integrar ferramentas de explicabilidade ao uso de modelos de inteligência artificial traz benefícios ao processo de tomada de decisão e monitoramento do comportamento do modelo, tendo em vista a prevenção de vieses. Entretanto, ainda são poucos os estudos de casos concretos na área de Explainable Ai (XAI), especialmente no que se refere a Natural Processing Language (NLP). O presente estudo tem como objetivo apresentar um sistema de explicabilidade do modelo textual, integrado à própria ferramenta FARO [1] e específica para as necessidades dos auditores, de forma a subsidiar a monitoração do modelo e de seus resultados pelos auditores.

O sistema de explicabilidade é baseado na ferramenta LIME (Local Interpretable Model-Agnostic Explanations) devido às suas características, entre outras, de ser agnóstica ao modelo utilizado, à sua implementação intuitiva e compatibilidade com os demais sistemas que já encontram-se em ambiente de produção.

Além disso, o LIME apresenta visualizações interessantes que auxiliam os auditores, inclusive aqueles que tenham um menor contato com a área de tecnologia, no entendimento dos resultados gerados pelo modelo de classificação. Desse modo, este trabalho contribui para a monitoração e melhoria do modelo de classificação textual e pode servir como um modelo para trabalhos futuros que investiguem outras etapas utilizadas na ferramenta FARO, como por exemplo o tratamento de dados estruturados.

Palavras-chave: Machine Learning, XAI, LIME, Explainability, Machine Learning, Artificial Intelligence

Abstract

Present-day society is highly influenced by Artificial Intelligence (AI) systems in various contexts. Although AI has provided numerous contributions, it is important to build a transparent and responsible approach to AI models, so that people can benefit from their advantages while also preventing potential harm that the use of this new technology might cause to society.

Within the scope of the Office of the Comptroller General, the benefit of AI can be observed through the use of the FARO tool [1], which, among other functionalities, incorporates a text classification model for classifying complaints. This helps auditors in handling complaints submitted by citizens. However, currently, auditors have access to the results generated by the model but lack explainability mechanisms that could enhance the transparency of the process, increase the understanding of the generated results, and improve the decision-making process for classifying complaints.

The present study aims to introduce an explainability system for the text model, integrated into the FARO tool itself and tailored to the needs of the auditors, in order to support the monitoring of the model and its results by the auditors.

Keywords: Machine Learning, XAI, LIME, Explainability, Machine Learning, Artificial Intelligence

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Artificial intelligence (AI) systems find diverse applications across multiple contexts, such as healthcare, industry, marketing and numerous other fields. In the medical domain, AI systems can aid in achieving early diagnoses for diseases, i.e. breast cancer screening [7], detecting diabetic retinopathy and diabetic macular edema [8], diagnosing stroke [9], and classify genes [10]. The industry also benefits significantly from AI, leveraging it for tasks like enhancing energy efficiency, improving quality control, and forecasting demand [11]. Additionally, AI enables organizations to track real-time data, analyze and respond swiftly to customer requirements, detect fraud, determine credit score, prevent churn and gain essential consumer insights into consumer behavior [12].

It is a fact that today's society is profoundly influenced by the role of AI. Some researches state that there is evidence that such learning algorithms have reached or even surpassed the performance of humans in isolated tasks [13] emphasizing the significance of AI as a vital tool in advancing societal development.

However, the influence of machine learning and computer vision algorithms is not always beneficial. They can significantly impact people's lives, sometimes negatively, as for example the exacerbation of social or economic inequalities [14] [15], affecting not only individuals and organizations, but society as a whole [14]. For instance, there have been cases where AI systems demonstrated biased behavior such as an AI-based recruiting engine used by Amazon.com Inc., which reportedly downgraded resumes from female candidates in favor of male candidates [16], facial recognition systems misidentifying people of color, women, and young people at high rates [17], a Twitter Inc. faced issues with an AI-operated system that was verbally abusive when communicating with users and Google LLC's AI-powered image search which returned racist results in certain instances [18].

To address these issues, governments, corporations and international organisations alike are committing to an accountable, responsible, transparent approach to AI, where human values and ethical principles are leading [19]. In order to achieve a more transparent AI researchers are growing interest in interpreting machine learning models and gaining insights into their working mechanisms [20].

Several machine learning models are classified as "black-box" due to their opaque decision-making process, making it difficult for humans, including the developers of these models, to fully understand how they arrive at predictions [21]. Addressing this limitation, Explainable Artificial Intelligence (XAI) seeks to develop human-interpretable models, particularly in sensitive sectors like the military, banking, and healthcare applications [22].

In this work, the terms "interpretability" and "explainability" will be used interchangeably. By achieving interpretability or explainability, XAI can foster greater trust in the models among users and stakeholders. Additionally, it empowers developers to identify, monitor, and proactively address potential issues, thereby enhancing system safety and reliability.

Many authors worked on the development of frameworks in order to capitalise on the opportunities for more rigour, structure and normalisation in this field [23], but the task of finding an ideal methodology or framework to interpret or explain machine learning models, evaluate, measure and compare different explanations for these models, remains unanswered. The existing literature has yet to reach a consensus on a definitive framework for evaluating the explicability of AI models. Important questions such as “What constitutes an acceptable explanation?” and “how to establish user trust in AI-powered systems?” are still unresolved [22].

Although XAI techniques may help humans to better understand AI systems and increase it’s transparency, certain authors have expressed concerns regarding the explanations themselves. They argue that these explanations might carry considerable uncertainty, potentially undermining users’ trust in the predictions and raising doubts about the model’s overall robustness [21].

Comparing interpretable methods presents challenges due to the vast array of different metrics available for evaluating explanations, and the complexities of applying them consistently across multiple explanation methods. Moreover, the integration of XAI methods in industry can lead to increased project costs and time consumption. Additionally, there remains a shortage of studies that thoroughly evaluate the explanations provided for day-to-day applications, highlighting the need for further research in this area. Even though organisations agree on the need to consider ethical, legal and societal principles, how these are interpreted and applied in practice, varies significantly across the different recommendation documents and there is still much work needed to ensure that AI is developed and used in responsible ways that contribute to trust and well-being [19].

This is not a justification for disregarding or neglecting explainability practices, as they serve as valuable tools for enhancing the transparency of the model, its monitoring, and preventing misuse. In the context of public administration, there remains a deficiency in implementing more robust methods for monitoring models in production, despite the growing attention this issue has gained from experts, managers, and developers in recent times [24].

Despite there being several libraries and explainability modules available, such as SHAP [25], LIME [6], and others, the models used by the F.A.R.O. tool, developed by the Office of the Comptroller General (CGU) for the treatment of complaints, do not make use of any of these libraries integrated into its process, which hinders auditors' access to relevant explainability information.

The use of explainability tools integrated into the business process can assist those responsible in monitoring the model's operation, ensuring the use of responsible AI aligned with the specific needs of each situation, such as the treatment of sensitive data.

Conducting further studies in this area, particularly those focusing on real-world problems, can offer valuable insights into the responsible use of AI systems. Such research can help us understand how to leverage AI's capabilities while avoiding adverse impacts on society. Recent studies [26] [27] have indicated that integrating explainability tools into the use of artificial intelligence models brings about advantages in the decision-making process and in monitoring model behavior, with the aim of bias prevention. However, there is still a lack of concrete case studies in the field of explainable AI, particularly concerning Natural Language Processing (NLP).

1.2 Objectives

The present study aims to present an explainability system for the textual model used by F.A.R.O. [1], that can be integrated into the tool itself, and tailored to the needs of auditors, in order to support the monitoring of the model and its results by auditors, as well as to provide greater transparency for the process of handling complaints. The objective is to showcase how a XAI algorithm can be integrated into a current classification process while providing an viable alternative to how domain users can benefit from the visualization of a model´s explanation in the task of screening complaints.

1.3 Contributions

While the interest in XAI has been increasing over the past years, there is a scarcity of studies addressing tangible and applicable scenarios for integrating explanations into real-world contexts. Moreover, there is a limited focus on evaluations conducted by users, particularly in cases involving explanations for text models and assessments by users who possess domain expertise.

In light of this gap in research, our study contributes to the field of XAI by looking into the practical implications and advantages of interpretable AI within a real-world context, specifically in the context of classifying text complaints. This integration has shown promise in enhancing the decision-making process and monitoring model behavior, ultimately with the goal of preventing biases. This research seeks to shed light on these aspects, presenting a viable application of explainability techniques integrated to a textual model used in the process of screening complaints.

The aim is to support auditors in the complaint screening process by enhancing transparency in the textual model's scoring, facilitating comprehension of the model, and serving as a tool for monitoring and refining its performance. Additionally, the explainability system presented in this work can inspire other real-world applications of explainability systems in the context of the public administration, specifically in the form of integrating the explanations into the business processs, shedding light on the importance of explanability techniques to produce responsible artificial intelligence.

1.4 Structure of the dissertation

To delineate the scope of this dissertation, the following organizational structure is employed: Chapter one provides an introduction, elucidating the motivation, objectives, and contributions of this study. Chapters two contextualizes the reader with theoretical concepts pertaining to Explainable Artificial Intelligence (XAI) and its various approaches to explanations, offering concise descriptions of two prevalent algorithms in the field, namely LIME and SHAP.

Chapter three offers a comprehensive insight into the F.A.R.O. process, meticulously outlining its steps and giving the reader with a good understanding of the context in which the explanation module will be integrated. Specifically, it details the textual model, elucidating its characteristics in detail.

Chapter four outlines the methodology adopted for this study, delineating the considered processes and detailing their respective phases, looking into the formulation and execution of the study's experiments. Finally, the results of this study are presented in chapter five and the conclusions in chapter six.

This structured approach ensures a comprehensive exploration of the subject matter, guiding the reader through the motivation, theoretical foundations, algorithmic insights, methodological considerations, and experimental facets of the dissertation.

Chapter 2

Explainable AI

Artificial intelligence extends beyond being a mere automation technique. Instead, it can be better comprehended as a socio-technical ecosystem, acknowledging the intricate interplay between people and technology. This perspective recognizes how complex infrastructures affect and are affected by society and by human behaviour [28].

