WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

EFMD
EQUIS
ACCREDITED

Master Thesis

SHAP in Fake News Detection: Assessing Correctness, Output Completeness, and Continuity

Elena Guettl

Subject Area: Information Business

Studienkennzahl: h1605253

Supervisor: Dr. Sabrina Kirrane

Date of Submission: 17.09.2023

Department of Information Systems and Operations, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

**DEPARTMENT FÜR INFORMATIONS-
VERARBEITUNG UND PROZESS-
MANAGEMENT** DEPARTMENT
OF INFORMATION SYSTEMS AND
OPERATIONS

Contents

List of Figures

List of Tables

Acronyms

AI Artifical Intelligence.

AOPC Area over the Perturbation Curve.

BERT Bidirectional Encoder Representations for Transformers.

DARPA U.S. Defense Advanced Research Projects Agency.

DNN Deep Neural Network.

GAN Generative Adversarial Network.

GDPR General Data Protection Regulation.

IMDB International Movie Database.

LIME Local Interpretable Model-Agnostic Explanations.

MLM Masked Language Model.

NLP Natural Language Processing.

NSP Next Sentence Prediction.

ROAR Remove and Retrain.

SHAP SHapley Additive exPlanations.

XAI eXplainable Artificial Intelligence.

Abstract

This thesis evaluates the eXplainable Artificial Intelligence (XAI) method SHapley Additive exPlanations (SHAP) for interpreting fake news detection using the Bidirectional Encoder Representations for Transformers (BERT) model. The focus lies on evaluating SHAP's explanations on two fake news datasets based on three Co-12 criteria, which are criteria introduced to categorize the evaluation of XAI methods: correctness, output completeness, and continuity. Results indicate that SHAP provides correct, output complete, and continuous explanations for the fake news datasets. Future work is suggested to explore multiple models, explainability methods, and diverse datasets, as well as human-grounded and application-grounded evaluation approaches to enhance interpretability and thereby trust in AI systems.

1 Introduction

The impact of fake news on both our lives and global affairs has been vividly demonstrated by recent events such as the 2016 U.S. presidential elections [20], the COVID-19 'infodemic' [1], and the dissemination of disinformation during the Russia-Ukraine war [2]. Although misinformation in the media is not a novel occurrence, the emergence of social media platforms [135] alongside the rise in AI-generated fake news [144] has escalated its reach to unprecedented levels.

The sheer volume of fake news has rendered traditional detection methods, relying on manual verification and journalistic practices, inadequate to keep pace with the rapid dissemination of such content [13]. Understanding the significant role fake news can play in misleading the public and potentially influencing democratic processes and social cohesion emphasizes the necessity of automated fake news detection [156]. It is notable that fake news tends to spread at a faster rate than genuine news, heightening the urgency of effective detection mechanisms [140].

While numerous approaches have shown promising outcomes in terms of the accuracy of fake news classification [7], an important question arises: How can one place trust in such classifications? A study conducted by Ribeiro et al. [116] demonstrates that even a deliberately flawed classifier, which only uses the presence of snow in the background to categorize images of wolves and huskies, led one third of the test subjects (graduate students with at least one machine learning class) to express trust in the model for real-life applications. This indicates that even individuals with some familiarity with the subject might trust an inadequate classifier, highlighting the limited transparency surrounding the behavior of such models. This black box nature of most Artifical Intelligence (AI) models results in a lack of understanding regarding their decision-making processes [24].

Addressing this challenge, the field of eXplainable Artificial Intelligence (XAI) aims to shed light on the inner workings of these complex models. This becomes especially crucial when considering the interpretation of AI outcomes in sensitive domains, such as fake news detection, healthcare, or the justice system [40]. By introducing methodologies that provide comprehensible explanations for AI model decisions, eXplainable Artificial Intelligence (XAI) bridges the gap between the inherent complexity of these models and the need for understandable and trustworthy outcomes.

1.1 Research Questions

To explore the effectiveness of such XAI methodologies in the context of fake news detection, this research work focuses on evaluating the SHapley Additive exPlanations (SHAP) explainability method. This framework, proposed by Lundberg and Lee [88], offers a local and model-agnostic explanation method. The aim is to assess SHAP's explanations of fake news text data, specifically addressing its correctness, output completeness, and continuity. By investigating these aspects, we aim to contribute to the understanding of how well SHAP functions in the context of interpreting Bidirectional Encoder Representations for Transformers (BERT)-based fake news detection.

With this overarching goal, the following research questions have been formulated:

RQ How can we effectively evaluate the suitability of SHAP for interpreting text data, particularly in the context of fake news detection using BERT, with a focus on assessing correctness, output completeness, and continuity, and what are the results of these evaluations?

To address the overarching research questions, we determine the following sub-questions:

SQ1 To what extent does SHAP demonstrate correctness, align with the original model's behavior in terms of faithfulness, when applied to BERT's fake news detection?

SQ2 How does SHAP perform in terms of output completeness, providing sufficient information to explain the output of the model when explaining fake news detection by BERT?

SQ3 How well does SHAP demonstrate continuity, indicating the generalizability of its findings, for BERT's fake news detection?

Through this investigation, we aim to contribute to the development of more transparent and trustworthy AI systems, particularly in critical areas like fake news detection.

1.2 Design Science Methodology

In this thesis, we evaluate the explainability method SHAP on its correctness, output completeness, and continuity for two fake news datasets. As our methodology, we use Design Science Research [63]. Design Science Research has as one of its goals "*to generate knowledge on how things can and should be constructed or arranged*" [19]. The methodology design process depicted in

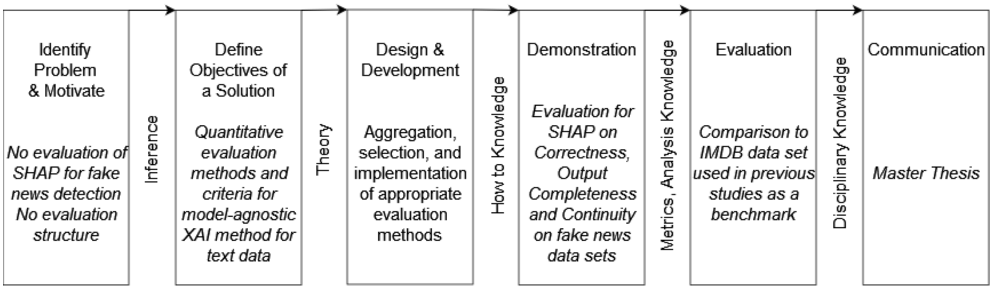

Identify Problem & Motivate
No evaluation of SHAP for fake news detection No evaluation structure
Inference
Define Objectives of a Solution
Quantitative evaluation methods and criteria for model-agnostic XAI method for text data
Theory
Design & Development
Aggregation, selection, and implementation of appropriate evaluation methods
How to Knowledge
Demonstration
Evaluation for SHAP on Correctness, Output Completeness and Continuity on fake news data sets
Metrics, Analysis Knowledge
Evaluation
Comparison to IMDB data set used in previous studies as a benchmark
Disciplinary Knowledge
Communication
Master Thesis

Figure 1: Design Science Research Process

Figure 1 has been adapted from Peffers et al.'s [106] design science research methodology process model, which is the most widely referenced process model [19]. This adapted model serves as the framework to effectively address our research questions.

First, we identify the problem. The evaluation of explainability methods on fake news data has yet to be explored widely. In general, there is currently no benchmark for the evaluation of XAI methods, irrespective of data type [21, 40, 47, 62, 66, 86]. This is a problem because it makes it difficult to compare different XAI methods and to assess their effectiveness. The objective of this thesis is to identify appropriate quantitative evaluation methods that can be used to evaluate model-agnostic XAI methods on fake news detection methods. The goal is to demonstrate the evaluation and thereby contributing to the field the results of the evaluation as well as the methodology used. We identify studies which aggregate and categorize quantitative XAI evaluation methods. We then choose several subcategories from Nauta et al.'s paper [100], which is one of the studies we identified, and analyze the included evaluation methods, identifying the ones applicable for text data. We also evaluate SHAP explaining fake news detection of two datasets using the identified methods and compare the results to established results. Finally, we communicate the results in this thesis.

1.3 Thesis Structure

The structure of this thesis is as follows:

In Chapter 2, we provide a background on the topic of eXplainable Artificial Intelligence. It starts with an introductory overview of XAI and proceeds to discuss the resurgence of the field. Furthermore, the intended audience, goals, and challenges addressed by XAI are examined. A definition of the term 'explainability' is presented, culminating in the introduction of our taxonomy for model-agnostic XAI. This taxonomy is applied to a table showcasing model-agnostic methods, providing a succinct overview of popular XAI methods.

In Chapter 3, our focus shifts to the introduction of BERT, a black box model, used for fake news detection. A concise theoretical foundation is laid before delving into its implementation within this thesis. This includes a detailed account of the utilized datasets, the pretrained models employed, and corresponding results obtained through their application to the datasets.

The topic of the upcoming Chapter 4 is SHAP, our chosen XAI method. Firstly, a theoretical overview of SHAP is presented to establish a solid foundation. Following that, the implementation details are outlined, showcasing the explanations obtained using SHAP for our research.

In Chapter 5 the reader is introduced to current research on XAI evaluation. An extensive analysis of evaluation methods for assessing of correctness, output completeness, and continuity is provided. This includes the availability of a GitHub repository, whether SHAP or Local Interpretable Model-Agnostic Explanations (LIME) have been evaluated using the presented method, the text dataset used as well as the task for the text data is presented. The implementation details and results of each evaluation tests are then presented, the code for the evaluations can be found at `https://github.com/elengue/eval_SHAP_BERT_text`.

The concluding Chapter 6, provides a summary of the main findings discussed throughout the thesis. Furthermore, it addresses the limitations encountered during the research process. Finally, potential avenues for future work and areas of improvement are explored.

2 Background

In this chapter, we introduce the current state of the art in XAI. It starts with a general introduction to XAI, followed by an exploration of model-agnostic XAI models.

2.1 XAI: the Renaissance of a Field

The denial of parole [145], the preference for male applicants in Amazon's recruitment tool [34] or Microsoft's racist chatbot [9] are only a handful of examples of AI gone wrong. These examples underscore the necessity of oversight through explainable and transparent models, highlighting the relevance of the field of XAI. While not a novel area of research, the field of XAI has experienced a recent renaissance. The initial interest in this subject dates back to the 1980s, primarily focused on providing explanations for decisions made by knowledge-based and expert systems [66, 67, 139]. Vilone and Longo [139] have observed a surge of interest in recent years (2016 to 2020) identifying 250 scientific articles, which is quadruple the amount compared to the previous 5-year period. This surge can be attributed to advancements in AI research, especially Deep Neural Network (DNN) which demonstrate high precision but are characterized by highly complexity and black box nature. Islam et al. [66] identify three events responsible for triggering this new wave of XAI research. Firstly, the U.S. Defense Advanced Research Projects Agency (DARPA)'s funding of the 'Explainable AI (XAI) Program' [59]. Secondly, to encourage high and strong explainability, China announced 'The Development Plan for New Generation of Artificial Intelligence' [146]. Finally, the European Union in its General Data Protection Regulation (GDPR) codified a 'right to explanation'[1]. Although, what exactly this right entails is still up to some debate [141], it certainly helped jumpstart the renewed interest in XAI research. In addition, to the GDPR the EU's High Level Expert Group on AI published 'Ethics Guidelines for Trustworthy Artificial Intelligence'[2] which includes as two of its seven key principles transparency and accountability. Furthermore, alongside these developments, the EU is actively working on the Artificial Intelligence Act, a comprehensive law regulating AI with a focus on ethical and safety consider-

[1] https://gdpr.eu/article-22-automated-individual-decision-making/

[2] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

ations, with the aim of reaching a final agreement by the end of the year[3],[4]. While, these events serve as the cornerstones of the new wave of XAI research, numerous other indications of interest have been observed for this field, such as initiatives from various countries (Netherlands [6], Portugal [65], and the United Kingdom[128]) as well as interest from various companies such as Google's AI design principles[5][24]. Of which one of the major pillars is interpretability with principles such as treating interpretability as a core part of the user experience, communicating explanations to models users, designing the model to be interpretable and understanding the trained model. Examples for other companies are IBM[6] who introduced AI Explainability 360 an open source toolkit helping people understand various machine learning models [7] and also established Trust and Transparency Principles[8] which include a requirement of new technologies such as AI systems to be transparent and explainable, H20 Driverless AI[9] which includes explanations as a part of its solution or companies with their whole business based on XAI like Galileox[10] or Minedxai[11].

2.2 Audience and Goals of XAI

The field of XAI is a multidisciplinary research area, as multiple disciplines contribute to its development. Naturally, Artificial Intelligence is a key component, along with data science and related disciplines such as mathematics, statistics, and computer science, as they provide the foundation to create AI models. However, XAI also benefits from the fields of human sciences or social studies, in particular human computer interaction and psychology, which contribute to the understanding of how users interact with explainable AI systems. Moreover, ethics and philosophy play an important role in voicing the need for explanations, thereby adding to the field [139]. Thus, improvements in on field contributing to XAI can catalyze progress in other fields as well as in the field of XAI [24].

Given the diverse disciplines involved in the field of XAI, differences in

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[4] https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[5] https://ai.google/responsibilities/responsible-ai-practices/

[6] https://research.ibm.com/topics/explainable-ai

[7] https://aix360.mybluemix.net/

[8] https://www.ibm.com/policy/trust-transparency-new/

[9] https://h2o.ai/platform/ai-cloud/make/h2o-driverless-ai/#Ingest-item-e50d9c2a21!

[10] https://galileox.ai/

[11] https://www.minedxai.com/

motives and starting points arise, prompting an exploration of 'Who uses XAI?' alongside the goals and motivations for its use ('Why?').

2.2.1 Audience

In order to define the concept of explainability, it is essential to take into consideration the intended audience for the explanation. Miller et al. [93] caution that in the field of XAI the developers are, for the most part, designing according to their own requirements and not the needs of the intended users of the explainability methods. In a recent study conducted by Cambria et al. [22], they observed that surveys, which consider different groups of users, cater to their intended audience. While, papers that do not explicitly consider the aspect of audience usually target technical users with a technical understanding. In their roadmap, Cambria et al. [22] incorporate the three most commonly mentioned types of users: ‘end user’, ‘developer’, and ‘decision maker’. While they do not formally define these roles, Mohseni et al. [95] utilize a similar division of target groups, calling them ‘AI novices’, ‘data experts’, and ‘AI experts’. The ‘AI novice’ is the end user using the model with little to no technical background knowledge. The ‘AI expert’ is the person who implemented the AI model or the explainability method. The ‘data expert’ is the domain expert who actively uses the AI model as a basis to make choices, do research, or analyze the data. As an example of a more fine-grained categorization of the intended audience of XAI methods by Arrieta et al. [12] include:

- Domain experts/users of the model such as insurance agents or medical doctors who need to be able to trust the model itself and also might want to derive scientific insights from the model.
- Users affected by the model decisions who want to be able to understand the decisions made about them as well as verify the fairness of said decision.
- Data scientists/developers/product owners who are interested in new functionalities and research of the model, as well as ways to improve and confirm the model works efficiently.
- Managers and executive board members who need to make sure all regulatory requirements are met as well as make consider other uses for the model.
- Regulatory entities/agencies who need to perform audits and ensure the models comply with legal obligations.

Considering this list of potential users/audiences, it becomes clear that each group has different reasons for wanting or needing explanations, as well as different thresholds of what is considered an explanation.

2.2.2 Goals of XAI

In addition, to considering the intended audience for explanations, it is also important to understand specific motivations. Although the ultimate goal of understanding black box models is quite clear, the specific reasons behind this need can differ, leading to variation in the nature of the explanation provided. Arrieta et al. [12] highlight that while, there is no consensus on the goals of XAI in the literature they reviewed, several recurring themes have emerged from the collective body of work. In order to provide additional insights, they have compiled a table presenting an overview of the identified goals in relation to the target audience and the respective papers where these goals were found. In total, they present a list of nine goals which are as follows: trustworthiness, causality, transferability, informativeness, confidence,fairness, accessibility, interactivity, and privacy awareness. Doshi and Kim [41] and Carvalho et al. [24] list a similar group of goals such as fairness, privacy, reliability/robustness, causality, and trust.