Questioning the utility of interpretable and explainable machine learning is a natural inclination, especially considering the prevalent and inaccurate belief that a trade-off exists between accuracy and interpretability [29]. This often raises doubts in software architects about why they should sacrifice performance in favor of transparency. It is essential to acknowledge that there are numerous instances where interpretability is entirely unnecessary, particularly when dealing with well-known problems that lack significant consequences [30]. Also, the presumed accuracy-interpretability trade-off may not necessarily be present in numerous datasets [29].

In addition, the advancing capabilities of AI models often increase the model's complexity contributing to it's opacity. Opacity refers to human scrutiny and understanding of an AI system's decision-making process [31] preventing humans to build appropriate trust in machine learning solutions. They often either blindly follow the system's decisions and recommendations or do not use them [32]. Also, opacity can lead to humans negligently rely on AI results and substitute their own judgment with potentially false decisions [15] but there is a concern that the lack of explainability may prevent the use of AI systems.

Explainable AI (XAI) addresses the opacity of AI systems by automatically generating explanations for their functioning and outcomes while maintaining the AI's high performance levels [20]. Interpretable and explainable machine learning techniques emerge from a need to design intelligible machine learning systems that can be comprehended by a human mind [33]. Although we can encounter pitfalls in some explanations provided by famous algorithms which are commonly used by developers and researches, their use is an important tool for developers, domain experts and users to start uncovering the problem of bias, improve the decision-making process and prevent unwanted outcomes.

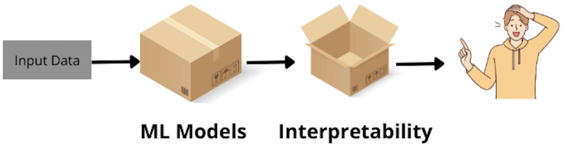


Figure 2.1: Interpretable Machine Learning (ML) [2].

Government, authorities and organizations are proposing different regulations e.g. the EU AI regulation that requires human oversight—to interpret and contest AI systems' outcomes—in "high-risk" applications such as recruiting or credit score evaluation [34] and Brazilian General Data Protection Law (LGPD) which determines that the personal data belongs to the person to whom it concerns and not to who is storing the data in the databases [35].

There is an ongoing debate within the machine learning community regarding definition of interpretability and the task of interpretation [30] [36]. According to Doshi et.al (2017), interpretability can be defined as the ability to explain or to present in understandable terms to a human. Some authors draw a clear line between interpretable and explainable ML stating that interpretable ML focuses on designing models that are inherently interpretable whereas explainable ML provides post-hoc explanations for existing black-box model [29]. A prevalent term in the literature is "explainability", a concept that is closely tied with interpretability. The notion of interpretability often depends on the domain of application [29] and the target explainee [37], e.g. the recipient of interpretations and explanations. Many authors do not differentiate between the two [37] and this work will use them interchangebly.

Since there is no general definition of either interpretability or explainability, researchers have elicited various desiderata, diverse and often contradicting, to provide motivation for different techniques [38] that can be summarized as the following goals to be achieved with interpretability:

- **Trust:** Lipton [36] decomposes trust into knowing "how often a model is right" and "for which examples it is right". It is easier for humans to trust a system that explains its decisions rather than a black box that just outputs the decision itself [37].

- **Causality:** supervised learning models are only optimized directly to make associations but researchers often use them in the hope of inferring properties or generating hypotheses about the natural world [36] although there is no guarantee to these associations to reflect causal relationships [36].

- **Reliability/Robustness/Transferability:** systems should be resistant to noisy inputs and (reasonable) domain shifts [38]. This line of research is closely connected to the challenges of domain adaptation and transfer learning [39] e.g. additive models for pneumonia risk prediction having domain adaptation facilitated by interpretability [40].
- **Fairness:** Traditionally, the fairness of a machine learning system has been evaluated by checking the models' predictions and errors across certain demographic segments such as groups of a specific ethnicity or gender, focusing on the social and ethical impact of machine learning algorithms in terms of impartiality and discrimination [41].
- **Privacy:** Especially for systems that rely on personal data, it is crucial to ensure the protection of sensitive information present in the data [37].

Explainability techniques have emerged as a crucial tool for pursuing these objectives as they aim to do evaluate and improve the system to detect its flaws and prevent unwanted behaviour, justify its decisions by improving transparency and accountability and learn from the system by identifying unknown correlations that could indicate causal relationships in the underlaying data [3].

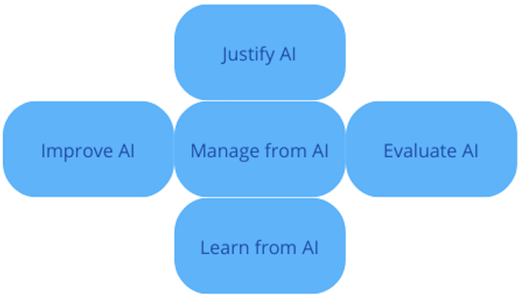


Figure 2.2: Generalized objectives of explainable artificial intelligence [3].

2.1 Evaluation of Interpretability

The issue of interpretability taxonomy remains unresolved, as experts have not yet reached a unanimous agreement. Various methods and approaches exist to make a model interpretable, and understanding these methods, their distinctions and assessments is crucial.

It allows us to categorize and group different techniques, ultimately enabling us to select the most appropriate one or a combination of methods suitable for each specific problem and context. One other aspect to be considered is the target explainee [37].

Explainees can be divided into three groups: developers and AI researchers, domain experts, and lay users. The first category includes data scientists, computer engineers, and researchers who build or maintain AI systems. The second category comprises individuals with expertise in the application domain based on formal education or professional experience. Finally, lay users are non-experts who are affected by AI decisions or interact with AI systems.

Apart from analyzing whether a real-world problem necessitates the use of explainability or stands to benefit from it, selecting an explanation method requires careful consideration of its usefulness for the intended users. Additionally, having a critical evaluation of the cost-benefit trade-offs associated with the implementation method is crucial. The focus should be on maximizing the benefits for each type of explainee, ensuring that the chosen approach aligns well with their needs and preferences.

Explanations are often qualitative [5], which leads to much discussion. It is not clear how to quantitatively evaluate and compare interpretability methods but validation is still paramount. Doshi-Velez and Kim [30] classified the evaluation of interpretability into three categories: functionally-grounded, human-grounded, and application-grounded. This categorization reflects the need of providing explanations that are both useful to humans (human-grounded) and accurately reflect the model's behavior (functionally-grounded). These categories each provide an important but different aspect for validating interpretability and should therefore be used in combination.

Application-grounded

Application-grounded evaluation calls for conducting human experiments within a real application [30]. The interpretability method is evaluated in the environment it will be deployed. e.g. if there is a concrete application in mind, such as working with doctors on diagnosing patients with a particular disease, the best way to show that the model works is to evaluate it with respect to the task: doctors performing diagnoses [30]. This rationale is in line with the evaluation methods commonly employed in the human-computer interaction and visualization communities, ensuring that the system effectively delivers on its intended task [30].

Examples of experiments include conducting a domain expert experiment with the exact application task, or alternatively, conducting a domain expert experiment with a simpler or partial task to reduce the experiment's duration and expand the pool of potentially willing participants [30].

Incorporating a baseline where explanations are provided by humans is crucial. However, due to the application-specific and time-consuming nature of this approach, application-grounded evaluation is rarely performed in NLP interpretability research. Instead, researchers often resort to more synthetic and general evaluation setups, which functionally-grounded and human-grounded evaluation methods encompass [5].

2.1.1 Human-grounded

Human-grounded metrics is about conducting simpler human-subject experiments that maintain the essence of the target application and can be completed with lay humans, which is appealing when experiments with the target community is challenging and allows for both a bigger subject pool and less expenses [30]. It is most appropriate when the researcher wishes to test more general notions of the quality of an explanation, without a specific end-goal such as identifying errors in a safety-oriented task or identifying relevant patterns in a science-oriented task [30].

Binary forced choice, forward simulation/prediction and counterfactual simulation are examples of human-grounded evaluation. In binary forced choice, humans are presented with pairs of explanations, and must choose the one of higher quality. Secondly, forward simulation/prediction presents humans with an explanation and an input where one must correctly simulate the model's output regardless of the true output. Finally, humans are presented with an explanation, an input, and an output in counter-factual simulation. They are asked what must be changed to change the method's prediction to a desired output.

Although human-grounded evaluation is much more efficient than application-grounded evaluation, it still requires time due to the human involvement. A common approach is to substitute the human with a simulated user, which can be problematic due to the fact that creating explanations that are truly informative to humans is a complex task and often necessitates interdisciplinary knowledge from fields like human-computer interaction and social science. Substituting a human with a simulated user can lead to overly optimistic results, which may not accurately reflect real human responses and understanding [30].

In general, common evaluation strategies are:

- Humans have to choose the best model based on an explanation [6].
- Humans have to predict the model's behavior on new data [42].
- Humans have to identify an outlier example called an intruder [43].

2.1.2 Functionally-grounded

Functionally-grounded is more commonly known as faithfulness [4] [44] [6] or fidelity. Although it might seem surprising that an explanation, which is directly produced from the model, would not reflect it, some interpretable methods, even intrinsically interpretable ones such as Attention and Neural Modular Networks, have been shown to not reflect the model [45].

Measuring if an interpretability method is functionally-grounded, for some tasks e.g. adversarial examples, is trivial. In this case, it is enough to show that the prediction changed and the adversarial example is a paraphrase. However, in other cases, most notably input features, providing a functionally grounded metric can be very challenging [45].

Common strategies to measure functionally-grounded are [45]:

- Comparing with an intrinsically interpretable model, such as logistic regression [6].
- Comparing with other post-hoc methods [45].
- Benchmarking against random explanations [46] [47].

Various authors have put forward different categorizations of interpretability methods. In the following chapters, we will present some of these categories.

2.2 Post-Hoc vs. Intrinsic

Du, Liu and Hu (2019) classifies existing models into two categories:

- **Traditional Machine Learning:** relies on feature engineering, a process that transforms raw data into features that better represent represent the predictive task.
- **Deep Learning:** discovers the mapping from representation to output and also learns representations from raw data

Post-hoc global explanation aims to offer a comprehensive understanding of the knowledge acquired by pre-trained models. It illuminates the parameters or learned representations in an intuitive manner for human comprehension [4]. They do not elucidate how black-box models work but instead help to provide useful information after the black-box model has already been deployed [48].