Using a different way of categorizing the reasons for XAI, Adadi et al. [3], along with Vilone and Longo [139] and Ding et al. [40] suggest four aspects motivate the use of XAI. Which are to explain to:

Justify in light of issues arising from biased or discriminatory decisions made by AI models, there is a growing demand to understand the reasoning behind the model's decisions. Specifically, the focus is on explaining how a model arrives at a particular decision, rather than providing descriptions of the model's structure or inner workings. This is particularly essential when the model produces unexpected results, as the explanation could function as a justification, supports audits, and acts as proof of the model's fairness and ethical compliance to instill trust. Additionally, it helps to fulfill legal obligations such as the EU GDPR's 'right to an explanation'.

Control explainability offers increased transparency, enabling better control over AI models. It provides insights into the model's functioning, helping to identify mistakes and facilitating the debugging process to prevent errors.

Improve explanations of the model's decision-making process enable iterative improvements, leading to enhanced performance and increased accuracy over time.

Discover the power to explain also opens the door to uncovering new knowledge. By providing explanations, XAI allows to uncover previously unknown causes, theories, or facts.

2.3 Explainability: Attempts at a Definition

There is currently no consensus on the exact definition of the term explainability as well as related terms such as interpretability, transparency, or understandability [40]. However, surveys on XAI consistently highlight the importance of establishing a common definition and suggest possible ways and versions to achieve this [3, 12, 24]. Arrieta et al. [12] shed light on the lack of common definitions within the community, attributing it to the incorrect usage of the terms 'explainability' and 'interpretability' interchangeably. They propose a clear distinction between these terms: 'interpretability' refers to the inherent quality of a model that makes it understandable to users (passive), while 'explainability' encompasses deliberate actions taken to make the model clearer (active). In their work, Palacio et al. [104], conducted a comprehensive analysis of the definition of 'explanation' and 'interpretation' in XAI literature and created a table presenting the various definitions found. Additionally, they examined definitions provided by several common dictionaries. Based on their findings, the authors formulated their own definition for 'explanation' as "*the process of describing one or more facts, such that it facilitates the understanding of aspects related to said facts (by a human consumer)*". Habiba et al. [60] also provide a tabular overview of the different definitions used by Lipton [84], Doshi-Velez and Kim [41], Gilpin et al. [53], Miller [92], and Montavon et al. [96].

Nauta et al. [100], try to take a comprehensive approach by not making a distinction between explainability and interpretability to ensure that no aspects are excluded. They also treat explainable artificial intelligence and interpretable machine learning as equivalent terms. Their definition of 'explanation' is: "*a presentation of (aspects of) the reasoning, functioning and/or behavior of a machine learning model in human-understandable terms*". Their definition is based on several other definitions such as van Lendt [136] who coined the term XAI as well as adopting 'human-understandable terms' from Doshi-Velez and Kim [41]. Inspired by the definition of Gilpin [53] 'reasoning, functioning and/or behavior' are used to emphasize that different methods of explanation are possible. Reasoning includes the process of reaching an explanation, functioning refers to the data structure and inner workings of the AI model, and behavior highlights the high-level global behavior of the model. This thesis follows the same definition, since we use Nauta et al. [100] as the basis of some of our work in the following chapters.

2.4 Categorizing XAI Methods

The resurgence of interest in XAI has resulted in a boom of proposed explainability methods. Consequently, there is a need to categorize and organize these different methods in order to gain a comprehensive understanding of the available options and select the most appropriate ones for each situation. Such categorization provides a valuable tool for researchers and practitioners to navigate the vast landscape of XAI methods effectively. Categorizing XAI methods can be approached in various ways, with one fundamental consideration being the stage at which interpretability is achieved. This distinction indicates whether interpretability is inherent to the AI method itself, termed 'intrinsic' or 'ante-hoc' interpretability, or whether an explanation is required after the model is trained, known as 'post-hoc' interpretability [42]. Intrinsic interpretability is achieved by building self-explanatory models, such as decision trees, rule-based models, and linear models, which are inherently interpretable due to their structures. Post-hoc interpretability, on the other hand, involves developing a second model to provide explanations for an existing model [116]. The key distinction between these two types is the trade-off between model accuracy and explanation precision. This trade-off refers to the phenomenon that intrinsically interpretable methods often exhibit lower accuracy compared to black box models, which require post-hoc explainability methods[3].

Another important consideration in categorizing XAI methods is their applicability, specifically whether they are 'model-agnostic' or 'model specific', which is only relevant for post-hoc explanations. When a method is model-agnostic, it can be applied to any AI method, irrespective of its underlying architecture. In contrast, a model specific method is tailored to only work with a specific type of AI method, such as support vector machines or DNN [130]. For the purposes of this thesis, we exclusively focus on post-hoc model-agnostic XAI methods. This choice is motivated by their broader applicability and their enduring relevance, which stems from their independence from specific black box models.

Other than the stage and applicability of XAI methods, there are numerous other factors, that can be employed to categorize the different methods. Sokol and Flach [129] identified 32 different criteria for categorizing and contrasting various XAI methods. However, while it is desirable to have a clear understanding of the distinctions among model-agnostic XAI methods, using such an extensive list of criteria may be overwhelming and counterproductive, potentially leading to more confusion. This raises the question of what are the most helpful criteria are for comparing a multiple XAI methods. Speith [130] conducted a review of eleven papers introducing taxonomies for categorizing XAI methods. He identified four different approaches to taxonomies: functioning-based, result-based, conceptual, and mixed. Each of these approaches serves different purposes and provides valuable insights for different types of users. As a result of his study, Speith introduced his own taxonomy, which serves as a basis for the taxonomy presented in the next chapter.

2.4.1 Overview of Model-Agnostic XAI Methods

Overall we found 15 different surveys [3, 18, 24, 33, 38, 40, 43, 67, 66, 69, 100, 113, 119, 120, 138] that categorize XAI methods, which in total led to over 150 unique model-agnostic XAI methods. LIME [116] was mentioned in 14 of the 15 papers. The goal of this thesis is to evaluate XAI on fake news data. Thus, we do not provide a comprehensive categorization of all the methods we encountered. Instead, we focus on methods mentioned multiple times, as this suggests their significance. We excluded methods mentioned which are more general concepts than concrete methods such as feature importance, feature interaction, global surrogates, model distillation, saliency maps, sensitivity analysis, and surrogate models, additionally we excluded DALEX [17] which is an R and Python package of multiple explainability methods. This leaves us with a total of 26 methods, which we categorized in Table 1, according to the forthcoming taxonomy.

The taxonomy we employ in this thesis considers the following four criteria: scope, input data, result, and output format. We introduce the criteria in the subsequent sections. It is important to note that the categories within this taxonomy are not mutually exclusive, allowing for flexibility in their application.

Scope. We distinguish two forms of explainability: global explainability and local explainability [42, 116].

- ‘Local’ explainability looks at an individual prediction of a model to see why it makes the decision it does [152].

- ‘Global’ interpretability looks at how the model operates globally by inspecting the mechanisms and parameters of a complex model [3].

There are also models that offer both interpretability at the local and the global level [40, 66].

Input Data. Input data describes which data types the XAI method accepts. We use the following categories as used by other surveys [18, 58], ‘tabular’ data which can either be numerical or categorical, ‘image’, ‘text’, or ‘any’ meaning the method can be used for data of any type might it be tabular, text, or image.

Others also include other categories such as time series [138], auditory [69], sensory [69] or, graph data [40, 100] as input data. However, for the sake of simplicity and due to their absence in the methods we categorized, we chose to omit them from our taxonomy. Schwalbe et al. [120] take a different approach all together and split the input data into symbolic data, which includes tabular, natural language, and graph data and non-symbolic data which includes images, point clouds, and audio.

Result. This category focuses on the output or result of the XAI method. There are three categories for result which were used by McDermid [91].

- ‘Surrogate Models’ are interpretable models, which were created based on the black box models.

- ‘Feature Relevance’ give information on which features influences the result in which way.

- ‘Examples’ are given to illustrate the black box model’s reasoning.

Output Format. The category Output Format describes in what format the explanation of the XAI method is delivered. It included in the taxonomy due to its significance in method selection, as different output formats may be more suitable based on factors such as speed, ease of use, and user type. Within this category, the following formats, as used by Vilone and Longo [138], are differentiated:

- ‘Numerical’ explanations presented in numerical form.
- ‘Rule’ explanations expressed in one of two forms, either as decisions trees or as IF-THEN rules that include AND/OR operators.
- ‘Textual’ explanation provided in written text form.
- ‘Visual’ explanations conveyed through visual representations, such as plots.
- ‘Mixed’ explanations encompass combinations of the aforementioned output formats.

Vilone and Longo [138] argue that textual explanations may not be ideal for XAI novices.

Table 1: Table of model agnostic methods mentioned more than once by our chosen surveys

Method	Scope	Input Data	Result	Output Format	Found in
ALE (Accumulated Local Effects) Plot [11]	global	tabular	feature summary	visual	[24, 43, 67]
Anchors [117]	local, global	any (text, image, tabular)	feature summary	rules	[18, 24, 40, 43, 66, 69, 100, 138]
BreakDown [131]	local	tabular	feature summary	mixed (text, visual)	[24, 67]
CaCE (Causal Concept Effect) [55]	local, global	image	feature summary	numerical	[33, 40]
CAM (Class Activation Maps) [132, 157]	global, local	image	feature Summary	visual	[33, 66]
CEM (Contrastive Explanations Approach) [37]	local	tabular, image	examples	visual	[120]
CIU (Contextual Importance and Utility) [10]	local	tabular	feature summary	mixed (visual, text)	[18, 40]
Counterfactuals CERTIFAI [141, 123, 72, 48]	local	tabular, image	examples	text	[3, 24, 38, 43, 66, 67]
DeConvolutional Nets [154, 103]	local	image	feature Summary	visual	[33, 66]
DGN-AM (Activation Maximization Deep Generator Network) [46]	global	image	feature Summary	visual	[3, 38]
DiCE (Diverse Counterfactual Explanations) [98]	local	any (text, image, tabular)	examples	rules	[18, 40]
FACE (Feasible and Actionable Counterfactual Explanations) [110]	local	any (text, image, tabular)	examples	visual	[18, 66]
GLocalX (Global through Local Explainability) [122]	local, global	tabular	surrogate model	rules	[18, 40]
Grad-CAM++ [26]	local	image	feature summary	visual	[33, 69]

ICE (Individual Conditional Expectation) [54]	global, local	tabular	surrogate model	visual	[3, 24, 43, 67, 138]
Influence Function [75]		image	examples	text	[24, 38, 67, 138]
LRP (Layer-wise Relevance Propagation) [14]	local	any (text, image, tabular)	feature summary	visual	[3, 18, 40, 138]
LIME (Local Interpretable Model-agnostic Explanations) [116, 115]	local	any (text, image, tabular)	surrogate model	mixed (visual, text, numerical)	[3, 18, 24, 33, 38, 40, 43, 58, 67, 69, 100, 113, 120, 138]
LORE (LOcal Rule-based Explainer) [57]	local	tabular	surrogate model	rules	[18, 40]
MAPLE (Model Agnostic SuPervised Local Explanations) [108]	local	tabular	surrogate model	numerical	[18, 38, 40, 100]
MMD-critic (Maximum Mean Discrepancy) also called Prototypes and Criticisms [74]	local, global	image	examples	mixed (visual, text)	[3, 24, 40, 67, 100, 138]
PDP (Partial Dependence Plot) [15, 45, 50, 56]	global	tabular	surrogate model	visual	[3, 24, 38, 43, 67, 120]
RISE (Randomized Input Sampling for Explanation) [107]	local	image	feature Summary	visual	[33, 120]
SHAP (SHapley Additive exPlanations) [88, 87]	local, global	any (text, image, tabular)	feature summary	mixed	[3, 18, 24, 33, 38, 43, 67, 69, 100, 113, 138]
SKOPERULE [51]	local, global	tabular	feature summary	visual	[18, 40]
SpRAy (Spectral Relevance Analysis) [79]	local, global	image	surrogate model manner	visual	[33, 138]

3 Black box model: BERT

In a benchmark study, Khan et al. [73] considered approaches from traditional machine learning, deep learning models, and advanced pretrained language models, which in total lead to 19 different machine learning approaches tested on three different fake news datasets. They found advanced pretrained language models, especially BERT based models [35], performed better on all three datasets. Therefore, we use a BERT based model as the black box model.

This chapter details the implementation of the black box model for fake news classification. We provide a short theoretical background on the chosen models and the dataset. We chose BERT and fake news datasets as the data and black box model to be explained due to their high relevance. BERT has proven to be very effective for texts tasks and fake news is a highly relevant topic.

3.1 Theoretical Background

BERT [35], which stands for Bidirectional Encoder Representations for Transformers, is an advanced pre-trained language model [73]. BERT whose architecture is based on the transformer architecture introduced by Vaswani et al. [137] has made waves in the Natural Language Processing (NLP) community since its introduction by Google in 2018 [134].

A universal language representation can be learned through pre-training on a large dataset. Pre-training can help the use of smaller datasets as it helps to prevent overfitting as regularization factor [112].

Pretraining for BERT was done on two different tasks. *Masked Language Model (MLM)*, which is sometimes also referred to as the clove task. For MLM 15% of the tokens are at random replaced with a special [MASK] token. In order to prevent a mismatch between pretraining and fine-tuning where the [MASK] token will not be present, the tokens are only replaced 80% of the time with the [MASK] token in 10% of the cases the token will be replaced by another random token and in the remaining 10% the token will not be replaced. The goal is to predict the masked token [35]. *Next Sentence Prediction (NSP)* is the other task BERT is trained on and used to make the model understand the relationship between sentences. For this task, sentence pairs (sentence A and B) are extracted from a corpus. In 50% of the cases, the training example's sentence B is the actual next sentence and in the rest it is a random sentence [35]. Bidirectionality from both sides is necessary.

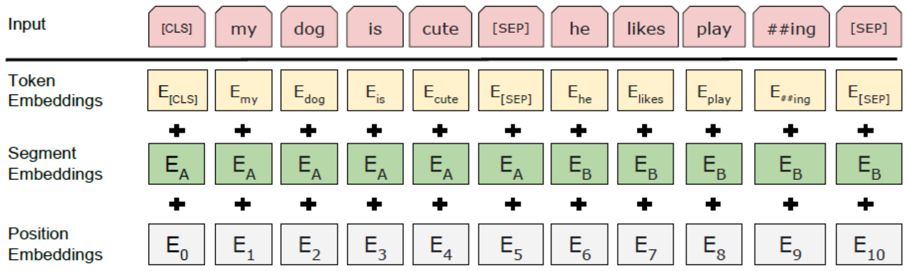


Figure 2: BERT input representation [35]

Figure 2 depicts the input representation, the first layer is the token embedding, including special tokens such as the [CLS] token for the beginning of each sequence and the [SEP] token which indicates the separation between two sentences or segments. The second layer is the segment embedding, which additionally to the [SEP] token denotes to which segment a token belongs. The third layer is the positional embedding, which is used to give information about the position of the tokens, since no recurrence or convolution is used that would provide such information [35].

Attention, as the title of the paper 'Attention is all you Need' already suggests, is an important part of the model. In the paper, Vaswani et al. [137] point out three ways in which the model uses attention. Firstly, in the layers of the encoder-decoder where input is the output of the previous decoder layer as well as the output from the encoder layer. It enables the consideration of all positions of the input sequence. Secondly and thirdly in the self-attention layer of the encoder as well as the decoder, allowing for every position of the encoder to consider every position of the previous layer and for the decoder to consider, including up to the position of the decoder.