In traditional machine learning, model-agnostic feature importance is widely applicable to models by treating it as a black-box and not inspecting internal model parameters.

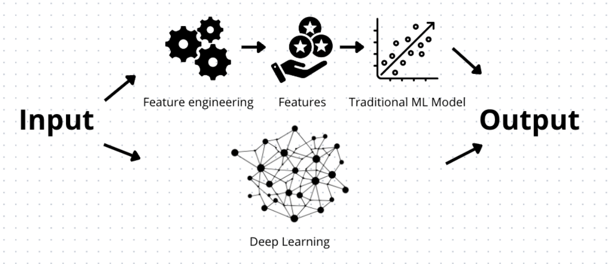


Figure 2.3: Traditional machine learning pipeline and a deep learning pipeline [4].

Permutation Feature Importance, for instance, is based on the idea that the that the importance of a specific feature to the overall performance of a model can be determined by calculating how the model prediction accuracy deviates after permuting the values of that feature [4].

On the contrary of model-agnostic explanations, post-hoc model specific explanations are designed specifically for each different model. Usually, they derive explanations by examining internal model structures and parameters [4]. While it is true that many post-hoc methods are model-agnostic, it is essential to note that this property is not always a requirement, and in certain cases, it only applies to a specific category of models [5].

Intrinsic transparency pertains to a model's capacity to be completely understood and comprehended by humans. As a result, models with high complexity such as deep neural networks or random forests cannot be categorizes as transparent. According to Du, liu and Hu (2019), models can provide an accurate explanation of how models intrinsically work, however, accuracy might have to be sacrified as complexity decreases to ensure explainability. On the other hand, post-hoc interpretability sacrifice performance but they work on model approximations and end up with limitations as of how close they can mimic or explain the predictions of the original model. Post-Hoc interpretability provides a significant advantage as it does no require to modify the model itself.

When it comes to explaining any black-box model, two methods stand out as the most comprehensive and dominant across the literature: LIME [6] and SHAP [49]. These methods are widely used for visualizing feature interactions and feature importance. Both LIME and SHAP are not only model-agnostic, but they have been demonstrated to be applicable to any type of data. White-box highly performing models are very hard to create, especially in computer vision and natural language processing, where the gap in performance against deep learning models is unbridgeable [41].

In the interpretability literature, it is customary to categorize communication strategies into three types: local explanations, global explanations, and class explanations. Local explanations focus on explaining a single observation, while global explanations aim to explain the entire model. Additionally, class explanations are a distinct category of methods that explain an entire output-class.

Madsen et. al (2022) proposed the following categories w.r.t. post-hoc interpretability methods [5], which will be explored in more detail later on.

- **Local explanations:** explain a single observation
  - Input Features
  - Adversarial Examples
  - Influential Examples
  - Counterfactuals
  - Natural Language
- **Class explanations:** summarize the model, but only with regard to one selected class
  - Concepts
- **Global explanations:** summarize the entire model with regards to a specific aspect
  - Vocabulary
  - Ensemble
  - Linguistic information
  - Rules

2.3 Local explanations

2.3.1 Input Features

It's a local explanation that aims to determine how important an input feature is for a given predicion. Input feature explanations can only explain one scalar e.g., one class at one timestep. In a sequence-to-sequence application, the explanation is replicated for each timestep [50] [51] though this approach may not adequately address the combinatorial complexities.

This approach aims to answers which tokens are most important for the prediction. It exhibits high adaptability to various problems, as the input features are always known and often hold meaningful interpretations for humans [5]. Gradient [50] and LIME [6] are examples of this strategy. In Natural Language Processing (NLP), the input features will often represent words, sub-words, or characters. An input feature explanation of the input x is represented as

$$E(x, c) : I^d \to R^d$$

where c is the desired class, I is the input domain and d is the dimensionality. When the output is a score of importance, the explanation is called an importance measure [5].

2.3.2 Adversarial Examples

An adversarial example is an input that causes a model to produce an incorrect prediction due to inherent limitations in the model. Typically, an adversarial example is generated from an existing example for which the model gives a correct prediction. The goal is to understand what conditions could disrupt the model's prediction. The adversarial method A maps the input x to the adversarial example $\tilde{x}$.

$$A(x) \to \tilde{x}$$

Several surveys about adversarial examples have been conducted, such as the ones by Wang et al. [30] and Glass [52]. These methods help us identify the support boundaries of a given example, which, in turn, provides insights into the underlying logic of the model, leading to interpretability. Interestingly, these adversarial explanations can be similar to input features. However, there is a significant distinction: adversarial explanations are contrastive, meaning they explain by comparing with another example, whereas input features only explain in the context of the original example [5].

To ensure that an adversarial example method is functionally-grounded, one only needs to assert that the predicted label changes while the gold label remains the same. Additionally, it is desirable for the original and adversarial examples to be similar in many applications this can be framed as paraphrasing. These explanations might not generalize easily to sequence-to-sequence problems [5].

The HotFlip [53] algorithm uses gradients to estimate the effect of changing a specific token to another one. To constrain the possible changes, so that the adversarial sentence is a paraphrase, hotflip uses word-embeddings, such that the adversarial word and the original word are constrained to have a cosine similarity of at least 0.8 [53].

In figure 2.4, the highlight indicates the gradient w.r.t. the input, which is used to select which token to change. The x indicates the original sentence, $\tilde{x}$ indicates the adversarial sentence, y is the desired class and $p(y|x; \theta)$ is the probability of the input x to belonging to class y [5].

One common limitation of adversarial example methods is their lack of control over the search direction. For instance, although changing "unpredictable" to "unforeseeable" could lead to the largest error due to robustness issues, it might be more interesting to

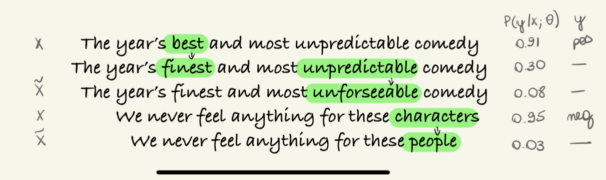

P(y|x;θ)
y
x
The year's best and most unpredictable comedy
0.91
pos
The year's finest and most unpredictable comedy
0.30
—
x̃
The year's finest and most unforseeable comedy
0.08
—
x
We never feel anything for these characters
0.95
neg
x̃
We never feel anything for these people
0.03
—

Figure 2.4: Hotflip visualization [5].

find out that modifying a word that changes the gender or the racial aspect of the input also results in a label flip [5].

2.3.3 Influential Examples

Influential examples explanations involve identifying examples from the training dataset that significantly influenced the model's predictions. These explanations are commonly utilized to uncover mislabeled observations. Among the various categories, influential example explanations stand out due to their non-trivial but suitable functionally-grounded metric, known as the label-correction experiment. This metric is relatively consistent across articles.

However, this experiment has not been extensively applied to NLP tasks, and in general, there has been limited functionally-grounded validation in NLP research. As a result, there is a need for more rigorous and extensive evaluation of functionally-grounded explanations in NLP to enhance interpretability and ensure model reliability.

2.3.4 Counterfactuals

Counterfactual explanations share many similarities with adversarial explanations, and sometimes these terms are confused in some works. The critical difference is that adversarial examples should have the same goal as the original example, while counterfactual examples should have a different gold label (often opposite) as the original example [54]. Counterfactuals essentially answer the question "how would the input need to change for the prediction to be different?" [5].

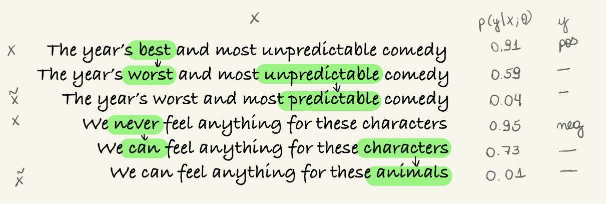

x
P(y|x;θ)
y
x
The year's best and most unpredictable comedy
0.91
pos
The year's worst and most unpredictable comedy
0.59
—
x̃
The year's worst and most predictable comedy
0.04
—
x
We never feel anything for these characters
0.95
neg
We can feel anything for these characters
0.73
—
x̃
We can feel anything for these animals
0.01
—

Figure 2.5: Counterfactual visualization [5].

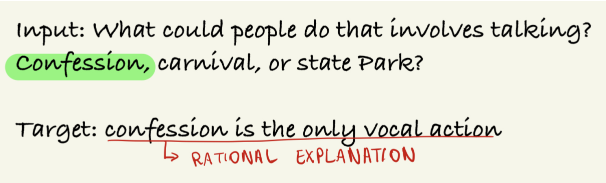
Figure 2.5 shows $\tilde{x}$ as a counterfactual from an original sentence x.

2.3.5 Natural Language

Explanation methods presented may be difficult to understand for people without specialized knowledge. Therefore, it might be interesting to directly generate an explanation in the form of natural language, which can be understood by simply reading the explanation for a given example.

Most research in the area of natural language explanation uses the explanations to improve the predictive performance of the model itself. The idea is that by enforcing the model to reason about its behavior, the model can generalize better [55] [42]. These approaches are, however, in the category of intrinsic methods.

Rationalization methods, however, are post-hoc methods in the sense that they attempt to explain after a prediction has been made [42]. Figure 2.6 shows a visualization of rationalizing Commonsense Auto-generated Explanations (CAGE) [5].



Input: What could people do that involves talking? Confession, carnival, or state Park?

Target: confession is the only vocal action

Figure 2.6: Rational explanation visualization [5].

This sub-field of natural natural language explanations has received criticism in NLP for not evaluating functionally-grounded [56]. This issue is even more problematic because the annotated explanations are provided by humans who have no insights into the model's behavior [57].

The explanation model therefore just learns about humans' thought processes rather than the model's logical process. This issue is somewhat unique to the NLP literature and is better treated in other fields [58].

2.4 Class explanations

2.4.1 Concepts

The term "concept" finds more frequent use in computer vision [59] [60] [61] than in NLP. Concept explanations attempts to explain how a model operates by using an abstraction of the input, called a concept. A classical example in computer visionis to explain how the concept of stripes affects the classification of a zebra. A computer vision model could classify a zebra based on a horse-like shape and a Savana background. While this may result in a high accuracy score, it is logically incorrect.

In NLP, the focus often revolves around bias detection. For instance, Vig et al. [62] utilize the concept of "occupation-words" like "nurse" and analyze its connection to the classification of pronouns like "he" and "she."

Regardless of the field, in both NLP and CV, only a single class or small subset of classes are analyzed. For this reason, concept explanation belong in its own category of class explanations.Vig et al. [62] applied Natural Indirect Effect to a small GPT-2 model, where the mediator is an attention head. By doing this, Vig et al. [62] can identify which attention heads are most responsible for the gender bias, when considering the occupation concept. By doing this, they can identify which attention heads are most responsible for the gender bias.

2.5 Global explanations

2.5.1 Vocabulary

Vocabulary explanation is a method that explains the whole model in relation to each word in the vocabulary, making it a global explanation. This approach is applicable to both sequence-to-class and sequence-to-sequence models. In the context of sentiment classification, words could be categorized as positive or negative, and identifying words that don't fit into either of these groups may reveal biases in the dataset.

In the field of NLP models, most research on vocabulary explanations is focused on these pre-trained word embeddings [63]. These pre-trained embeddings play a crucial role in the model's language understanding capabilities and contribute significantly to the explanation process.

2.5.2 Ensemble

Ensemble explanations aim to create a comprehensive global explanation by gathering multiple local explanations, each representing different aspects or modes of the model. The key challenge for ensemble explanations lies in strategically selecting representative examples and their corresponding local explanations. However, despite their potential significance, only a limited number of ensemble methods have been proposed in practice, with the majority being applicable exclusively to tabular data [64] [65].

The effectiveness of ensemble explanations heavily relies on the functionally-grounded nature of the local explanations.

2.5.3 Linguistic information

A widely adopted approach for validating the reasonability of a natural language model involves aligning it with the wealth of linguistic theory developed over centuries. Methods within this category employ two main strategies: behavioral probes (or behavioral analysis) and structural probes. They involve strategically modifying the model's input to observe its reaction and understand how it processes language and aim to establish alignment between a latent representation of the model and some linguistic representation, helping to reveal the model's underlying linguistic understanding.

The applicability of these strategies to specific types of models varies based on the method used. Generally, behavioral probes are applied to sequence-to-class models, while structural probes can be utilized for both sequence-to-class and sequence-to-sequence models.

One especially noteworthy subcategory of Structural Probes is BERTology which specifically focuses on explaining the BERT-like models [66] [67]. The research being done in behavioral probes, also called behavioral analysis, is not just for interpretability but also to measure the robustness and generalization ability of the model.

One of the pioneering articles that explores interpretability through behavioral probes is authored by Linzen et al. [68]. In their research, they investigate a language model's capacity to accurately reason about subject-verb agreement.

Clouatre et al. [69] and Sinha et al. [70] conducted studies where they examined the impact of destroying syntax by shuffling words on a Natural Language Inference (NLI) task. Their findings revealed that this manipulation did not significantly affect the model's performance indicating that the model does not achieve natural language understanding [5].

McCoy et al. [71] look at NLI, a task where a premise and a hypothesis are provided and the model should inform if these sentences are in agreement (called entailment), contradiction or neutral. For example:

- **premise:** The judge was paid by the actor
- **hypothesis:** The actor paid the judge

They propose three heuristics based on the linguistic properties: lexical overlap, subsequence, and constituent [71].

2.5.4 Rules

Rule explanations aim to provide a simple set of rules to explain complex models. However, due to the inherent complexity of these models, reducing them to concise rules is often impossible. Therefore, methods attempting rule explanations typically focus on explaining specific simplified aspects of the model rather than offering a complete representation.

Semantically Equivalent Adversaries Rules (SEAR) is an extension of the Semantically Equivalent Adversaries (SEA) method [72], where they developed a sampling algorithm for finding adversarial examples. Hence, the rule-generation objective is simplified, as only rules that describe what breaks the model needs to be generated. Because the category of rule explanations can be very diverse, groundedness evaluation would likely depend on the specific explanation method. However, generally, functionally groundedness can be measured by asserting if the rule holds true by evaluating it on the dataset and compare with the model response. Additionally, human-groundedness can be evaluated by asking humans to predict the model's output or choose the better model [5].

2.6 LIME

This research endeavors to explore the benefits of integrating XAI within the domain of public administration. It involves a qualitative assessment of XAI in the classification process of a complaint text classification model. The study aims to illuminate the practical implications and advantages of interpretable AI in a real-world setting, with a specific emphasis on post-hoc techniques.

For security and private reasons, evaluation will encompass a collection of synthetic text complaints adapted from actual cases, undergoing analysis through a widely used explainable algorithm known as LIME. Despite the availability of more recent explainable algorithms, LIME was selected due to its post-hoc characteristic, making it seamlessly integrable into the existing production process. Another notable advantage of LIME is its ability to generate human-friendly visualizations, which significantly aids the domain expert in evaluating the explanations.

In the context of the NLP model, LIME is employed to unveil the most influential words in each text complaint or local instance. This is achieved by approximating the complex model through the creation of a linear surrogate model, which enables an exploration of the words with the greatest influence on the local instance.

The algorithm samples words within the text, systematically masking them to investigate their contribution to the complaint's classification. The user has the flexibility to determine both the number of samples considered by the algorithm and the quantity of the most influential words presented in the visualization.

Local Interpretable Model-Agnostic Explanations (LIME) has the overall goal to identify an interpretable model over the interpretable representation that is locally faithful to the classifier [6]. The model to be explained is denoted by $f : R^d \to R$ and the explanation produced by LIME is obtained by the following:

$$\xi(x) = \underset{g \epsilon G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

On the second part of the equation, $\mathcal{L}(f, g, \pi_x) + \Omega(g)$ is the measure of how unfaithful the model $g \in G$ is, where G is a class of potentially interpretable models such as linear models or decision trees, in approximating f in the locality defined by π_x. In order to ensure both interpretability and local fidelity, we must minimize this measure while having the measure of complexity $\Omega(g)$ be low enough to be interpretable by human. [6].

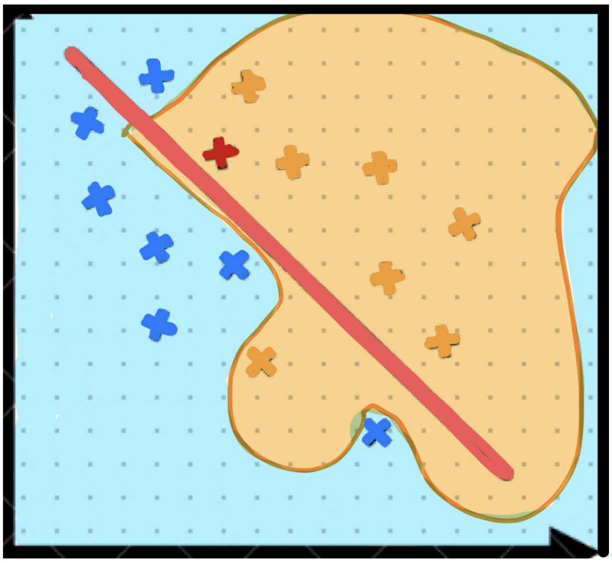


Figure 2.7: LIME [6].

In order to learn the local behavior of f as the interpretable inputs vary, we approximate $\mathcal{L}(f, g, \pi_x) + \Omega(g)$ by drawing samples, weighted by π_x. While the overall complexity of the original model may be too complex to explain globally, LIME provides locally faithful explanations [6]. Figure 2.7 represents an intuition of LIME where the black-box model's complex decision function f is represented by the blue/orange background and the bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained. The pink line is the learned explanation that is locally, but not globally, faithful.

First, the LIME algorithm perturbs the input to generate additional points with the model. This process creates neighborhood data by randomly masking features or words from the instance and then making predictions with the classifier. By default, LIME uses 'UNKWORDZ' as the mask for hidden features.

Feature selection in LIME can be conducted in various ways. In this study, the default parameter ’auto’ was used. With ’auto’, LIME employs forward_selection if the number of features is 6 or fewer, and ’highest_weights’ otherwise. Forward_selection iteratively adds features to the model, while ’highest_weights’ selects features based on the highest product of absolute weight and the original data point when using all features.

A Sklearn regressor is employed to explain the local point.

LIME uses cosine distance to compute the distances between the original and perturbed instances.

2.7 SHAP

A limitation of LIME is that the weights in a linear model may not be intrinsically interpretable. In situations where multicollinearity exists, for example when input features are linearly correlated with each other, the model weights can be scaled arbitrarily, leading to a misleading sense of importance [5].

To address this issue, one approach is to compute Shapley values [25] which are derived from game theory. The central idea involves fitting a linear model for every permutation of features [5].

While the Shapley value computation method may work in theory, it can be clearly intractable. However, Lundberg and Lee [49] have introduced a framework that offers a more tractable approach to produce Shapley values called Kernel SHAP. It combines the ideas of reducing the number of features by a mapping function $h_x(z)$, using squared-loss instead of cross-entropy via working on logits and it weighting each observation by how many features there are enabled.

SHAP and Shapley values in general are heavily used in the industry [73]. In NLP literature SHAP has been used by Wu et al.[74]. This popularity is likely due to their mathematical foundation and the shap library.

Chapter 3

Complaint Text Classification Model

Corruption has become one of the primary challenges for public administration and democracy in Brazil, a topic frequently discussed within Brazilian society. Regularly, as we watch the news, we are confronted with numerous cases of corruption at the federal, state, and municipal levels, involving both public and private figures. It is a social phenomenon that has severely eroded the public assets of the Brazilian state and significantly impaired the effectiveness and efficiency of public policies in the country.