3.2 BERT Implementation

We used Flores and Hao's [49] BERT fine-tuned classifiers retrieved from GitHub[12] as the black box model to be explained. In their paper, they create three adversial attacks as benchmarks for fake news detection and test them on two BERT classifiers.

[12] https://github.com/ljyflores/fake-news-adversarial-benchmark

Id	**Statement**	**Label**
2359.json	I kicked crooked cops and government officials off the public pension rolls.	1
7313.json	State revenue projections have missed the mark month after month.	0
7234.json	Obamacare includes a $63 charge every American will begin paying (in 2013) as a way to cover some of the increased costs associated with providing health insurance to those with pre-existing conditions.	1
11243.json	Buying the naming rights to the new Falcons stadium is the largest marketing deal in Mercedes-Benz history.	0
9119.json	If the legislature refuses to expand Medicaid, 27 states are going to get our money Virginia taxpayer money into their states to provide health care for their citizens, paid for by us.	1

Table 2: Five random rows of the LIAR dataset

3.2.1 Datasets

The Datasets used are the LIAR dataset[13] and the Kaggle Fake News dataset[14]. The LIAR dataset [143] consists of statements made on social media (such as Facebook posts or Tweets) as well as statements made in political debates, TV ads, interviews, or news releases etc. The labels for this dataset were created by the editors of POLITIFACT.COM and are fine-grained with 6 categories: 'pants-fire', false', 'barely-true', 'half-true', 'mostly-true', and 'true'. In the context of Flores and Hao's work, they consider both multi-label and binary classification. For this thesis, only binary labels are considered. Therefore, the labels were collapsed to only reflect true or false with the labels 'pants-fire, 'false', and 'barely-true' being converted to 'false' or the value 0 and the labels 'half-true', 'mostly-true', and 'true' changing to true or the value 1. Table 2 shows an excerpt of 5 random rows of the LIAR dataset, consisting of the 'id', the 'statement', and the 'label'.

The other fake news dataset used is the Kaggle Fake News[15] which was created for the task of developing a machine learning program, which can identify if an article is fake news. The Dataset is part of a Kaggle competition run by the UTK Machine Learning Club. The version of the data used by Flores and Hao and subsequently used by us contains three columns an 'id' column which is an ascending number to uniquely identify each input row, a 'statement' which is the title column of the original dataset and contains the title of a news story to be classified and a 'label' column which has binary

[13] https://github.com/tfs4/liar_dataset

[14] https://www.kaggle.com/competitions/fake-news/overview

[15] https://www.kaggle.com/competitions/fake-news/overview

Label	LIAR	Fake News	IMDB
0	false	reliable (true)	negative
1	true	unreliable (false)	positive

Table 3: Statement labels and their meaning per dataset

Id	Statement	Label
3260	20 wines for under $20: the fall edition - the new york times	0
15337	have the dallas police improved? depends on whom you ask - the new york times	0
11578	cannabis aficionados develop thc-a crystalline: the strongest hash in the world at 99.99% thc	1
18850	how we can win the propaganda war	1
11722	russian grannies make deal with lipton	1

Table 4: Five random rows of the Fake News dataset

values which according to the dataset description mark 0 as reliable and 1 as unreliable. Which is the opposite label as compared to the LIAR dataset, we decided not to change this difference but caution the reader to it and remind the reader whenever relevant. As an overview, Table 3 shows the label values and the corresponding meaning for each dataset. Table 4 shows the first five rows of the dataset to get an impression of its content.

The datasets were obtained from the Google Drive link[16] provided by Flores and Hao's on their GitHub page[17]. For the rest of this thesis, these are the LIAR or Fake News dataset we refer to. Due to computational limitations 1,000 randomly selected rows of the 'fake_news.csv'[18], the 'liar_valid.csv', and the 'liar_valid.csv' located in the raw data folder were used for all subsequent experiments. The datasets were loaded using the pandas `read_csv()`[19] function and sampled using the `.sample()`[20] method with 1000 rows and `random_state=42` specified to ensure reproducibility.

In addition, to the LIAR and Fake News dataset, we included the International Movie Database (IMDB) dataset in our evaluations. The IMDB dataset is designed for sentiment classification tasks and widely used to eval-

[16] https://drive.google.com/drive/folders/1OzdrFakmNSOeOmQufYwQvTiESwP8pNyz

[17] https://github.com/ljyflores/fake-news-adversarial-benchmark

[18] The naming for the csv files for the Fake News dataset may be confusing as only fake_news.csv and fake_news_train.csv are present. However, we verified that the data from fake_news_test.csv is not contained in fake_news.csv and the data in the encoded fake_news_test.pt corresponds to the contents of fake_news.csv

[19] https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html

[20] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html

Statement	**Label**
Stop me if you hard this one before, some cheerleaders, their coach and a couple guys are trapped within a cabin in the woods ...	0
i enjoyed this film immensely, due to pungent scenes (humorous as well as ironic, some even ”tragical”), believable performan ...	1
While The Twilight Zone was a wonderful show, it was also very uneven–with some great episodes, some lousy ones and many in ...	0
algernon4's comment that Ms Paget's ”ultra lewd dance in (this film) is the most erotic in the history of films” is certainly one ...	1
Family guy. When the show first aired, it was fresh, original, and actually quite funny. Now, I have stopped watching it. It has ...	0

Table 5: Five random rows of the IMDB dataset

uate XAI methods for text data, as demonstrated in Chapter 5.1 Table 9. Using this dataset provides us with the opportunity to compare the results of our experiments with those of others on the same dataset.

The dataset consists of movie reviews from the movie review website, International Movie Database (IMDB). We retrieved the dataset function from Huggingface's datasets package[21]. Specifically, we loaded the test set of the IMDB dataset using the 'imdb' and 'test' parameters in the `load_dataset()`[22] function. Similar to the other two datasets, we shuffled the data, setting the seed to 42 to ensure reproducibility, using the `shuffle()`[23] method. The shuffled dataset was then formatted to a pandas dataframe using the `set_format`[24] method and the first 1000 rows were selected and stored in a csv file. From now on, we refer to this data as the IMDB dataset. The dataset consists of a column originally called 'text', which we renamed 'statement' to ensure consistency over all three datasets. The 'statement' column contains the movie reviews ant the 'label' indicates the sentiment of the reviews, with 0 representing negative or 1 representing positive reviews. Table 5 shows five random rows of the dataset. Only the first few lines displayed, since the reviews can be quite lengthy.

[21] https://huggingface.co/docs/datasets/index

[22] https://huggingface.co/docs/datasets/v2.9.0/en/package_reference/loading_methods#datasets.load_dataset

[23] https://huggingface.co/docs/datasets/v2.9.0/en/package_reference/main_classes#datasets.Dataset.shuffle

[24] https://huggingface.co/docs/datasets/v2.9.0/en/package_reference/main_classes#datasets.Dataset.set_format

3.2.2 Model and Tokenization

We use he already fine-tuned models, by Flores and Hao, which were retrieved from the Google Drive Link[25] provided on the Github page[26]. The models, the liar_model2 (trained on the LIAR dataset with binary labels) and the fn_model (trained on the Fake News dataset), were instantiated using the `BertForSequenceClassification.from_pretrained()`[27],[28] method. The 'bert-base-uncased'[29] tokenizer[30], which was also used during training, and is deployed in tandem with the models. Regarding the IMDB dataset, we use the fine-tuned model 'textattack/bert-base-uncased-imdb'[31] from the text attack package. The model was loaded in a similar manner to the other models and tokenizers, with the only difference being the usage of the `AutoModelForSequenceClassification`[32] and `AutoTokenizer`[33] classes and 'textattack/bert-base-uncased-imdb' instead of the usage of the model path for both the model and the tokenizer. Figure 3 gives a visual representation of the steps taken to create the SHAP values for each dataset and highlights the differences based on color.

Table 6 gives an overview of the tokens for the 1,000 row samples of the datasets, including information on the number of unique and total tokens, as well as the average amount of tokens per statement. These figures confirm that the statements for the IMDB Dataset are significantly longer than the other two datasets. When comparing the two datasets for fake news classification, we observe that the LIAR dataset to have more total tokens, whereas the Fake News dataset contains more unique tokens. Table 7 shows the first example of the Fake News dataset as tokens and words split up to token level. The [CLS] and [SEP] tokens representing the beginning and the end of a sequence respectively, and are special tokens within the tokenizer and the model.

[25] https://drive.google.com/drive/folders/1XFoYNmYP-DD3Bj7zg9AXzDT7VmGtf0aG

[26] https://github.com/ljyflores/fake-news-adversarial-benchmark

[27] https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification

[28] https://huggingface.co/docs/transformers/main_classes/model#transformers.PreTrainedModel.from_pretrained

[29] https://huggingface.co/bert-base-uncased

[30] https://huggingface.co/docs/transformers/v4.26.1/en/model_doc/bert#transformers.BertTokenizer

[31] https://huggingface.co/textattack/bert-base-uncased-imdb

[32] https://huggingface.co/docs/transformers/model_doc/auto#transformers.AutoModel

[33] https://huggingface.co/docs/transformers/v4.26.0/en/model_doc/auto#transformers.AutoTokenizer

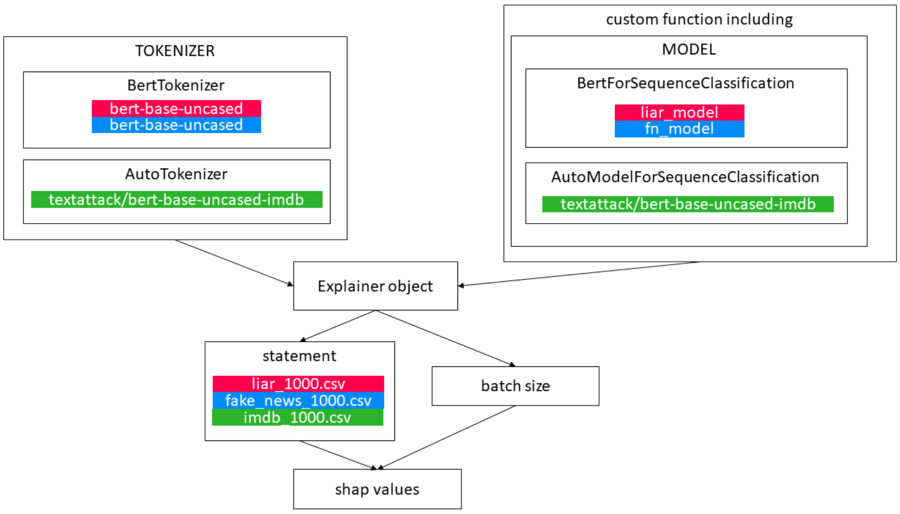


Figure 3: Outline of how SHAP values were computed

Dataset	Unique Tokens	Total Tokens	Avg Token per Statement
LIAR	3,985	24,148	24.15
Fake News	4,517	18,669	18.67
IMDB	22,407	288,544	288.55

Table 6: Statistics about the amount of unique tokens, total tokens and average tokens per statement for each dataset

3.2.3 Results

Table 8 shows the results for precision, recall, f1-score, support (indicating the number of instances in the sample have the label 0 and 1, for label descriptions, refer to Table 3), and accuracy across the 1,000 samples of the LIAR, Fake News and IMDB datasets. These results were computed using `classification_report`[34] from the sklearn package. The results clearly show the differences in the model's ability to classify the inputs between the datasets. For the LIAR dataset, the results are only slightly better than chance for a binary classification task. However, these results are on par with the best results mentioned by Flores and Hao [49]. They report a 98.9 accuracy score for the Fake News dataset by Kaliyar et al. [71] and a 27.3 accuracy score by Ding et al. [39] for the 6-label LIAR dataset, which

[34] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

Tokens	[101, 2322, 14746, 2005, 2104, 1002, 2322, 1024, 1996, 2991, 3179, 1011, 1996, 2047, 2259, 2335, 102]
Words	[’[CLS]’, ’ 20’, ’ wines’, ’ for’, ’ under’, ’ $’, ’ 20’, ’ :’, ’ the’, ’ fall’, ’ edition’, ’ -’, ’ the’, ’ new’, ’ york’, ’ times’, ’[SEP]’]

Table 7: Example of tokenization

	LIAR		Fake News		IMDB	
	0	1	0	1	0	1
Precision	0.58	0.54	0.97	**0.99**	0.89	0.88
Recall	0.53	0.59	**0.99**	0.98	0.88	0.89
F1-score	0.55	0.57	**0.98**	**0.98**	0.88	0.88
Support	515	485	495	505	512	488
Accuracy	0.56		**0.98**		0.88	

Table 8: Precision, recall, f1-score, support, and accuracy for the datasets

Flores and Hao’s model outperforms with an accuracy score of 29.4. In contrast, Flores and Hao only report their own accuracy score of 57.4 for the binary classification of the LIAR dataset, which is the version we are using exclusively used in this thesis. We, on the other hand, achieve an accuracy score of 56 % for the same binary version of the LIAR dataset, which is close to the result reported by Flores and Hao.

4 XAI Model: SHAP

This chapter details the implementation of the XAI model. We provide a short theoretical background on SHAP, before presenting our implementation of SHAP on BERT and the datasets presented in the previous chapter.

4.1 Theoretical Background

SHapley Additive exPlanations, abbreviated as SHAP, is a local and model-agnostic explanation method, introduced by Lundberg and Lee [88]. Its significance in the field of eXplainable Artificial Intelligence is in part attributed to its robust theoretical foundation [97]. It is based on Shapely values, originating from cooperative game theory [18] and measures how to fairly distribute a reward among a group of players based on their contribution to the outcome [97].

For model explanation SHAP defines the problem as the game in which the features act as the players and the model's prediction represents the outcome. Thus, Shapley values are employed to compute feature importance, as they fairly distribute the impact of each feature to the model's prediction [16]. However, the exact computation of Shapley values becomes infeasible quickly due to the need to calculate all possible feature combinations. Therefore, approximations are utilized to make the computation manageable [97].

4.1.1 Properties and Versions of SHAP

SHAP as one such approximation, uses additive feature attribution methods, to create an explanation [138]. It possesses three major properties: *local accuracy*, ensuring its output matches with the output of the original model with simplified input when explaining a particular instance; *missingness*, which prevents missing input features from impacting SHAP values; and *consistency*, ensuring that an increase in the model's marginal contribution leads to a corresponding increase in the SHAP value [88]. Lundberg and Lee [88] propose five different methods for computing SHAP values. The first is Kernel SHAP, a model-agnostic approach that combines LIME and Shapley values. The other four methods namely Linear SHAP, Low-Order SHAP, Max SHAP, and Deep SHAP are model-specific. In addition, the authors released an official library[35], which has been continually expanded. New additions to the library include Gradient SHAP, Tree SHAP (both model specific), and

[35] `https://github.com/slundberg/shap`

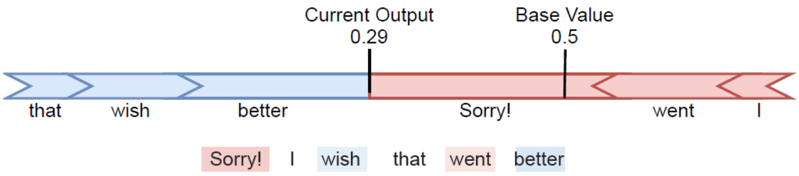


Figure 4: Local SHAP graph explaining a sentiment analysis example [97]

Partition SHAP[36], a faster version of Kernel SHAP that incorporates hierarchies and leverages Owen values from game theory for computation [97]. For detailed insights into the workings of the Partition Explainer, refer to [142].