The Brazilian legal system grants every citizen the opportunity to report irregularities occurring within the Public Administration. The Brazilian Office of the Comptroller General (CGU) is the internal control body of the Brazilian government responsible for defending public assets and for increasing transparency, through audit, internal affairs and corruption prevention and fighting [75].

At the federal level, CGU is responsible for receiving and handling complaints regarding public agents, organs, and entities of the Federal Executive Power [76]. A citizen can submit a complaint, suggestion, or request via the Fala.BR platform, where they have the option to attach various file types, including images, spreadsheets, and more. During the analysis of a complaint, a domain expert carefully reviews both the textual content and any attached files. They conduct searches in the organization's databases, if necessary. The final determination classifies the complaint as either 'Suitable' or 'Not Suitable' for further processing within the standard investigative procedures.

On average, only 30 percent of the incoming reports meet the criteria set by domain experts. This leads to a substantial allocation of effort by the team to address reports lacking the essential information for any investigation. The large volume of reports poses a formidable challenge, necessitating the involvement of a sizable team dedicated to the meticulous screening of these submissions [1].

In order for the complaints to be properly investigated, they must provide coherent information about the alleged incidents. As a result, the first step in handling the complaints is known as the suitability analysis, which aims to assess, based on all the received material, whether a complaint should be deemed eligible for further consideration or not.

During the assessment, the staff must read the text of the complaint and access and analyze each attached file. These attachments may come in various formats such as spreadsheets, images, presentations, text files, etc. [77].

After analyzing the complaint texts, the staff needs to verify the information reported in these documents against databases and corporate systems. Based on these analyses, they conclude whether the complaints are suitable for further investigation or not.

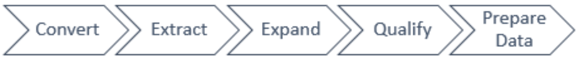


Figure 3.1: Steps in the Treatment of Complaint Texts.

3.1 Conversion

Figure 3.1 shows the steps previous to the auditor assessment of evaluation of the complaint. The first phase of the process is the Conversion phase. The main function of this phase is to access the attached files and convert them into an appropriate format for machine reading.

Therefore, during this phase, all the textual content of these files is transformed into plain text format, enabling their utilization in the subsequent stages of processing [77].

3.2 Extraction

The second phase involves Information Extraction from the texts. This methodology identifies and extracts a set of elements considered relevant for the task of handling complaints such as names of individuals, CPFs (Brazilian national identification numbers), CNPJs (Brazilian corporate identification numbers), contract numbers, agreement numbers, monetary values, etc. [77].

3.3 Expansion

The Expansion phase uses the entities identified in the previous phase and attempts to find new information about them in other databases. This search aims to validate the existence of the identified entities and discover new elements that are linked to the entities identified earlier.

3.4 Entity Qualification

The next phase of the process is dedicated to Entity Qualification, with the objective of verifying and categorizing the entities identified in the preceding stages. For instance, for a given CPF, the system checks if it belongs to a public servant or if the individual is a beneficiary of any social program, among other relevant criteria. Each type of entity goes through this qualification process, where specific sets of qualifiers are carefully examined and applied.

3.5 Data Preparation

Lastly, the final phase is Data Preparation. During this stage, all the information obtained in the previous phases is aggregated to create a structured dataset that can be used for model training [77].

The previously outlined phases are part of a automated process called *Ferramenta de Análise de Risco em Ouvidoria* (FARO) developed and presented by CGU as a part of the Anticorruption Plan. FARO is an innovative solution that leverages machine learning and natural processing languages (NLP) techniques, significantly aiding auditors in efficiently screening complaints.

Two distinct models have been devised based on prior complaints. The first model utilizes structured data acquired during the Expansion and Entity Qualification phases while the second model focuses on the actual text of the complaint.

This study will explore the second model in detail, aiming to investigate the seamless integration of LIME explainability into the complaint screening process.

Chapter 4

Formulation and Experiment

Research in the field of XAI has predominantly concentrated on the technical aspects of explanations, giving little attention to users' needs [78]. While explanations offer insights into model approximations, numerous articles and studies have highlighted limitations in these tools. Recent studies are making efforts to highlight the benefits derived from the synergy of explainable tools with human interactions. Nevertheless, there are few researches that analyze the application of explainable AI tools in real-world scenarios, especially when it comes to NLP models, with examples of real-world scenarios.

The effectiveness of explanations is intricately linked to users' perceptions. For this study, we will apply LIME algorithm for text considering the six most influential words. Furthermore, explanations and visualizations will be generated multiple times to scrutinize the consistency of the chosen influential words and their respective weights.

Given the need to adhere to regulations and uphold citizens' privacy regarding their actual complaints, this study will utilize newly generated complaints inspired by selected real-world cases. The selection process involves searching the data for situations where the integration of the explainability tool holds particular relevance in decision-making. This approach aims to illustrate the tool's role in the decision-making process and assess its impact on the task of screening complaints.

After generating the new texts of complaints and the corresponding explanations through LIME, the visualizations, along with the actual text, will be presented in the explanation module, including the classification score suggested by FARO's model. The auditor's task is to evaluate whether the complaint is deemed suitable or non-suitable. This evaluation mirrors real-world scenarios where the individual screening the complaint assigns a score ranging from 10 to 100, in increments of 10. This score, termed *Grau de Aptidão* or aptitude degree, reflects the auditor's judgment regarding the complaint's suitability for progressing through the regular investigation process. In cases where the complaint is considered non-suitable, its aptitude degree will be closer to 10. Conversely, if the complaint meets the prerequisites for suitability, its grade will approach 100.

The investigation will focus on proposing an explanation module that can be integrated to the actual software used presently in the process of screening complaints, where the auditor can delve into the details of the model's decision-making via LIME's visualization. By doing so, this research seeks to provide valuable insights and recommendations for improving the explainability and interpretability of machine learning models in the domain of Brazilian complaints.

4.1 Text-based Model

For the text-based model, the current approach for representing text as vectors employs the Term Frequency–Inverse Document Frequency (TF-IDF) technique. This method establishes the specificity of a term by directly correlating it with its frequency in the given document while inversely linking it to the term's commonality across a collection of documents. The selection of TF-IDF for the model was driven by its simplicity and minimal computational requirements at the time of development.

The Portuguese stopwords, such as 'de,' 'a,' 'o,' 'que,' and others were eliminated in the preparation of the data. Additionally, uniform lowercase formatting was applied to ensure consistent treatment of all words.

The initial dataset used to train the model consisted of 1489 labeled complaint records categorized as suitable or non-suitable. An 80/20 split was employed, with 80% of the complaints utilized for training the model and the remaining 20% reserved for validation. The proportion of records in the original dataset was maintained for both classes throughout the training and validation.

The chosen metric for assessing the model's performance was the Area Under the Receiver Operating Characteristic Curve (ROC AUC), where the XGBClassifier achieved a score of 0.83 [1].

4.2 Requirements of a Suitable Complaint

The Office of the General Ombudsman (OGU) does not impose a specific format for citizens to submit their complaints. However, certain essential attributes are expected to be present in a suitable complaint.

- Relate to a federal entity or involve concerns regarding public resources of federal origin.
- Describe an irregularity that signifies harm or poses a threat to public assets.

- Contain minimal justification to enable its investigation.

Additional details such as the citizen's full name, email, address, and any pertinent information enabling OGU to establish contact should be included. Also, any information the citizen wishes to keep confidential must be specified. The citizen can outline the steps taken to address the issue, and highlight any unresolved aspects. The complaint should clearly describe the situation, providing a comprehensive account of the facts and identifying individuals involved, when possible.

4.3 Complaints evaluation

This study proposes for the integration of an explainable tool into the process of screening complaints. Presently, auditors have access to diverse data pertaining to the complaint and the model, including the model's score, which ranges from zero to one. A higher score signifies the model's suggestion that the complaint should be classified as suitable.

In addition to the model's score, auditors can access derived data obtained during the extraction, expansion, and qualification phases. The abundance of information is currently not organized in an intuitive manner. The LIME explanation visualization has the potential to enhance user-friendliness in the interface by presenting this information in a more structured and accessible manner.

For the purposes of this study, complaints will be generated based on real-case scenarios. The actual complaints will not be disclosed in this work to uphold the confidentiality of sensitive data, including the citizen's identification reporting the irregularity, entities involved in the report, etc.

Following the selection and generation of the complaints, explainable visualizations of the text will be created using the LIME tool. These visualizations will consist of the most influential words identified by the surrogate model that LIME establishes to approximate the local classification. In addition to showcasing the most influential words, their respective weights will also be presented. These words can carry either positive or negative meanings, influencing the classification for both classes.

Chapter 5

Results

5.1 Explanation Module

As one of the objectives of this study, a software module called the 'explanation module' is presented. This module provides visualizations of local explanations for each complaint text selected by the user. It aims to serve as a tool for auditors to generate insights that can improve the text classification model of complaints, enhance understanding of some scores suggested by the model, and assist in the decision-making process for screening complaints.

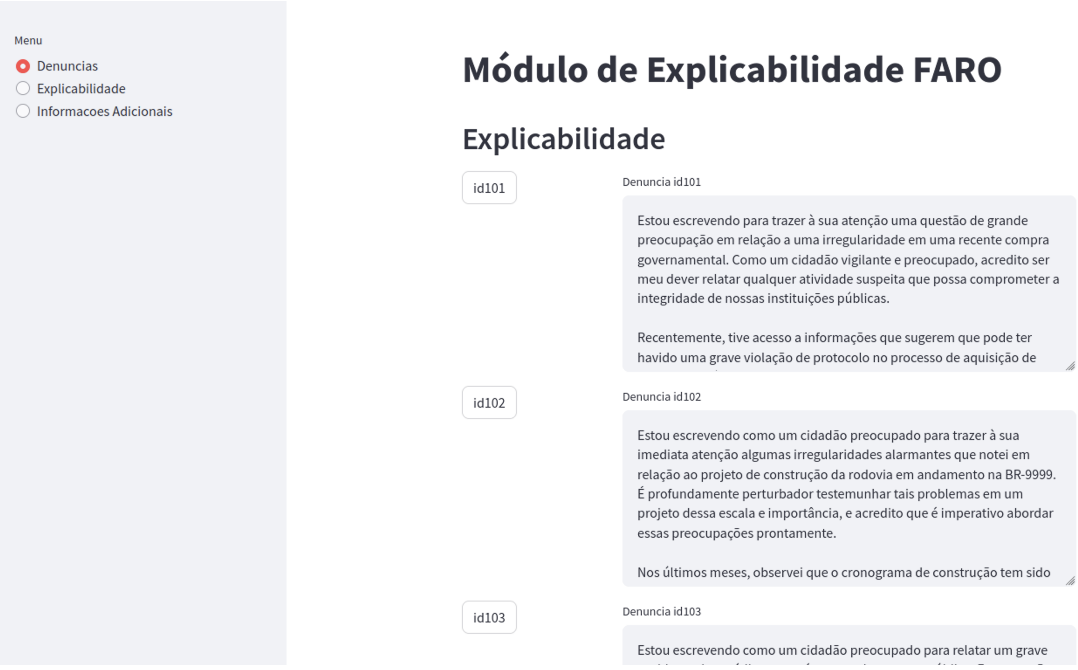
By integrating this module into the existing complaint screening process, auditors will gain a valuable tool for monitoring the text classification model and exploring specific complaint scores generated by the model. One significant advantage is that the model's explainability will become accessible to a broader range of users, as there will be no need for extensive knowledge of software development to understand the explanations generated by the LIME algorithm.

The explanation module was developed using the Python programming language due to its simplicity and support for various machine learning libraries. Additionally, the existing F.A.R.O. process was also developed in Python. For the front-end development, the Streamlit framework was utilized, and the LIME algorithm was employed to provide local explanations.

The explanation module is organized as follows:

- *Start screen (Home)* : The first part of the explanation module displays a list of the complaint texts which can be selected by the user. Each complaint will have the score suggested by the text model of F.A.R.O. next to its text.

[0] https://github.com/stellameireles/modulo_explicabilidade



Módulo de Explicabilidade FARO

Explicabilidade

Denuncia id101

Estou escrevendo para trazer à sua atenção uma questão de grande preocupação em relação a uma irregularidade em uma recente compra governamental. Como um cidadão vigilante e preocupado, acredito ser meu dever relatar qualquer atividade suspeita que possa comprometer a integridade de nossas instituições públicas.

Recentemente, tive acesso a informações que sugerem que pode ter havido uma grave violação de protocolo no processo de aquisição de

Denuncia id102

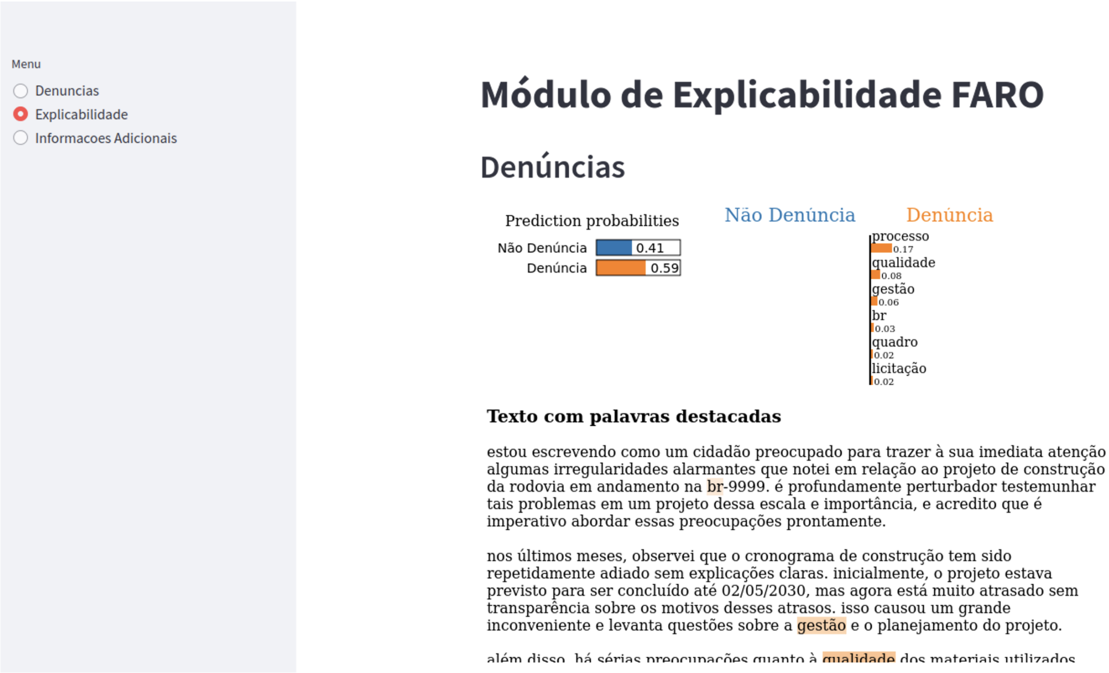
Estou escrevendo como um cidadão preocupado para trazer à sua imediata atenção algumas irregularidades alarmantes que notei em relação ao projeto de construção da rodovia em andamento na BR-9999. É profundamente perturbador testemunhar tais problemas em um projeto dessa escala e importância, e acredito que é imperativo abordar essas preocupações prontamente.

Nos últimos meses, observei que o cronograma de construção tem sido

Denuncia id103

Estou escrevendo como um cidadão preocupado para relatar um grave

Figure 5.1: Start screen.



Módulo de Explicabilidade FARO

Denúncias

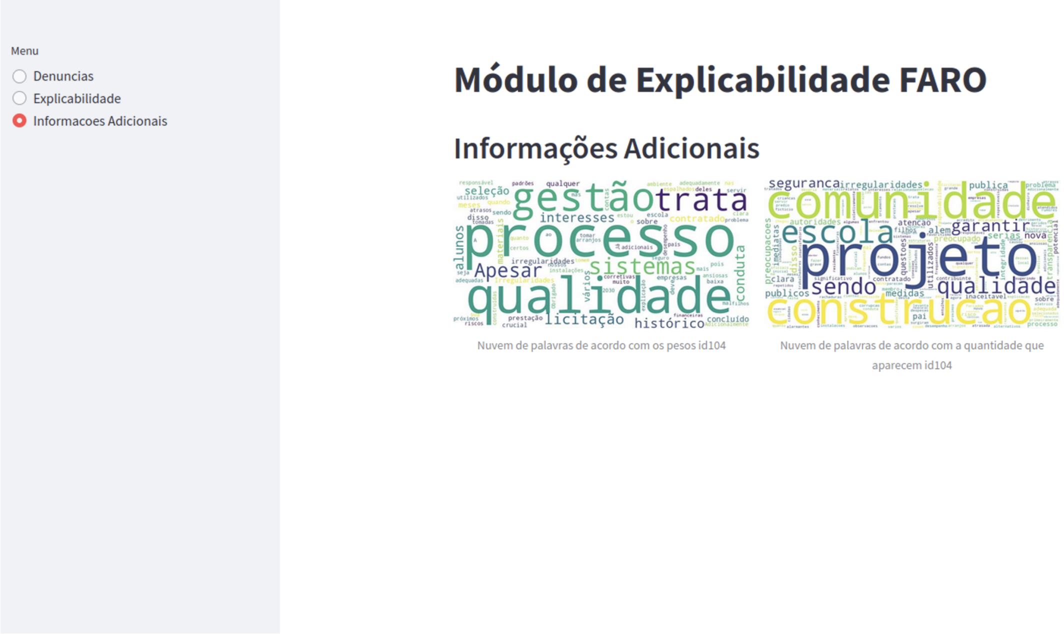
Texto com palavras destacadas

estou escrevendo como um cidadão preocupado para trazer à sua imediata atenção algumas irregularidades alarmantes que notei em relação ao projeto de construção da rodovia em andamento na br-9999. é profundamente perturbador testemunhar tais problemas em um projeto dessa escala e importância, e acredito que é imperativo abordar essas preocupações prontamente.

nos últimos meses, observei que o cronograma de construção tem sido repetidamente adiado sem explicações claras. inicialmente, o projeto estava previsto para ser concluído até 02/05/2030, mas agora está muito atrasado sem transparência sobre os motivos desses atrasos. isso causou um grande inconveniente e levanta questões sobre a gestão e o planejamento do projeto.

além disso, há sérias preocupações quanto à qualidade dos materiais utilizados

Figure 5.2: LIME.



Menu

○ Denuncias

○ Explicabilidade

● Informacoes Adicionais

Módulo de Explicabilidade FARO

Informações Adicionais

Nuvem de palavras de acordo com os pesos id104

Nuvem de palavras de acordo com a quantidade que aparecem id104

Figure 5.3: Visualization.

- *Second screen (LIME)* : After selecting a complaint, the user is taken to the detailed explanation generated by the LIME algorithm. In this section, the most influential words for both classes (suitable complaint and non-suitable complaint) are displayed, along with their weights. The complete text of the complaint is also shown, with influential words highlighted according to the weights.
- *Third screen (Visualization)*: The third screen displays a word cloud with the words of the complaint text. This is a more intuitive and familiar visualization that gives an idea of the most frequent words in the document, alongside with the most influential words, as the LIME algorithm works by perturbing in the model's input.

The described sections were chosen for their relevance and utility to the user. Additional screens and functionalities may be incorporated in future updates to enhance the module further.

5.2 Software Architecture

To propose an explainability module tailored to the needs of auditors, this study analyzed the current software architecture used by the F.A.R.O. process. The existing process regularly handles batches of complaints, processing them to generate suggested scores that are stored in a database.

Figure 5.4 illustrates the current organization of the complaint screening process. Each day, citizens submit complaints, which are then stored in a dedicated complaints database.

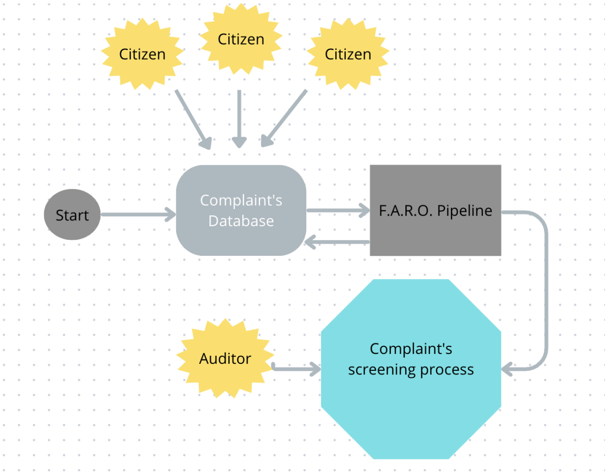


Figure 5.4: Current Pipeline.