4.1.2 SHAP's Visualisations

The SHAP package offers various visualization options for local as well as global explanations. Thereby providing the options to create graphs explaining individual instances as well as overall trends. Figure 4, created by Mosca et al. [97], is a schematic example of a graph for a local explanation, specifically applied to sentiment analysis. The base value represents the model's average prediction. Each token or word is treated as a feature that contributes to the difference between the base value and the current output, which is the result. For the given instance, 'Sorry!', 'went', and 'I' have negative contributions, pushing the SHAP value towards the current output. On the other hand, 'better', 'wish', and 'that' represent negative contributions, shifting the result closer to the base value. The graph shows how individual features affect the final prediction.

While SHAP is primarily designed as a local XAI method, it does offer the option to create global explanations. As an example of such a global visualization, we present Figure 5, which depicts global explanations. In order to ensure better understanding of the global graphs, we chose a graph presenting the results for a model utilizing tabular data, as they are in our opinion more intuitive. In this graph, extracted from the 'An introduction to explainable AI with Shapley values' article in the SHAP documentation[37] the average most influential factors for an adult to earn over $50,000 annually are depicted. Additionally, it assesses the average impact of each variable on the

[36] https://github.com/slundberg/shap/blob/b6e90c859fdfc6bc145242d9a8082d4ad844e995/shap/explainers/_partition.py

[37] https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html#nlp_model

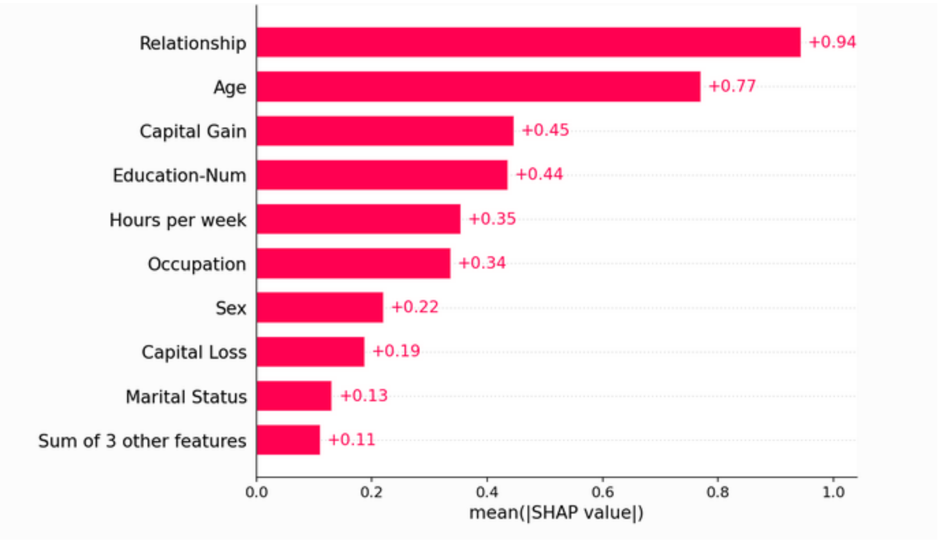


Figure 5: Example of a summary plot with tabular data to facilitate better understanding retrieved from the SHAP documentation[38]

SHAP values. It shows that the 'relationship' variable has the most impact, it shows on the bottom that the three least influential categories have been omitted. Summary plots for text data differ from those of tabular data due to the nature of the data type. Text data does not always contain the same variables per instance as tabular data does. In the example above, every entry in the dataset has a value for relationship, age etc. Whereas for text data the variables are the tokens, which make up words, every instance will therefore only have a fraction of the overall variables present, as not all the words present in the whole dataset will also be present in each instance. This results in a higher amount of variables and a relatively low average impact per variable. In Figure 6, we provide the summary plots created for this thesis, following the pattern for text data explained above.

4.2 SHAP Implementation

In this thesis, we implement SHAP using the SHAP python package[39]. We adapt an example from the package's documentation[40] to suit the specific requirements of the study. The computation of the SHAP values is executed on

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

[40] https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html#nlp_model

```
    .values = array([ 0.,  0.00490105,  0.00268453, 0.00688738,
    ↪  0.00365818, 0.00770843,  0.0048399 ,  0.00829434, 0.0130564 ,
    ↪  0.00249624, 0.00745153,  0.0097388 , 0.00669724, -0.00151104,
    ↪  -0.00126433, -0.00331485,  0.])

    .base_values = 10.309349060058594

    .data = array(['', ' 20', ' wines', ' for', ' under', ' $', ' 20', '
    ↪  :', ' the',' fall', ' edition', ' -', 'the', ' new', ' york', '
    ↪  times', ''], dtype='<U8')
```

Listing 1: Example of the content of shap_values

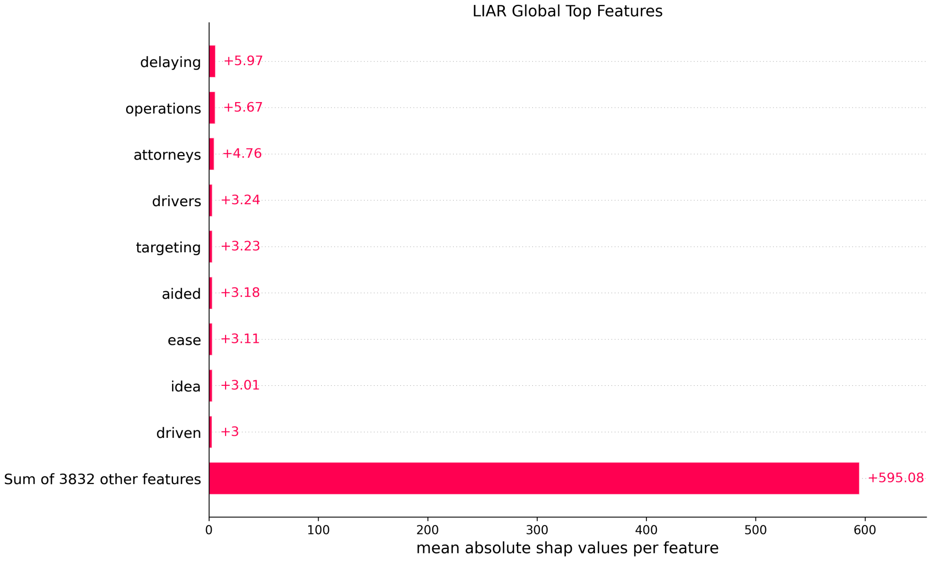
a NVIDIA A100-SXM4-40GB, utilizing Google Colab Pro. For each dataset containing 1,000 samples and using a batch size of 20, the computation times for the SHAP values are as follows: 54 minutes and 29 seconds for the LIAR dataset, 36 minutes and 18 seconds for the Fake News dataset, and1 hour 10 minutes and 36 seconds for the IMDB dataset. The Partition Explainer[41], a model-agnostic explainer that calculates Owen values as an approximation for SHAP values, was employed.

Listing 1 illustrates the output for a single instance of the SHAP values. The first list, denoted as '.values', contains the SHAP values, which correspond to the .data list. These values signify the influence each token (.data element) exerts on the overall classification. The '.base_value' represents the mean prediction of the model. The Figures 7a, 7b, and 7c show the local explanation for one randomly selected sentence from each dataset. As a brief reminder of the dataset introductions provided in Section 3.2.1 the meaning of 0 and 1 differs per dataset. In the LIAR datasets, 0 corresponds to fake news, and 1 corresponds to real news. Conversely, in the Fake News datasets, 0 represents reliable/real news and 1 represents unreliable/fake news. For the IMDB dataset, 0 indicates a negative review, while 1 denotes a positive review, see Table 3. These Figures can be interpreted similarly to the example graph shown in Figure 4. Specifically, Figure 7a features a base value of -7.5. Notably, the most influential tokens are 'official' and 'government', which are depicted with the most solid color and occupy the longest space on the chart. Overall, the blue/negative arrows have more value than the red/pos-

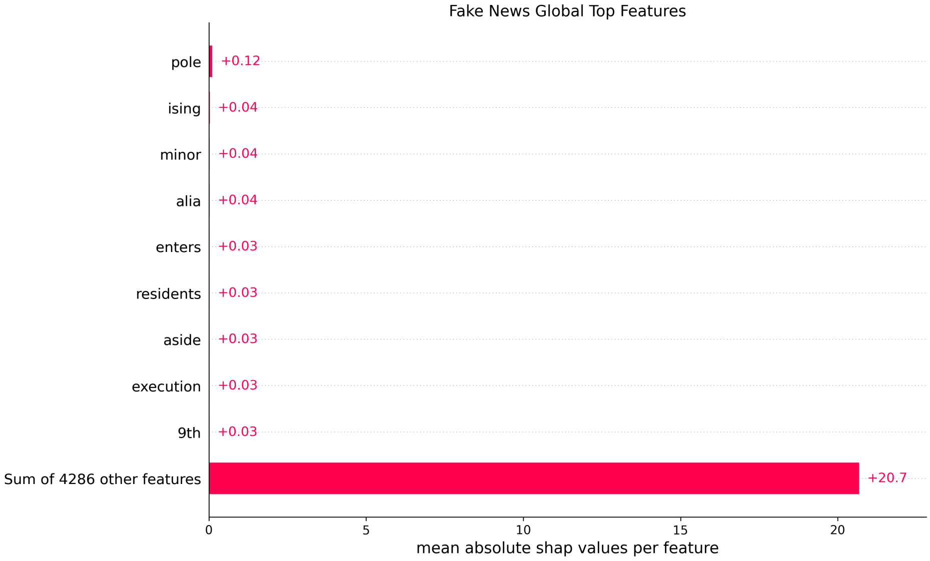
[41] https://github.com/slundberg/shap/blob/b6e90c859fdfc6bc145242d9a8082d4ad844e995/shap/explainers/_partition.py

itive arrows, and the ‘f(inputs)’ is lower than the base value. According to Table 2, the example in question holds a true label of 1, indicating a true statement. In Figure 7b, we observe the local explanation for a randomly selected example of the Fake News dataset, which, as shown in Table 4 bears a true label of 0, indicating a reliable statement. The base value is 10.309, and the ‘f(inputs)’ is 10.381 with the majority of tokens being red/positive. Finally, Figure 7c shows the local explanation of the IMDB example, which holds a true label of 0, signifying a negative review as shown in Table 5. The base value is -2.1, and the ‘f(inputs)’ is -8.3 with the big majority of tokens being blue/negative.

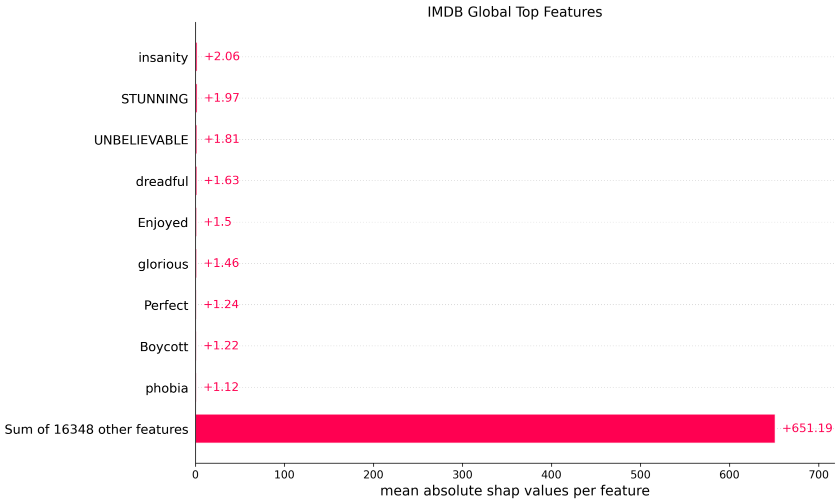
Figure 6 presents the global explanations obtained for our analysis. The graph illustrates the most influential tokens per dataset, where token importance is computed as the mean of the local SHAP values for each token. For the sentiment analysis of the IMDB dataset, Figure 6c showcases the three most influential tokens: ‘insanity’, ‘STUNNING’, and ‘UNBELIEVABLE’. These tokens are plausible choices as they convey strong emotions, aligning well with sentiment analysis objectives. Moving on to the fake news datasets, Figures 6a and 6b depict the top global features representing the most influential tokens for the LIAR and Fake News datasets, respectively. For the LIAR dataset, the top three tokens are ‘delaying’, ‘operation’, and ‘attorneys’ while for the Fake News dataset’s most influential tokens are ‘pole’, ‘ising’, and ‘minor’. Given the nature of sentiment analysis and fake news detection, it is more straightforward to perform a brief sanity check on the sentiment analysis results. However, assessing the validity of fake news detection results becomes more challenging, especially considering the low accuracy score of the Fake News dataset.



(a) Top global SHAP features - LIAR dataset

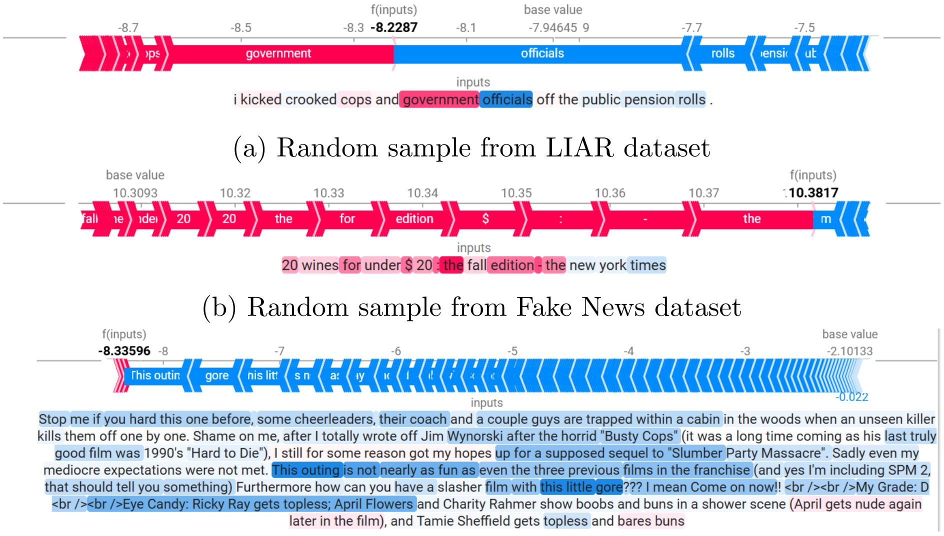


(b) Top global SHAP features - Fake News dataset



(c) Top global SHAP features - IMDB dataset

Figure 6: Global most impactful SHAP Values



(a) Random sample from LIAR dataset

(b) Random sample from Fake News dataset

(c) Random sample from IMDB dataset

Figure 7: SHAP text plot for one random example per dataset

5 Evaluation of XAI method

This chapter provides an overview of how XAI methods are currently quantitatively evaluated and assesses SHAP explanations of the BERT fake news classification model based on selected criteria.

5.1 Background: Evaluation of XAI Methods

The necessity to evaluate XAI methods is underscored by two primary reasons [90, 158]. Firstly, the need arises for establishing metrics that facilitate the comparison of various methods, enabling users to make informed decisions based on evaluation outcome. Especially due to the lack of ground-truth for post-hoc explainability, since the inner workings of the model are unclear [40, 62]. Secondly, evaluations are imperative to ascertain the fulfillment of defined objectives and thereby the accomplishment of explainability within a given application.