The texts of the complaints and any attachments undergo preprocessing, including steps such as conversion, extraction, expansion, entity qualification, and data preparation as previously outlined. Following these preprocessing steps, the complaint classification model assigns a score, and the Complaints Database is updated with the scores and other relevant information generated throughout the pipeline.

Integrating explainability into the day-to-day software can be challenging, as some algorithms require a considerable amount of time to generate explanations. One solution is to process the explanations separately and store them. Once processed, these explanations can be displayed to the user without delay. Considering this, the following organization of the software is proposed to integrate the explanation module into the existing pipeline.

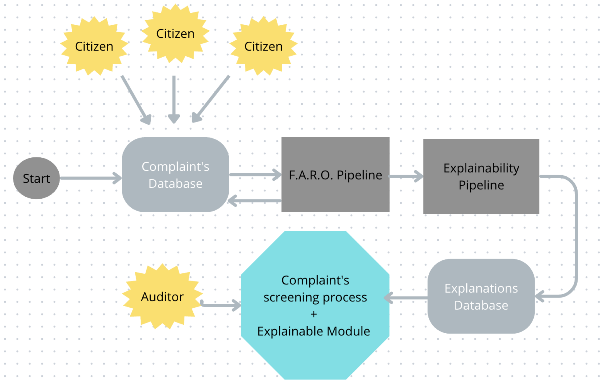


Figure 5.5: Proposed Pipeline.

In the new proposed pipeline, there is an extra step called explanation pipeline that uses the complaint's batches and the FARO model to generate LIME explanations and visualizations. These explanations are stored in a database to be used afterwards in the explanation module.

In this study, 8,393 complaints were analyzed, and their corresponding LIME explanations were generated. The LIME parameters used were 2,500 for the *samples* and 6 for the *features*, indicating that 2,500 perturbation samples were generated and the six most influential words were identified for each explanation. On average, generating explanations for a batch of 200 complaints took 212 minutes, resulting in over a minute per explanation.

Given the substantial time required to produce each explanation, the decision to process them in batches was made. This approach mitigates the potential disruption to users who would otherwise face long waits to view individual explanations. Additionally, users often need to examine multiple explanations to investigate potential biases and gain insights into the model, further justifying the batch processing method.

5.3 Explanations

Quantifying or measuring the quality of explainability, particularly in natural language processing cases, remains a challenging task. Although various frameworks have been proposed [79] to address this issue, there is no consensus in academia on a standard approach, nor is there a widely accepted framework deemed relevant for this study.

Some studies conduct qualitative evaluations of the most influential words generated by LIME explanations, analyzing relevant examples and exploring the semantics of the words. This may involve the assistance of an expert with deep knowledge of the business process, as seen in the analysis of the predicting patient admissions at the emergency department using triage notes [80]. Similarly, in this work, unique and relevant examples of influential words are highlighted and analyzed to provide deeper insights.

In this study, the six most influential words were identified for each of the 8,393 complaints. These words were categorized into positive words, which are associated with the class of suitable complaints, and negative words, which are related to the class of non-suitable complaints.

Some of the most influential words are non-surprisingly related to the complaint concept, such as politics, managment, mayor, public, enterprise. These words can easily be placed when someone is repoprting an irregularity. This positive words were compared with a private list of words selected by domain expert auditors that, accordding to their vision, indicate that the document containing this words is indeed a suitable complaint. These words are used in another part of the process of screening complaints, out of scope of this study.

It is possible to notice some semantical similarity between the positive words in the LIME explanations and the words selected by the expert. The positive words that are not present in the private list were presented to be added in the list, aiming to improve the list.

Surprisingly, some words that are not semantically related to the complaint context were listed as the most influential words in the LIME explanations. Examples include words such as "yet," "avenue," "telephone," and a common personal name. This finding does not necessarily indicate that the model is malfunctioning, as the relevance of each word in a document can be minimal. However, it can suggest that an additional preprocessing step, such as deleting or altering proper names, could be beneficial.

Another potential issue in the explanations is the presence of grammatically incorrect words within the texts. These words may not be properly considered, affecting the accuracy and interpretability of the explanations.

Chapter 6

Conclusions

Interpretable and explainable machine learning is a dynamic and evolving research field. With the recent significant progress in creating high-performance predictive models and the widespread integration of machine learning across various domains, the impact of algorithmic decision-making is profound. It is crucial for these algorithms to be comprehensible and trustworthy to human end-users [33], specially in some critical areas like medicine, healthcare, credit score and others.

This subject holds significant societal relevance as it contributes to improving public services and building trust among citizens regarding the use of AI models.

Interpretable and explainable machine learning could significantly benefit from the adoption of improved empirical research practices, as is often the case with developing research areas, as many works still rely on purely qualitative or even anecdotal evidence [33].

Meaningful adaptations of the discussed methods to real-world machine learning systems and data analysis problems are mostly yet to be explored, representing a significant focus for the future. In order to facilitate widespread and effective utilization of interpretable and explainable ML techniques, it is crucial to involve stakeholders in meaningful discussions.

This research offers valuable insights to the field by analyzing interpretability methods in a real-world scenario within a specific business context. The findings and conclusions of this work are expected to contribute to future research efforts, fostering the construction of a more robust knowledge base in the field of explainable AI. By shedding light on the effectiveness and applicability of these methods in practical settings, this study serves as a stepping stone towards better understanding and implementing explainable AI solutions in various domains.

Referências

[1] *Relatorio. controladoria geral da uniao.* https://repositorio.cgu.gov.br/bitstream/1/69535/1/Relatorio_Gerencial_Acao_CGU_35_Planoanticorrupcao.pdf (Accessed Nov. 14, 2023). vii, viii, 3, 22, 26

[2] Gao, Lei and Ling Guan: *Interpretability of machine learning: Recent advances and future prospects.* IEEE MultiMedia, 2023. xi, 7

[3] Meske, Christian, Enrico Bunde, Johannes Schneider, and Martin Gersch: *Explainable artificial intelligence: objectives, stakeholders, and future research opportunities.* Information Systems Management, 39(1):53–63, 2022. xi, 8

[4] Du, Mengnan, Ninghao Liu, and Xia Hu: *Techniques for interpretable machine learning.* Communications of the ACM, 63(1):68–77, 2019. xi, 11, 12

[5] Madsen, Andreas, Siva Reddy, and Sarath Chandar: *Post-hoc interpretability for neural nlp: A survey.* ACM Computing Surveys, 55(8):1–42, 2022. xi, 9, 10, 12, 13, 14, 15, 16, 18, 19, 21

[6] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin: *" 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*, pages 1135–1144, 2016. xi, 3, 10, 11, 12, 13, 20

[7] McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, *et al.*: *International evaluation of an ai system for breast cancer screening.* Nature, 577(7788):89–94, 2020. 1

[8] Gulshan, Varun, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, *et al.*: *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs.* jama, 316(22):2402–2410, 2016. 1

[9] Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang: *Artificial intelligence in healthcare: past, present and future.* Stroke and vascular neurology, 2(4), 2017. 1

[10] Brown, Michael PS, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S Furey, Manuel Ares Jr, and David Haussler: *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proceedings of the National Academy of Sciences, 97(1):262–267, 2000. 1

[11] Peres, Ricardo Silva, Xiaodong Jia, Jay Lee, Keyi Sun, Armando Walter Colombo, and Jose Barata: *Industrial artificial intelligence in industry 4.0-systematic review, challenges and outlook.* IEEE Access, 8:220121–220139, 2020. 1

[12] Verma, Sanjeev, Rohit Sharma, Subhamay Deb, and Debojit Maitra: *Artificial intelligence in marketing: Systematic review and future research direction.* International Journal of Information Management Data Insights, 1(1):100002, 2021. 1

[13] Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans: *When will ai exceed human performance? evidence from ai experts.* Journal of Artificial Intelligence Research, 62:729–754, 2018. 1

[14] Mirbabaie, Milad, Alfred B Brendel, and Lennart Hofeditz: *Ethics and ai in information systems research.* Communications of the Association for Information Systems, 50(1):38, 2022. 1

[15] Robert, Lionel, Gaurav Bansal, and Christoph Lutge: *Icis 2019 sighci workshop panel report: Human computer interaction challenges and opportunities for fair, trustworthy and ethical artificial intelligence.* 2020. 1, 6

[16] *Gonzales, guadalupe. how amazon accidentally invented a sexist hiring algorithm a company experiment to use artificial intelligence in hiring inadvertently favored male candidates.* https://www.inc.com/guadalupe-gonzalez/amazon-artificial-intelligence-ai-hiring-tool-hr.html(Accessed Aug. 22, 2023). 1

[17] Buolamwini, J: *Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers (master of science). massachusetts institute of technology, cambridge, ma*, 2017. 1

[18] Yampolskiy, Roman V: *Predicting future ai failures from historic examples.* foresight, 21(1):138–152, 2019. 1

[19] Dignum, Virginia: *Responsible artificial intelligence–from principles to practice.* arXiv preprint arXiv:2205.10785, 2022. 2, 3

[20] Brasse, Julia, Hanna Rebecca Broder, Maximilian Förster, Mathias Klier, and Irina Sigler: *Explainable artificial intelligence in information systems: A review of the status quo and future research directions.* Electronic Markets, 33(1):26, 2023. 2, 6

[21] Zhang, Yujia, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell: *" why should you trust my explanation?" understanding uncertainty in lime explanations.* arXiv preprint arXiv:1904.12991, 2019. 2

[22] Ali, Sajid, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera: *Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence.* Information Fusion, 99:101805, 2023. 2

[23] Mill, Eleanor Ruth, Wolfgang Garn, Nicholas F Ryman-Tubb, and Christopher Turner: *The sage framework for explaining context in explainable artificial intelligence.* Applied Artificial Intelligence, 2023. 2