5.1.1 An Absence of Standardized Evaluation Approaches

The absence of a benchmark or standardized approach for evaluating eXplainable Artificial Intelligence (XAI) frameworks is a significant concern [21, 40, 47, 62, 66, 86]. This lack of a common evaluation standard especially hinders effective comparison between different frameworks, making it challenging to assess the performance of such methods, as noted in various surveys [12, 21, 41, 47, 85, 100]. Hedström et al. [62] even argue that currently XAI papers employ questionable and one-sided evaluation methodologies. These practices not only have the potential to limit the access to the current state of the art, but may also in the long run negatively impact the reputation of the XAI field. Their observation is supported by Nauta et al. [100], who conducted a systematic review of over 300 papers on the topic of XAI and found that one third of the surveyed papers relied solely on anecdotal evidence to evaluate the explainability of the examined model. Commonly in connection with the lack of evaluation standards, a lack of common definitions and terms is mentioned since using the same term for (slightly) different concepts only adds to the confusion [3, 12, 40, 85, 104]. Consequently, as we build part of this thesis on Nauta et al.'s work [100], we generally depend on their definitions, unless otherwise specified. For the definition of the term 'explainable' for this thesis, see Section 2.3. The study by Doshi-Velez and Kim [41] is a highly influential work frequently cited in discussions regarding the categorization of the evaluation of explainable artificial intelligence (XAI) [3, 86, 158]. In their work, they propose three distinct categories of evaluation methods. The first category is called 'functionally-grounded' evaluation, which involves assessing XAI methods using proxy tasks without human involvement. The second category, termed 'human-grounded', engages real human participants in attempting to solve simplified tasks using the XAI methods. Lastly, the third category, referred to as 'application-grounded', focuses on evaluating XAI methods by involving real humans in solving real-world tasks. Overall, a separation between tasks involving human participants and those without human involvement is common [40, 66, 86, 139, 158]. The scope of this thesis lies on evaluations without human participants and with quantitative measures. Recent studies try to organize existing evaluation methods into categories to enable a way to organize results and to classify which aspects of explainability are tested [85, 86, 100].

5.1.2 Classification Approaches of XAI Evaluation

Löfström et al. [85], conducted a semi-structured meta survey of fifteen surveys, wherein they categorize the evaluation criteria into three main aspects: model, explanation, and user. Under the model explanation criterion, the identified 'performance', 'fairness', and 'privacy'. The explanation aspect encompasses 'reliability', 'identity', 'separability', 'novelty', 'representativeness', and 'fidelity'. While the user aspect includes 'appropriate trust', and 'explanation satisfaction'.

Furthermore, Lopes et al. [86] propose a taxonomy that splits evaluation methods into two categories: human-centered and computer-centered. With the human-centered methods, they further identified four subcategories: 'explanation usefulness' and 'satisfaction', 'performance', 'understandability', and 'trust' with the latter two subdivided into three additional subcategories each. In contrast, the computer-centered category encompasses 'interpretability' with subcategories 'clarity', 'broadness', and 'simplicity' and 'fidelity' with the subcategories completeness and soundness.

Quantus[42] an evaluation framework for neural network explanations introduced by Hedström et al. [62] aims to facilitate the quantification of XAI and allows for the evaluation of image, time-series, and tabular data. While usage for text data is in development, at the time of writing this thesis, it is not yet available. The framework encompasses over 30 evaluation metrics, categorized into six different categories: 'faithfulness', 'robustness', 'localization', 'complexity', 'randomization', or 'axiomatic' metrics.

Additionally, Nauta et al. [100] performed a systematic literature re-

[42] https://github.com/understandable-machine-intelligence-lab/Quantus/

view of 312 papers that introduced a XAI method and performed a quantitative evaluation of the proposed method. In the context of their survey, Nauta et al. create a definition of explainability consisting of 12 distinct features, under which they grouped the quantitative evaluation methods used in the surveyed papers. The authors argue that explainability should not be treated as a binary characteristic but rather a collection of properties, the features of their definition of explainability, each of which can be met to varying degrees. The criteria, which they call Co-12 properties are 'correctness', 'output completeness', 'consistency', 'continuity', 'contrastivity', 'covariate complexity', 'compactness', 'compositionality', 'confidence', 'context', 'coherence', and 'controllability'. The properties sometimes contradict each other, meaning that fulfilling one criterion such as completeness may compromise another, like compactness. The authors emphasize that the survey results offer several benefits, such as enabling the examination of XAI methods, providing a benchmark and criteria for evaluating different methods based on the Co-12 properties, identifying the strengths and weaknesses of various methods, and offering for developing new methods with the focus on certain properties. In the context of this master thesis, we chose to evaluate the correctness, output completeness and continuity properties of Nauta et al.'s Co-12 criteria [100]. The reason for this decision lies in the belief that these properties hold central significance. In our opinion, an explanation cannot be considered useable even if all other Co-12 criteria are fulfilled when it lacks correctness, completeness, or consistency.

5.1.3 Evaluation of Correctness, Output Completeness, and Continuity

To comprehensively understand the evaluation criteria correctness, output completeness, and continuity investigated in this study, we conducted a thorough analysis of all the papers listed by the Nauta et al. [100] for these properties. Table 9 presents a summary of papers from Nauta et al.'s survey [100] that pertain to the evaluation of XAI methods specifically explaining models using text data. The table includes the author name and citations, the Co-12 properties evaluated, the availability of a GitHub repository, whether SHAP or LIME were evaluated, the text dataset was used in the evaluation and the corresponding task the model was designed to solve. Notably, the table highlights that none of the surveyed papers evaluated an explanation method for a text-based model using fidelity. Consequently, we excluded this metric from our evaluations.

The following sections introduce the categories and methods collected and defined by Nauta et al. [100], which we employ for our evaluation.

Author	**Evaluation Methods**	**Git-Hub**	**SHAP 1 Lime +**	**Dataset for Text Data**	**Task for Text Data**
Camburu et al. [23]	Stability for Slight Variation	x	x	SNLI	Natural Language Inference; Model that provides natural language explanations
Chen et al. [27]	Incremental Deletion	1	1, +	IMDB	Sentiment Classification
Chen et al. [28]	Preservation Check	1	1, +	IMDB	Sentiment Classification
Chen et al. [29]	Incremental Deletion	1	1, +	IMDB, AG News, Yahoo Answers	Classification
Cheng et al. [31]	Single Deletion	1	x	Yelp, Movielens	Recommender Systems
Han et al. [61]	Incremental Deletion	1	x	Stanford Sentiment Treebank, Multi-Genre NLI (MNLI), HANS	Sentiment Classification, Natural Language Inference
Kumar and Talukdar [76]	Preservation Check; Deletion Check	1	x	Stanford NLI dataset	Natural Language Inference; Model that provides natural language explanations
Li et al. [81]	Preservation Check; Stability for Slight Variation	x	x	Machine Translation	Translation
Liang et al. [83]	Model Parameter Randomization Check; Preservation Check; Deletion Check; Stability for Slight Variations	1	1	IMDB	Sentiment Classification

Luo et al. [89]	Preservation Check	1	x	Real World Chinese Financial Website Documents	Sentiment Classification on Financial Data
Mohankumar et al. [94]	Incremental Deletion	1	x	IMDB, Stanford Sentiment Treebank, Yelp, Amazon, Twitter ADR, 20 Newsgroups, MIMIC ICD9 (Anemia, Diabetis), SNLI, Quora Question Paraphrase, bAbI, CNN News Articles	Sentiment Classification, Classification, Natural Language Inference, Paraphrase Detection, Question Answering
Ramamurthy et al. [99]	Incremental Deletion	x	+	ATIS	Classification
Serrano and Smith [121]	Single Deletion; Incremental Deletion	1	x	IMBD, Yelp, Amazon, Yahoo	Sentiment Classification, Classification
Yeh et al. [149]	Preservation Check	1	1	IMDB	Sentiment Classification
Yuan et al. [151]	Preservation Check	1	x	MR (Movie Reviews), AG News	Sentiment Classification, Classification

Table 9: Table of papers listed by Nauta et al. [100] who evaluate text data

5.1.4 Correctness

In the context of evaluation methods, the correctness category seeks to address the question of how faithful, or how closely, the explanations mimic the underlying model they aim to explain. It is crucial to distinguish this notion of correctness from the user's perception of how reasonable the explanations are.

Model Parameter Randomization Check. The Model Parameter Randomization Check, proposed by Adebayo et al. [4], serves as a 'sanity check' to assess the dependence of explanations on a trained model's parameters. This check involves comparing the explanations obtained for the trained

Author	Metric
Liang et al. [83]	Model randomization test as suggested by Adebayo et al. [4]: to evaluate calculate cosine correlation between binary masks (the one created from randomization and the original one).

Table 10: List of metrics for Model Parameter Randomization Check for text data

Author	Metric
Cheng et al. [31]	Evaluate the effectiveness: Pearson correlation coefficient between the influence reported by the method and the influence computed from leave-one-out retraining (removing a feature and retraining the model multiple times).
Serrano and Smith [121]	Investigate if zeroing the highest attention weight results as compared to randomly chosen weight result in a decision flip.

Table 11: List of metrics for Single Deletion for text data

model with those of a model of the same architecture but with randomly initialized parameters. The rationale behind this check is that the explanations should vary based on the model's trained parameters. Hence, if the explanations for both models appear similar, it indicates that the explanation method may not rely significantly on the model parameters. This is of course only a 'sanity check' and does not definitively determine the reliance of the explanation method on the model parameters. As suggested by Nauta et al. [100], to enhance the assessment's robustness, it may be prudent to repeat the parameter reinitialization process multiple times, to account for the possibility of chance similarities between the reinitialized and trained models. Table 10 lists the metrics for the Model Parameter Check for text data.

Single Deletion. Single Deletion measures the change in model output when one feature is deleted, based on the explanation's feature importance ranking. The underlying assumption is that removing or replacing a high ranking feature will have a more pronounced impact on the model's output compared to a feature with lower importance according to the explanation method. Additionally, this evaluation approach allows for the identification of 'null attributes' - features that have no impact on the model's output. If the explanation is accurate, these null attributes should also be assigned an importance score of zero [77]. Metrics for the Single Deletion test are presented in Table 11.

Incremental Deletion (or Incremental Addition). Similar to the Single Deletion method, the Incremental Deletion (or Incremental Addition) method involves the deletion or addition of features, but does so iteratively. The process can be performed in a top-down manner, starting with the most important features according to the explanation method, or in a bottom up manner, starting with the least important features. Since manipulating input features individually might be computationally expensive, due to the presence of numerous features, some researchers choose to group the deletions or additions, to reduce the costs, for example by considering the top or bottom 10% of features. Several authors have been mentioned in this context. Shrikumar et al. [124] consider the difference in log-odds score to be calculated between the original model output and the deletion iterations. Samek et al. [118] propose the use of the Area over the Perturbation Curve (AOPC) as a metric. However, Nauta et al. [100] caution against solely relying on AOPC, as it assumes that only a handful of features account for the majority of the importance, which might be valid for softmax scores, but not necessarily for other measures. For our evaluation, we report both Log-odds scores and AOPC scores.

Log-odds Score. The Log-odds score is computed by calculating the negative logarithmic probabilities of the predicted class. This involves taking the logarithm of the model's probability for the predicted class and subtracting the logarithm of the probability of not being the predicted class, see Listing 2. The negative logarithmic probability after masking the data is subtracted from the for the original outputs and averaged over all samples for our implementation. Lower log-odds scores are preferable, as they indicate better performance [27]. Table 12 outlines the metrics related to Incremental Deletion for text data.

AOPC. The Area over the Perturbation Curve (AOPC) quantifies the average change in prediction probability for the predicted class when deleting a percentage 'k' features from the inputs [102, 118]. Please refer to Listing 3 for our implementation details. Higher AOPC values are considered favorable, as they suggest that the deleted words have a substantial impact on the model's decision [27].

Criticism of Single and Incremental Deletion. It is essential to consider criticisms regarding Single Deletion and Incremental Deletion, which can inadvertently create out-of-distribution examples when features are deleted from the input [25, 64, 68]. One possible solution for this problem is to retrain the model on the small input (e.g. Remove and Retrain (ROAR) proposed by Hooker et al. [64]). However, this approach introduces a new issue, as the evaluation would no longer be conducted on the original model. Alternatively, other strategies involve using replacement values from the original data distribution, employing a synthetic dataset [68] or generating new data, using for example a Generative Adversarial Network (GAN) [25].

Incremental Deletion/Addition and Output Completeness. Incremental Deletion/Addition can evaluate not only evaluate the correctness criterion but also output completeness, assessing whether the important features are sufficient to explain the behavior of the model. Incremental Addition tests this by using only the important features as input, and a complete output explanation should produce results similar to the original input. In contrast, Incremental Deletion tests for output completeness, when all features are deleted, where a wrong decision from the model would be expected. Additionally, Incremental Deletion/Addition can also assess Compactness, a Co-12 feature not evaluated in this thesis.

5.1.5 Output Completeness

The output completeness category evaluates the extent to which the model behavior is explained by the explanation method.

Preservation and Deletion Check. The Preservation and Deletion Check assesses the output completeness of the explanation method. This evaluation method involves selecting the top x features presented by the explanation method and then either using only those top features (preservation) or excluding them while using the rest (deletion) as input to the model. The resulting accuracy score is expected to either show minimal change in case of preservation, indicating that the top features are indeed important, as suggested by the explanation method. Conversely, for deletion, a significant drop is anticipated, highlighting that the top features are crucial for the model's decision. Kumar and Talukdar [76] and Liang et al. [83] conduct both Preservation and Deletion Check, while the other authors listed in Table 13 only conduct a Preservation Check.

Author	**Metric**
Chen et al. [27]	Deleting top k% words and calculate the average change in the prediction probability of the predicted class over the entire test data (AOPC); Log-odds score: average difference of negative logarithmic probabilities of the predicted class of the original input and the input where top k% of features are masked with special token <pad>; cohesion-score evaluate the interactions between words within a given text span.
Chen et al. [29]	Calculate the change in log-odds scores of the original input and after masking the top k% features based on their importance scores (k=0-20%).
Han et al. [61]	Sanity check: remove the 10% most positively, negatively, least influential or random training examples and report the average change in prediction confidence of the retrained model to the original model
Mohankumar et al. [94]	To assess importance ranking within hidden representations, the method involves random permutations of attention weights and the examination of resultant differences in the model's output (total variation distance), with the computation of Pearson correlation and JS divergence between the attribution and the attention distribution.
Ramamurthy et al. [99]	Explanation infidelity (following Yeh [148]): calculates how faithful the explanations of the black box model under perturbations are (changing numerical values to 0 or least frequent categorical value, no mention of text data) averaging the result over all test data.
Serrano and Smith [121]	Multiple weights test: erasing representations from the top of the ranking downward until the model's decision changes; Jensen-Shannon (JS) divergences—of the model's original output distribution vs output distribution after removing (highest attention vs random component) visualize the result of the subtraction of the two JS divergences over the difference in attention weights.

Table 12: List of metrics for Incremental Deletion for text data

Author	**Metric**
Chen et al. [28]	Post-hoc accuracy: compute the accuracy of the model the top k (10) word with the with (other) unselected words masked by zero paddings.
Kumar and Talukdar [76]	Sensitivity Analysis based on DeYoung [36] (based on Yu [150]): comprehensiveness: examining the impact of removing the explanation from the inputs and sufficiency: examining the impact of keeping only the explanations.
Li et al. [81]	Principled metric based on fidelity: the potential of constructing an optimal model (behaves similar to target model) based on the relevant words selected by the explanation method from the input. Different versions of this metric based on proxy models: multi-layer feedforward network (FN), recurrent network (RN), self-attention network (SA) and combination of all three as well as baseline of well-trained NMT model.
Liang et al. [83]	Fidelity for selected features: FS-M consistency between model's output with unselected features masked by zeros vs original; FS-A consistency between model's output and an approximator; FU-M, FU-A: same as above but with selected features masked all following Chen et al. [28].
Luo et al. [89]	Acc-reduced: accuracy of the trained model when using the top k weighted sentences as input; Over 50 rounds visualize average trends and the standard deviation (reflects the distributions of Acc-reduced) of top k weighted sentences (k=1-10) versus random sentences as input for the trained model.
Yeh et al. [149]	Append 5 nearest neighbors (out of 500-nearest neighbors) of each concept (one concept at a time) to the end of all testing instances versus 5 random sentences and compare prediction score.
Yuan et al. [151]	Matching rate: create new sentence out of 5 nearest neighbors representing 3 locations with the highest contribution to the decision, feed the new input to model, to obtain another classification result. Calculate the rate of new and the old classification result being equal (matching).

Table 13: List of metrics for Deletion and Preservation Check for text data

5.1.6 Continuity

The continuity evaluation method aims to assess the generalizability of an explanation.

Stability for Slight Variations. Stability for Slight Variations measures the continuity of an explanation by investigating similarity between the explanation for the original sample and a slightly altered sample. This metric is referred to using different terms in the literature, such as 'stability' [8], 'sensitivity' [83, 148], and 'robustness' [111, 126]. The measurement of similarity can vary, depending on the type of input data. For our evaluation, we use Local Lipschitz values [8], which quantify the difference between the explanations for the original and the input data with small perturbations. This

Author	**Metric**
Camburu et al. [23]	Sanity check: framework for creating inconsistent natural language explanations via adversarial generation of new input.
Li et al. [81]	we could not find such an evaluation in this paper.
Liang et al. [83]	SEN: sensitivity score of the influence of adversarial examples on the feature importance scores, proposed by Yeh et al. [148]; not reported for text dataset.

Table 14: List of metrics for Stability for Slight Variations for text data

approach is based on the idea that small changes in the input data should not cause significant changes in the explanations. Naylor et al. [101] report using this metric to assess the explanation's robustness. Listed in Table 14 are the metrics relevant to evaluating Stability for Slight Variation with text data.

5.2 Evaluation of SHAP Explanations of the Fake News Evaluation Model

This section describes the implementation and the results of the evaluation of the correctness, output completeness and continuity of SHAP on fake news and sentiment classification tasks. The code for this implementation can be found at `https://github.com/elengue/eval_SHAP_BERT_text`.

When an evaluation method requires the deletion or masking of tokens, we always mask the original tokens with the special [PAD] token, as it seemed the most appropriate approach. This also ensured that the length of the input data remained unchanged. Although we did not evaluate the effect of the input data length on classification, we deemed it appropriate to maintain a constant length. In cases where a replacement scheme based on the output of the SHAP values was used, we referred to it as the 'selected' treatment, where specific tokens were selected based on their corresponding SHAP values. The opposite of this is the randomized treatment, where the tokens to be masked are randomly chosen. Additionally, the term 'document' is also sometimes used and refers to a single statement from the input data.

For the correctness criterion, we conducted both the Model Parameter Randomization Check and the Incremental Deletion test. Since an Incremental Deletion Test is made up of multiple single deletions with different variables, we do not report results for Single Deletion separately. Instead, the results of the Incremental Deletion test can be considered as indicative of the single deletion test.

5.2.1 Model Parameter Randomization Check

As a preliminary step, we perform the Model Parameter Randomization Check as a 'sanity check'. While Nauta et al. [100] recommend performing this check multiple times, we could only do so once per dataset due to computational limitations. Nevertheless, the results showed a noticeable difference between the fine-tuned and the randomly initialized models, indicating that the risk of coincidentally obtaining similar weights from random re-initialization did not occur in our case.

During the implementation of this check, careful consideration was given to which parameters should be randomized. Since our model is a fine-tuned model of an already pretrained 'bert-base-uncased' model, there were several options for randomization. For simplicity, we opted for a fully reinitialized model, having the same architecture as both our fine-tuned and the 'bert-base-uncased' model, but with weights are randomly initialized.

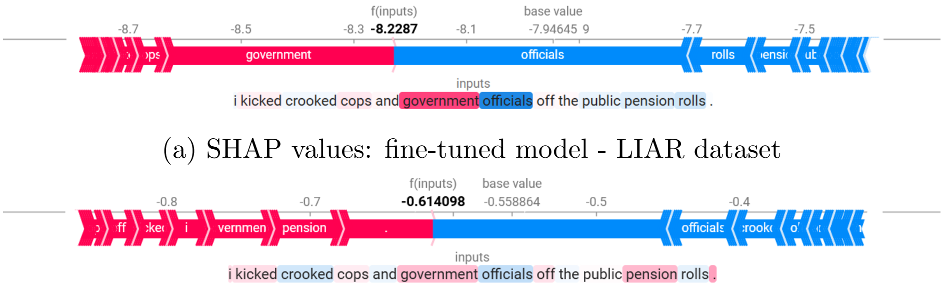
Implementation. We employed the `BertConfig`[43] configuration class to initialize the default model without pretrained weights. The `BertForSequenceClassification` was instantiated using the empty configuration, as the default configuration corresponds to the 'bert-base-uncased' model architecture utilized for fine-tuning our models. Subsequently, the randomized model was employed when loading the SHAP explainer and all other steps remained consistent, including the use of the same 1,000 data instances per dataset, as utilized for computing the original SHAP values. The computations were performed using a NVIDIA A100-SXM4-40GB GPU via Google Colab Pro, with run times comparable to the computation of the SHAP values with 38 minutes and 17 seconds for the LIAR dataset, 57 minutes and 23 seconds for the Fake News dataset and one hour 15 minutes and 19 seconds for the IMDB dataset.

Results. In order to analyze the results, we first compare the text plots of the SHAP values for one randomly chosen example per dataset of the fine-tuned model and the corresponding example from both the fine-tuned model and the corresponding example from the randomized model. Figure 8a displays the results of the fine-tuned model for the LIAR dataset, while Figure 8b shows the results of the randomly initialized model. While, Figure 8 exhibits the most similarities in the proportions of positive and negative values (represented by red and blue coloring), a closer examination reveals distinct numeric values and highlighting. Similar comparisons are conducted for the Fake News dataset in Figure 9, and the IMDB dataset in Figure 10. These comparisons highlight a consistent pattern: there are clear distinctions between the results produced by the fine-tuned models and the randomized model. In essence, the local examples demonstrate clear differences between the two models for all datasets.

Next, we conduct a comparison at the global level, revealing substantial differences between the fine-tuned and randomized SHAP values. No single word is in both the top tokens of the fine-tuned and randomized SHAP values, as demonstrated in Figure 11, Figure 12, and Figure 13. In order to quantitatively assess the difference in SHAP values between the two models, we calculated the Spearman Correlation Coefficient as proposed by Liang et al. [83]. The results were produced using the `eval_correlation`[44] function provided by Liang et al. in the GitHub repository for their paper. The results are reported in Table 15, with both the absolute values and the original sign

[43] `https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertConfig`

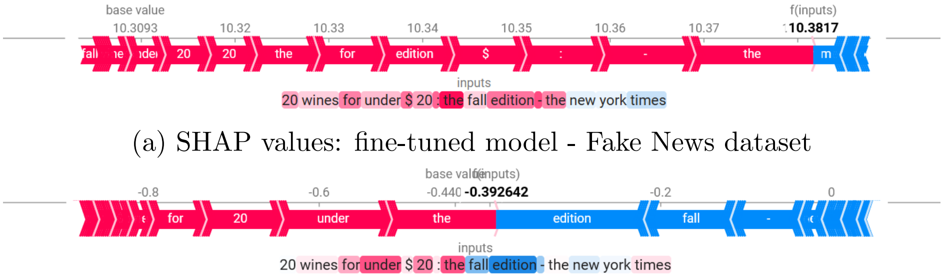
[44] `https://github.com/langlrsw/MEED/blob/master/imdb/eval_methods.py`



(a) SHAP values: fine-tuned model - LIAR dataset

(b) SHAP values: randomized model - LIAR dataset

Figure 8: SHAP values: fine-tuned vs. randomized model - LIAR dataset

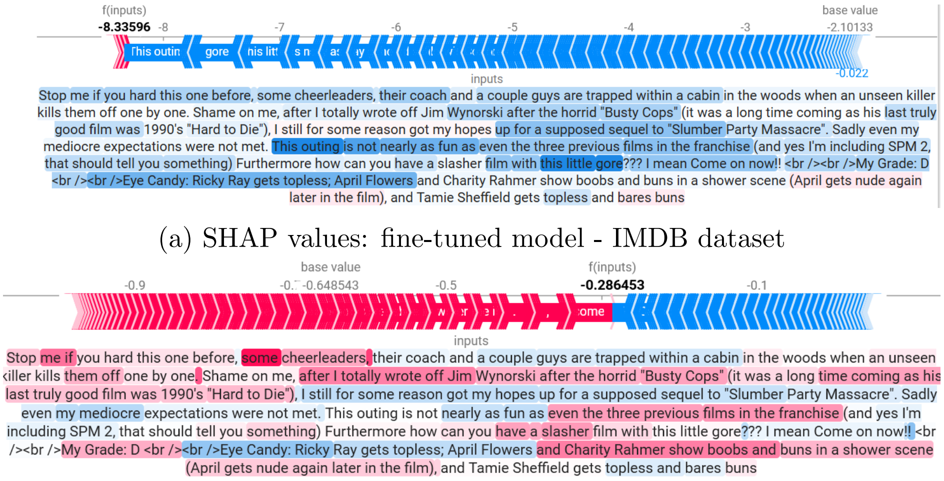


(a) SHAP values: fine-tuned model - Fake News dataset

(b) SHAP values: randomized model - Fake News dataset

Figure 9: SHAP values: fine-tuned vs. randomized model - Fake News dataset

values presented. Assuming that Liang et al. reported the results for the absolute values, we obtained a similar result for the IMDB dataset with 12.16% compared to their reported value of 9.39%. Notably, IMDB dataset's Spearman Correlation Coefficient is the highest among all datasets and both treatments, with -11.86%, being a close second, which corresponds to the value we report for the original values of the LIAR dataset. In summary, these findings for the Model Parameter Randomization Check indicate that the explanations are highly dependent on the model.



Stop me if you hard this one before, some cheerleaders, their coach and a couple guys are trapped within a cabin in the woods when an unseen killer kills them off one by one. Shame on me, after I totally wrote off Jim Wynorski after the horrid "Busty Cops" (it was a long time coming as his last truly good film was 1990's "Hard to Die"), I still for some reason got my hopes up for a supposed sequel to "Slumber Party Massacre". Sadly even my mediocre expectations were not met. This outing is not nearly as fun as even the three previous films in the franchise (and yes I'm including SPM 2, that should tell you something) Furthermore how can you have a slasher film with this little gore??? I mean Come on now!! <br /><br />My Grade: D <br /><br />Eye Candy: Ricky Ray gets topless; April Flowers and Charity Rahmer show boobs and buns in a shower scene (April gets nude again later in the film), and Tamie Sheffield gets topless and bares buns

(a) SHAP values: fine-tuned model - IMDB dataset

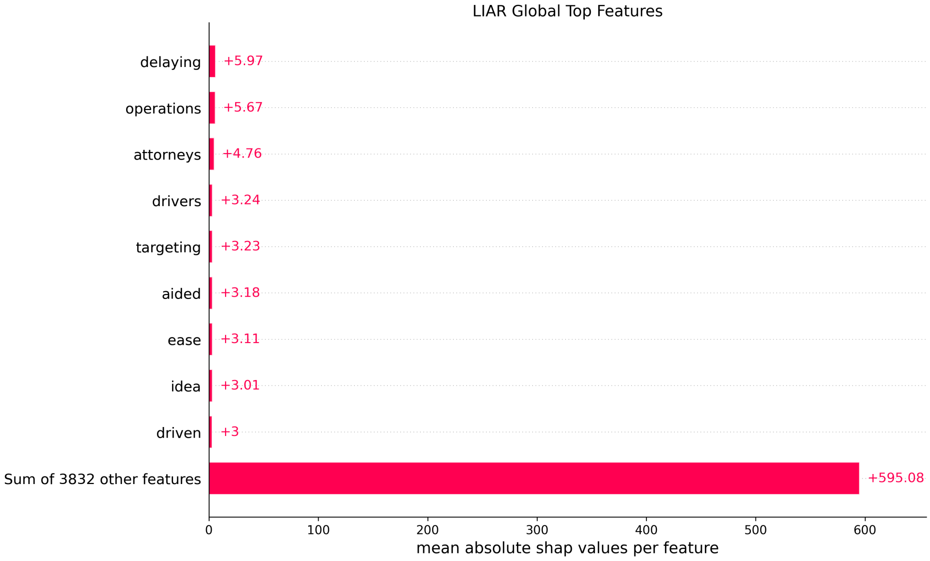
Stop me if you hard this one before, some cheerleaders, their coach and a couple guys are trapped within a cabin in the woods when an unseen killer kills them off one by one. Shame on me, after I totally wrote off Jim Wynorski after the horrid "Busty Cops" (it was a long time coming as his last truly good film was 1990's "Hard to Die"), I still for some reason got my hopes up for a supposed sequel to "Slumber Party Massacre". Sadly even my mediocre expectations were not met. This outing is not nearly as fun as even the three previous films in the franchise (and yes I'm including SPM 2, that should tell you something) Furthermore how can you have a slasher film with this little gore??? I mean Come on now!! <br /><br />My Grade: D <br /><br />Eye Candy: Ricky Ray gets topless; April Flowers and Charity Rahmer show boobs and buns in a shower scene (April gets nude again later in the film), and Tamie Sheffield gets topless and bares buns

(b) SHAP values: randomized model - IMDB dataset

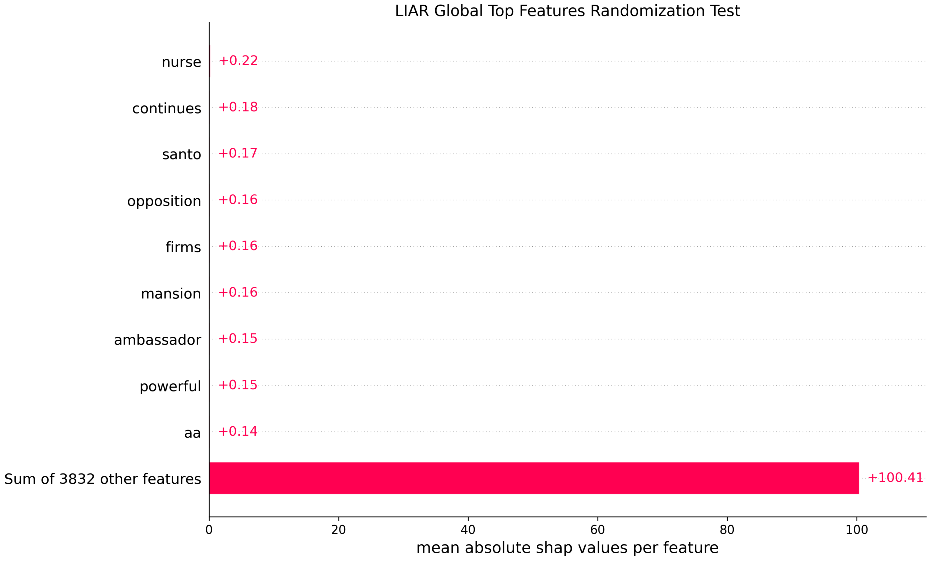
Figure 10: SHAP values: fine-tuned vs. randomized model - IMDB dataset

	Fake News	**LIAR**	**IMDB**
Absolute Values	-0.0063	0.0135	0.1216
Original Values	-0.0126	-0.1186	0.0017