[24] *Novas tecnologias e tributação.* https://www.revista.ibdt.org.br/index.php/RDTIAtual/article/vie (Accessed Nov. 21, 2024). 3

[25] Shapley, Lloyd S *et al.*: *A value for n-person games.* 1953. 3, 21

[26] Zeineldin, Ramy A, Mohamed E Karar, Ziad Elshaer, · Jan Coburger, Christian R Wirtz, Oliver Burgert, and Franziska Mathis-Ullrich: *Explainability of deep neural networks for mri analysis of brain tumors.* International journal of computer assisted radiology and surgery, 17(9):1673–1683, 2022. 3

[27] Song, Di, Jincao Yao, Yitao Jiang, Siyuan Shi, Chen Cui, Liping Wang, Lijing Wang, Huaiyu Wu, Hongtian Tian, Xiuqin Ye, *et al.*: *A new xai framework with feature explainability for tumors decision-making in ultrasound data: comparing with grad-cam.* Computer Methods and Programs in Biomedicine, 235:107527, 2023. 3

[28] Dignum, Virginia: *Responsible artificial intelligence: how to develop and use AI in a responsible way*, volume 2156. Springer, 2019. 6

[29] Rudin, Cynthia: *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature machine intelligence, 1(5):206–215, 2019. 6, 7

[30] Doshi-Velez, Finale and Been Kim: *Towards a rigorous science of interpretable machine learning.* arXiv preprint arXiv:1702.08608, 2017. 6, 7, 9, 10, 14

[31] Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, *et al.*: *Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai.* Information fusion, 58:82–115, 2020. 6

[32] Herse, Sarita, Jonathan Vitale, Meg Tonkin, Daniel Ebrahimian, Suman Ojha, Benjamin Johnston, William Judge, and Mary Anne Williams: *Do you trust me, blindly? factors influencing trust towards a robot recommender system.* In *2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 7–14. IEEE, 2018. 6

[33] Marcinkevičs, Ričards and Julia E Vogt: *Interpretable and explainable machine learning: A methods-centric overview with concrete examples.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, page e1493, 2023. 6, 34

[34] *Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.* https://eur- lex.europa.e/legalcontent/EN/TXT/HTML/? uri= CELEX: 52021 PC020 6 from= EN (Accessed Aug. 22, 2023). 7

[35] *Lei geral de proteção de dados pessoais.* https://www.planalto.gov.br/ccivil_03_ato2015-2018/2018/lei/l13709.htm (Accessed Aug. 22, 2023). 7

[36] Lipton, Zachary C: *The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.* Queue, 16(3):31–57, 2018. 7, 8

[37] Carvalho, Diogo V, Eduardo M Pereira, and Jaime S Cardoso: *Machine learning interpretability: A survey on methods and metrics.* Electronics, 8(8):832, 2019. 7, 8, 9

[38] Marcinkevičs, Ričards and Julia E Vogt: *Interpretability and explainability: A machine learning zoo mini-tour.* arXiv preprint arXiv:2012.01805, 2020. 7, 8

[39] Kouw, Wouter M and Marco Loog: *An introduction to domain adaptation and transfer learning.* arXiv preprint arXiv:1812.11806, 2018. 8

[40] Lou, Yin, Rich Caruana, and Johannes Gehrke: *Intelligible models for classification and regression.* In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158, 2012. 8

[41] Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis: *Explainable ai: A review of machine learning interpretability methods.* Entropy, 23(1):18, 2020. 8, 12

[42] Rajani, Nazneen Fatema, Bryan McCann, Caiming Xiong, and Richard Socher: *Explain yourself! leveraging language models for commonsense reasoning.* arXiv preprint arXiv:1906.02361, 2019. 10, 16

[43] Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei: *Reading tea leaves: How humans interpret topic models.* Advances in neural information processing systems, 22, 2009. 10

[44] Jacovi, Alon and Yoav Goldberg: *Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?* arXiv preprint arXiv:2004.03685, 2020. 11

[45] Wiegreffe, Sarah and Yuval Pinter: *Attention is not not explanation.* arXiv preprint arXiv:1908.04626, 2019. 11

[46] Hooker, Sara, Dumitru Erhan, Pieter Jan Kindermans, and Been Kim: *A benchmark for interpretability methods in deep neural networks.* Advances in neural information processing systems, 32, 2019. 11

[47] Madsen, Andreas, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy: *Evaluating the faithfulness of importance measures in nlp by recursively masking allegedly important tokens and retraining.* arXiv preprint arXiv:2110.08412, 2021. 11

[48] Du, Mengnan, Ninghao Liu, and Xia Hu: *Techniques for interpretable machine learning.* Communications of the ACM, 63(1):68–77, 2019. 11

[49] Lundberg, Scott M and Su In Lee: *A unified approach to interpreting model predictions.* Advances in neural information processing systems, 30, 2017. 12, 21

[50] Knight, Kevin, Ani Nenkova, and Owen Rambow: *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies.* In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016. 13

[51] Tang, Zhiyuan, Ying Shi, Dong Wang, Yang Feng, and Shiyue Zhang: *Memory visualization for gated recurrent neural networks in speech recognition.* In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2736–2740. IEEE, 2017. 13

[52] Belinkov, Yonatan and James Glass: *Analysis methods in neural language processing: A survey.* Transactions of the Association for Computational Linguistics, 7:49–72, 2019. 14

[53] Ebrahimi, Javid, Anyi Rao, Daniel Lowd, and Dejing Dou: *Hotflip: White-box adversarial examples for text classification.* arXiv preprint arXiv:1712.06751, 2017. 14

[54] Ross, Alexis, Ana Marasović, and Matthew E Peters: *Explaining nlp models via minimal contrastive editing (mice).* arXiv preprint arXiv:2012.13985, 2020. 15

[55] Latcinnik, Veronica and Jonathan Berant: *Explaining question answering models through text generation.* arXiv preprint arXiv:2004.05569, 2020. 16

[56] Hase, Peter, Shiyue Zhang, Harry Xie, and Mohit Bansal: *Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language?* arXiv preprint arXiv:2010.04119, 2020. 16

[57] Wiegreffe, Sarah and Ana Marasović: *Teach me to explain: A review of datasets for explainable natural language processing.* arXiv preprint arXiv:2102.12060, 2021. 16

[58] Andreas, Jacob, Anca Dragan, and Dan Klein: *Translating neuralese.* arXiv preprint arXiv:1704.06960, 2017. 16

[59] Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, *et al.*: *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).* In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 16

[60] Goyal, Yash, Amir Feder, Uri Shalit, and Been Kim: *Explaining classifiers with causal concept effect (cace).* arXiv preprint arXiv:1907.07165, 2019. 16

[61] Mu, Jesse and Jacob Andreas: *Compositional explanations of neurons.* Advances in Neural Information Processing Systems, 33:17153–17163, 2020. 16

[62] Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber: *Investigating gender bias in language models using causal mediation analysis.* Advances in neural information processing systems, 33:12388–12401, 2020. 17

[63] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean: *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781, 2013. 17

[64] Natesan Ramamurthy, Karthikeyan, Bhanukiran Vinzamuri, Yunfeng Zhang, and Amit Dhurandhar: *Model agnostic multilevel explanations.* Advances in neural information processing systems, 33:5968–5979, 2020. 17

[65] Sangroya, Amit, Mouli Rastogi, C Anantaram, and Lovekesh Vig: *Guided-lime: Structured sampling based hybrid approach towards explaining blackbox machine learning models.* In *CIKM (Workshops)*, 2020. 17

[66] Devlin, Jacob, Ming Wei Chang, Kenton Lee, and Kristina Toutanova: *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018. 18

[67] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, *et al.*: *Language models are few-shot learners.* Advances in neural information processing systems, 33:1877–1901, 2020. 18

[68] Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg: *Assessing the ability of lstms to learn syntax-sensitive dependencies.* Transactions of the Association for Computational Linguistics, 4:521–535, 2016. 18

[69] Clouatre, Louis, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar: *Local structure matters most: Perturbation study in nlu.* arXiv preprint arXiv:2107.13955, 2021. 18

[70] Sinha, Koustuv, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams: *Unnatural language inference.* arXiv preprint arXiv:2101.00010, 2020. 18

[71] McCoy, R Thomas, Ellie Pavlick, and Tal Linzen: *Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.* arXiv preprint arXiv:1902.01007, 2019. 18

[72] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin: *Semantically equivalent adversarial rules for debugging nlp models.* In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 856–865, 2018. 19

[73] Bhatt, Umang, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley: *Explainable machine learning in deployment.* In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657, 2020. 21

[74] Wu, Tongshuang, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld: *Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models.* arXiv preprint arXiv:2101.00288, 2021. 21

[75] *Controladoria geral da união.* https://www.gov.br/cgu/pt-br/centrais-de-conteudo/publicacoes/institucionais/arquivos/portifolio-ingles.pdf (Accessed Nov. 14, 2023). 22

[76] Paiva, Eduardo de, Fernando Sola Pereira, and Nelson Ebecken: *Sumarização de denúncias: Proposta e avaliação de métodos de geração de resumos.* In *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, pages 121–132. SBC, 2022. 22

[77] Paiva, Eduardo de and Nelson Ebecken: *Ferramenta para classificaçao de denuncias: Uma abordagem baseada em textos e dados estruturados.* In *Anais Estendidos do XVIII Simpósio Brasileiro de Sistemas de Informação*, pages 83–90. SBC, 2022. 23, 24

[78] Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin.: *Model-agnostic interpretability of machine learning.* 2016. 25

[79] Liu, Hui, Qingyu Yin, and William Yang Wang: *Towards explainable nlp: A generative explanation framework for text classification.* arXiv preprint arXiv:1811.00196, 2018. 32

[80] Arnaud, Emilien, Mahmoud Elbattah, Pedro A Moreno-Sánchez, Gilles Dequen, and Daniel Aiham Ghazali: *Explainable nlp model for predicting patient admissions at emergency department using triage notes.* In *2023 IEEE International Conference on Big Data (BigData)*, pages 4843–4847. IEEE, 2023. 32