Table 15: Spearman Correlation Coefficient for the Model Randomization Test

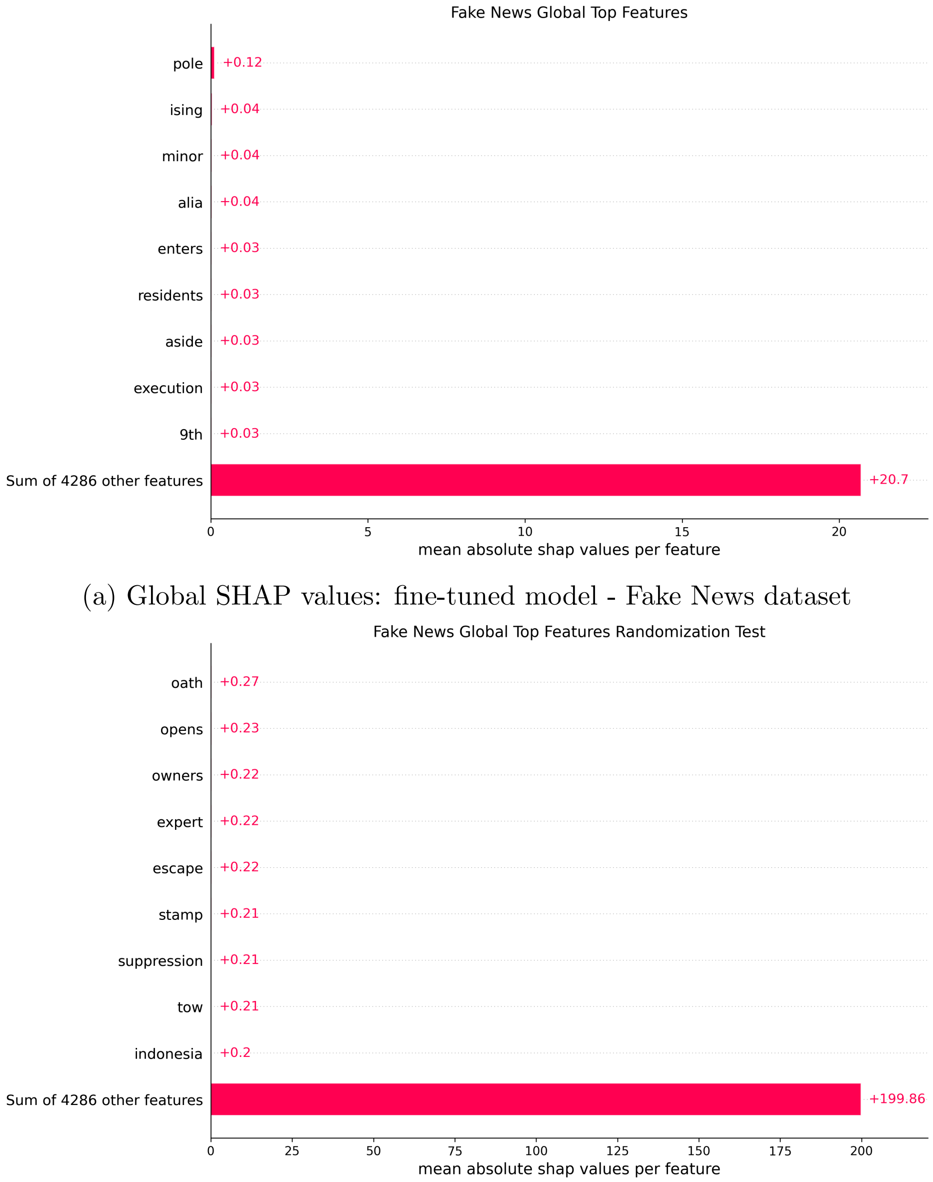


(a) Global SHAP values: fine-tuned model - LIAR dataset



(b) Global SHAP values: randomized model - LIAR dataset

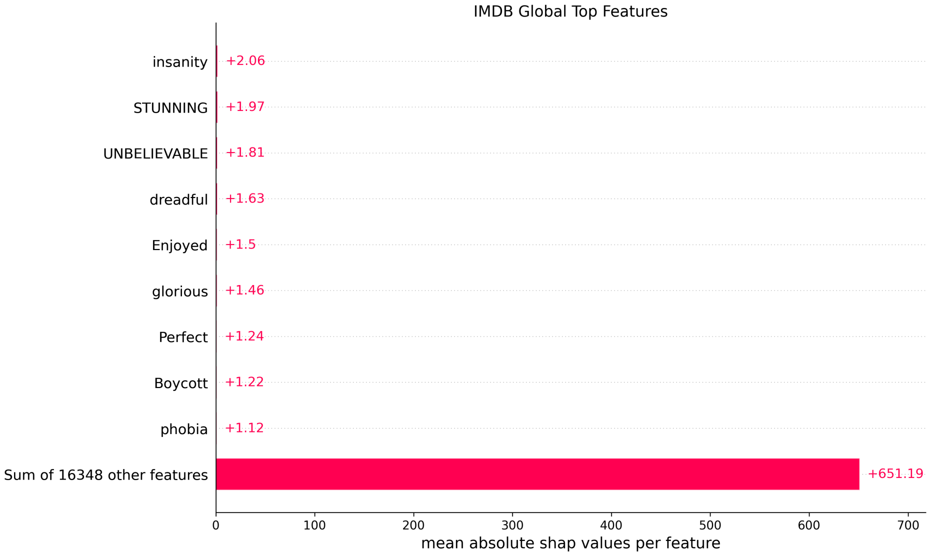
Figure 11: Global SHAP values: fine-tuned vs. randomized model - LIAR dataset



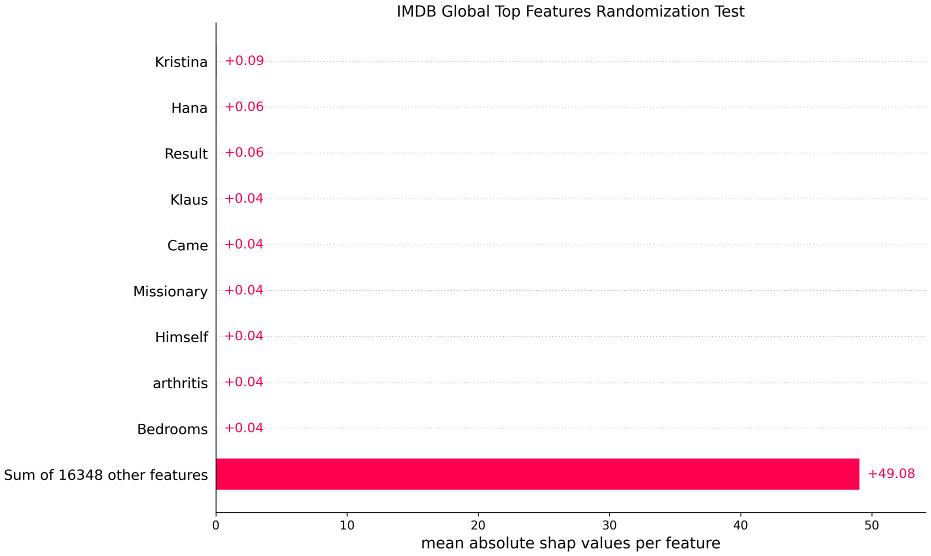
(a) Global SHAP values: fine-tuned model - Fake News dataset

(b) Global SHAP values: randomized model - Fake News dataset

Figure 12: Global SHAP values: fine-tuned vs. randomized model - Fake News dataset



(a) Global SHAP values: fine-tuned model - IMDB dataset



(b) Global SHAP values: randomized model - IMDB dataset

Figure 13: Global SHAP values: fine-tuned vs. randomized model - IMDB dataset

5.2.2 Incremental Deletion (or Incremental Addition)

In this evaluation, we conduct both an Incremental Deletion and an Incremental Addition test. The Incremental Deletion involves masking tokens in order of importance, starting with the most important tokens at a global level and progressively removing less important tokens using the [PAD] token. On the other hand, the Incremental Addition test consists of adding the globally most important tokens to an empty string composed of [PAD] tokens until all tokens are included. To measure the performance of these tests, we report the Log-odds scores and Area over the Perturbation Curve (AOPC)metrics, following the approach outlined by Chen et al. [27]. For detailed numerical results, please refer to Appendix A.1.1 and A.1.2.

Implementation. To conduct the Incremental Deletion and Incremental Addition tests, we first designed the 'get_global_feature_list' function. This function creates a global feature list containing all tokens in the dataset, ordered based on the average SHAP values. The SHAP values are computed for each token, representing their contributions to the model predictions. The global feature list facilitates understanding the relative importance of tokens across the entire dataset. The SHAP package does not offer a built-in functionality to generate global feature lists directly. However, for visualizations `shap.plots.bar(shap_values)`[45] provides the same results in graph form. Based on a suggestion from the SHAP GitHub repository[46], we implemented the 'get_global_feature_list' function. The `replace_global_highest_elements` function is utilized to perform both the Incremental Deletion and Incremental Addition tests. This function allows us to replace either the top k% of features in the entire 1,000-sample dataset (Incremental Deletion), or everything except the top k% of tokens (Incremental Addition). Additionally, we also create datasets in which the same amount of tokens are replaced as in the Incremental Deletion or Addition test. However, instead of choosing the tokens based on the SHAP values, we replace randomly selected tokens to establish a baseline for comparison.

The 'replacement_type' variable in the 'replace_global_highest_elements' function determines which test is executed. If 'replacement_type' is set to `top`, the Incremental Deletion test is performed, replacing the top k% of the global features or a random sample of the same size from the 'global_features' dataframe with the [PAD] token. On the other hand, if 'replacement_type' is set to `bottom`, the Incremental Addition test is performed, replacing the bottom 100-k% of the global features or a random sample of the same size

[45] which is equivalent to `shap.plots.bar(shap_values.abs.mean(0))`

[46] https://github.com/slundberg/shap/issues/632

```python
def neg_log_prob(modified_prob, original_prob, cls):
    return np.log(modified_prob[cls] + 1e-6) - np.log(1 - original_prob[cls] + 1e-6)
```

Listing 2: Calculate the negative logarithmic probabilities

```python
def AOPC(original_probs, modified_probs, cls):
    return (original_probs[cls]-modified_probs[cls]).item()
```

Listing 3: Calculate the AOPC

with the [PAD] token. In order to measure the performance of these tests, we report the average of the Log-odds score and the AOPC for each value of k for each dataset. The softmax function is applied to the model outputs to ensure their suitability for computation[47]. The Log-odds score and the AOPC for each document in the 1,000 input data are computed. We used the same implementation as Li et al.[48] to compute the 'lor_selected', see Listing 2. Listing 3 shows our implementation of the AOPC.

Results. For the purpose of the Incremental Deletion and Addition test, we examine the effects of global token deletion and addition on the Log-odds score and AOPC across varying values of k. These k values correspond to the percentage of global tokens, ranging from 0% to 100% in increments of 10%. In all figures, the solid lines represent the results of the selected treatment, where tokens were masked in order of importance based on the SHAP values. Conversely, the dashed and more transparent lines represent the results of the random treatment, where tokens were masked randomly.

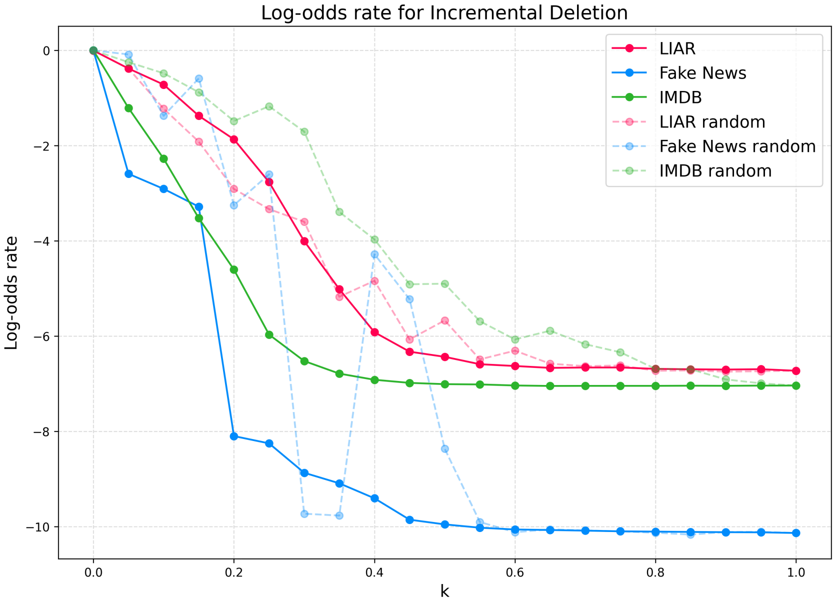
Incremental Deletion. The Log-odds scores and the AOPC scores for the Incremental Deletion evaluation are depicted in Figures 14a and 14b respectively. Both, the Fake News and the IMDB dataset exhibit lower Log-odds scores and higher AOPC scores compared to the randomized control, indicating better performance for the selected treatment. However, in the Fake News dataset, at k values of 0.3 and 0.35, there is a slight deviation with a drop in the Log-odds score

[47] https://github.com/icrto/xML/issues/1

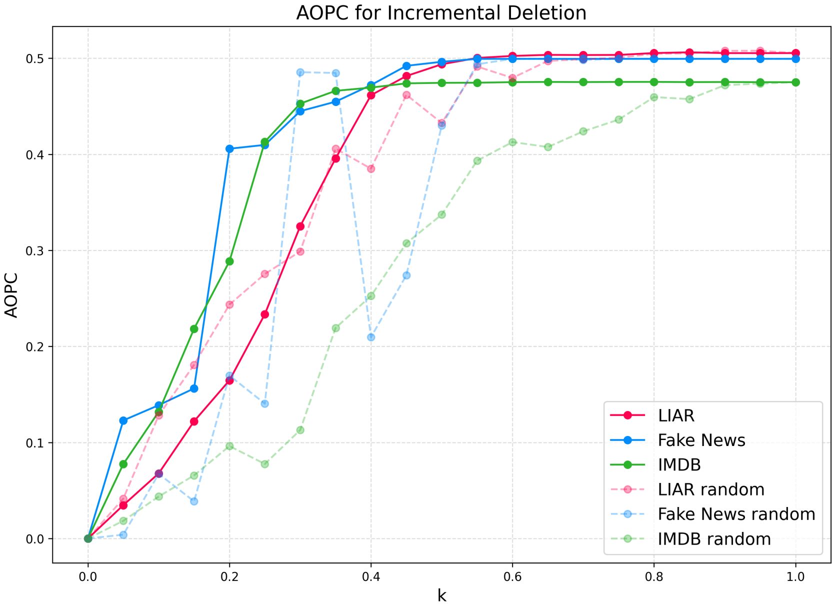
[48] https://github.com/Jianbo-Lab/LCShapley/blob/master/texts/utils.py

and a spike in the AOPC in the random control. We speculate that this anomaly may be due to random chance, leading to the selection of important tokens. Overall, the results demonstrate the expected behavior with convergence of selected and the randomized results at k value of 0.6 for the Fake News dataset and at 0.9 for the IMDB dataset. The difference in converge points may be attributed to the average token length disparity between the datasets, impacting classification.

For the LIAR dataset, the results do not entirely follow the expected trends. The random control exhibits lower Log-odds scores for k values below 0.25, and both selected and randomized results show similar scores at k values of 0.3 and 0.35. However, beyond this point, the Log-odds and AOPC scores align with the expected behavior of lower Log-odds scores and higher AOPC scores for the selected treatment compared to the randomized control. Notably, the selected and randomized values converge at a k value of 0.65. We suspect that the unusual behavior observed in the LIAR dataset is a result of the low accuracy score of the underlying model. This, in turn, impacts the quality of the explanations and, consequently, the behavior in this evaluation.



(a) Log-odds Score for Incremental Deletion for all three datasets



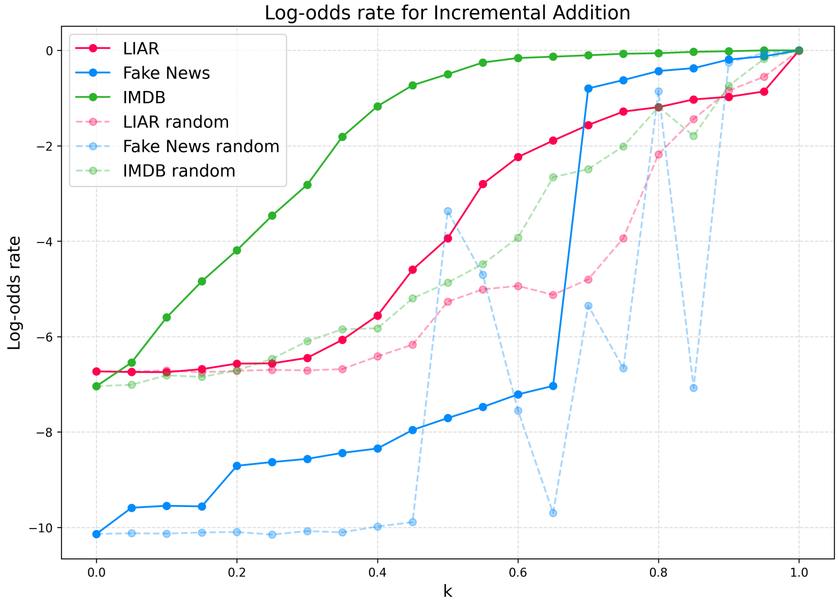
(b) AOPC for Incremental Deletion for all three datasets

Figure 14: Results for Incremental Deletion for all three datasets

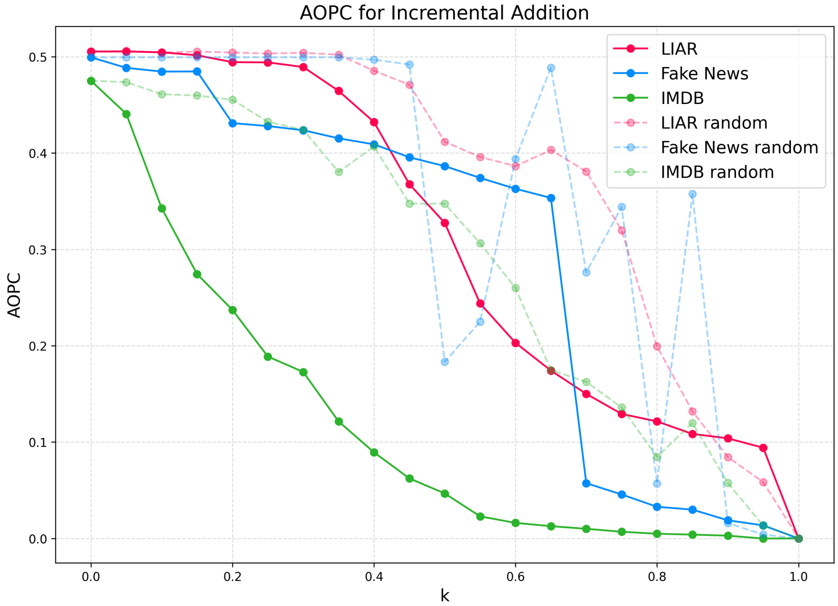
Incremental Addition. The results of the 'Incremental Addition' test are shown in Figures 15a and 15b for the Log-odds scores and AOPC, respectively. In this test, tokens are incrementally added to the dataset, requiring a different interpretation of results. Smaller AOPC values indicate that the addition of less relevant tokens has a less impact on the model's prediction, making the expected outcome. Similarly, higher Log-odds scores are considered more desirable.

The results of the Incremental Addition Test exhibit similarities to the Incremental Deletion test, albeit in reverse order. For the IMDB dataset, the selected treatment shows a rapid increase in Log-odds scores and a corresponding decrease in AOPC, while the randomized treatment maintains higher AOPC scores and lower Log-odds scores throughout all k values. The Fake News dataset, shows unstable behavior in the random treatment, with fluctuations starting at a k value of 0.5 and continuing to the end. This behavior may be due to the freshly randomized sampling of tokens for each k value, resulting in varying proportions of significant or insignificant tokens. The selected treatment in the Fake News dataset displays a more stable incline/decline, with a notable increase/decrease at the k value of 0.7. At k values of 0.5 and 0.6, the selected treatment's Log-odds and AOPC scores outperform the randomized treatment, corresponding to the first significant spike observed in the randomized treatment. In the LIAR dataset, both AOPC and Log-odds are similar up to a k value of 0.3, after which they develop as expected, with higher Log-odds scores of the selected treatment and lower AOPC scores. The treatments then switch positions at a k value of 0.9 until k reaches 1.

Overall, the results of the Incremental Deletion and Incremental Addition tests are as expected, the deletions/additions based on SHAP values have higher impact on the metrics as compared to the random deletions/additions, with noted and discussed exceptions for the LIAR dataset and some random behavior in the Fake News dataset.



(a) Log-odds Score for Incremental Addition for all three datasets



(b) AOPC for Incremental Addition for all three datasets

Figure 15: Results for Incremental Addition for all three datasets

5.2.3 Preservation and Deletion Check

The Preservation and Deletion Check presents a different approach to evaluating token importance, it evaluates the output completeness criterion of the Co-12 properties, which evaluates how much of a model's behavior is explained by the explainability method. It focuses on local importance within individual documents rather than global importance across the entire dataset, as seen in the Incremental Deletion/Addition test. This evaluation method involves deleting tokens from each document in the dataset in order of their importance, as determined by SHAP. The results of the Preservation and Deletion Check are reported using two evaluation metrics: The first metric, referred to as the 'acc-reduced' by Luo et al. [89], and referred to as 'post-hoc accuracy' by Chen et al. [28], measures the accuracy of the model after the input is altered. In this thesis, we adopt the terminology used by Luo et al. and refer to it as 'acc-reduced'. The second metric assesses whether the alteration of the input data leads to a different classification by the model. Yuan et al. [151] refer to this metric as 'matching rate', while Liang et al. [83] term it 'consistency'. For clarity and intuitiveness, we refer to it as 'label flip', as used by Flores and Hao [49], which is also similar to Serrano and Smith's [121] term 'decision flip'. For detailed numerical results, please refer to Appendix A.1.3 and A.1.4.

Implementation. The implementation of the Preservation and Deletion Check is based on the version provided by Chen et al. [28] [49] of these evaluations. For the Deletion Check, we created a function called `replace_k_highest_elements`, responsible for replacing k tokens with the [PAD] token. The 'replacement type' can be set to `top` to replace the k absolute highest SHAP values, thereby performing the Deletion Check. Alternatively, for the Preservation Check, the 'replacement type' is set to `bottom` replacing everything except the k absolute highest SHAP value tokens. Additionally, as a benchmark, we also replace k tokens randomly. After perturbing the data as specified above, the data is transformed, encoded, and fed back to the model for classification. We then compared the resulting classification of the perturbed data with the ground truth labels to calculate the 'acc-reduced' metric. Moreover, the comparison between the model's classification of the perturbed data and the original model's results yields the 'label flip' metric.

Results. To evaluate the Preservation and Deletion Check, we examine the effects of token deletion and addition on the model's accuracy and change

[49] https://github.com/Jianbo-Lab/L2X/tree/master/imdb-sent

in label prediction for varying values of k. For the LIAR and Fake News datasets, we performed the Deletion and Preservation Check with k values ranging from zero to 14 in one-token increments. As the IMDB dataset has significantly longer documents, we chose a different range of k values, deleting zero to 140 tokens in 10-token increments to ensure a comparable perturbation, see Table 6.

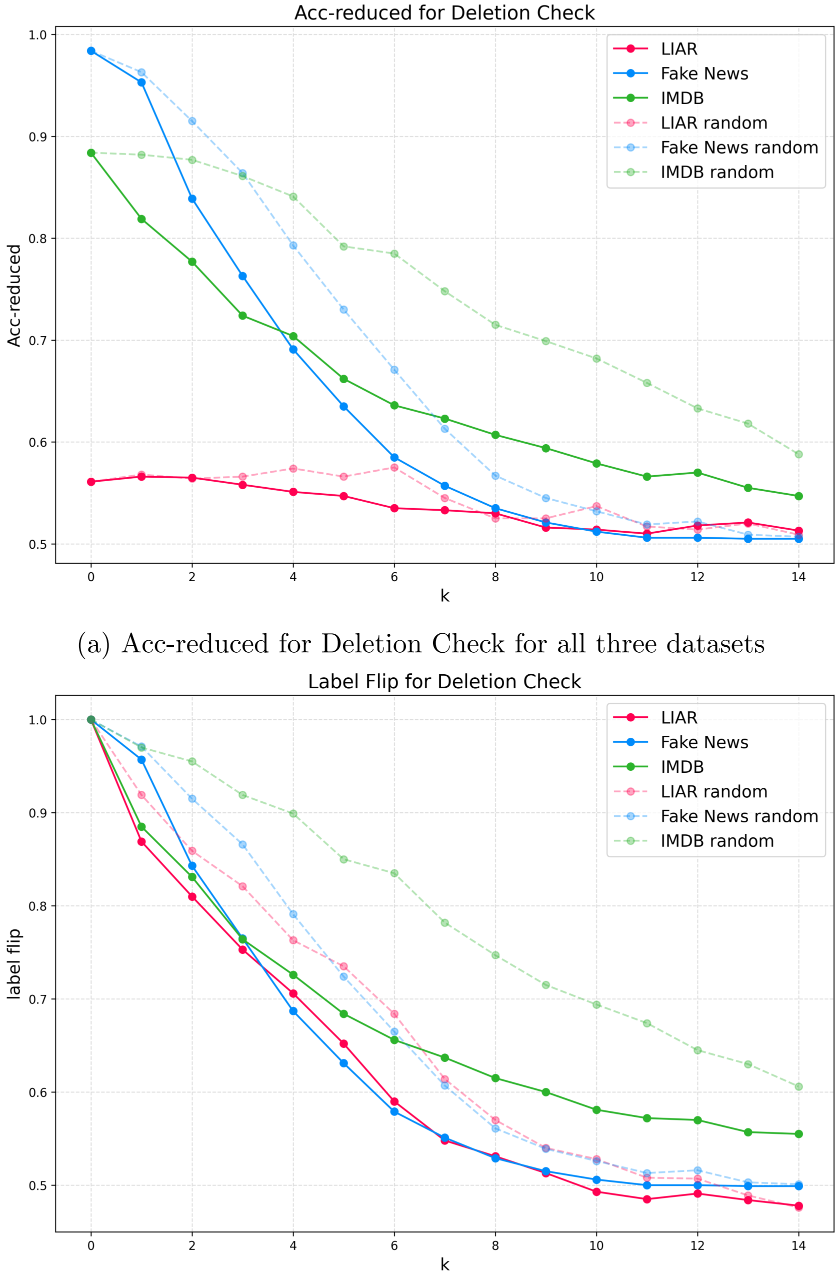
Deletion Check. The Deletion Check's expected behavior suggests that the acc-reduced scores will decrease more rapidly for tokens deleted based on importance SHAP values than randomly deleted tokens. Similarly, labels would flip sooner for inputs altered based on token importance compared to random alterations. For k=0 so no tokens having been changed, the flip rate is 1 indicating that no label has flipped.

The obtained results, as shown in Figure 16a for acc-reduced and Figure 16b for label flip, align with expected behavior. Each dataset has different starting points, which are the accuracy scores of the original model. The LIAR dataset has a poor original model accuracy of 0.56, the Fake News dataset has an accuracy score of 0.98, and the IMDB dataset one of 0.88, see Table 8 for detailed accuracy scores.

For the LIAR dataset, no significant difference is observed between the results based on the SHAP values and the ones based on random replacement, for k values one to three. For k values four to seven, the acc-reduced scores of random replacement are higher than the SHAP-based ones, as expected. The values become similar again for k values eight and nine, and once more, random replacement outperforms selected treatment for k 10 to 14.

For the Fake News and IMDB datasets, the acc-reduced scores clearly differentiate between random replacement and the selected treatment, consistent with the expected behavior. The difference in scores is more significant for the IMDB dataset compared to the Fake News dataset.

The label flip metric follows the expected behavior for all datasets, see Figure 16b, even for the LIAR dataset, where acc-reduced did not exhibit the expected trend. This might imply that the label flip metric is more informative when accuracy is low, making it a better metric in such cases. This finding suggests that SHAP can provide meaningful explanations even for models with poor accuracy.



(a) Acc-reduced for Deletion Check for all three datasets

(b) Label Flip for Deletion Check for all three datasets

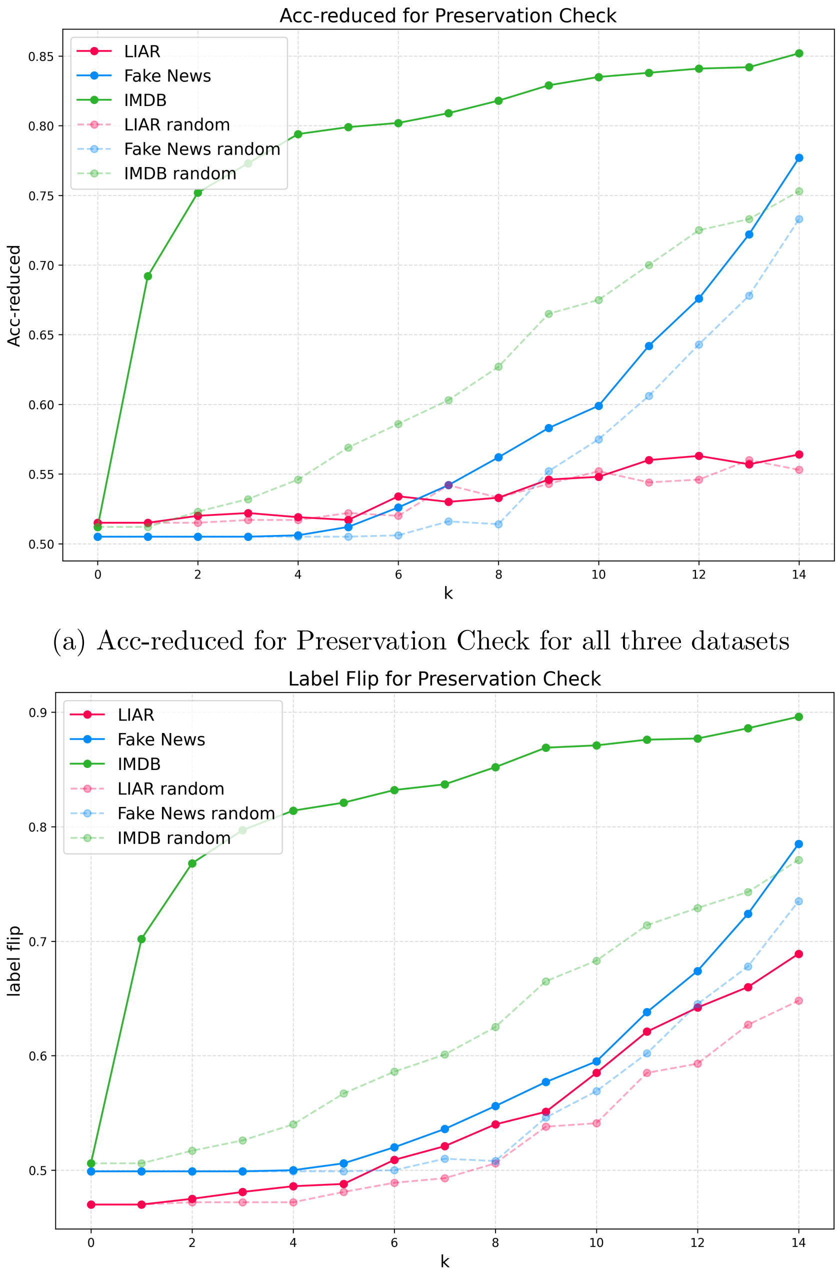
Figure 16: Results for Deletion Check for all three datasets

Preservation Check. Results for the Preservation Check need to be interpreted oppositely to the Deletion Check, as the evaluation adds the most important tokens incrementally to an otherwise empty input filled with [PAD] tokens.

The Preservation Check, see Figures 17a and 17b exhibits specific characteristics, resulting in a rapid increase in 'acc-reduced' scores for the IMDB dataset with selected features, compared to the datasets classifying fake news. This difference might be due to sentiment analysis's reliance on a few tokens to determine sentiment, while fake news classification may require more tokens for classification. Additionally, the 10-token increments used for the IMDB dataset could have a more substantial impact on the Preservation Check results.

Apart from the bigger difference between the random and the selected treatments for the IMDB dataset for both acc-reduced and label flip metrics, the results for the Preservation Check are similar to the Deletion Check. For the Fake News dataset, a clear difference between the random and selected treatment emerges starting at five tokens. For the LIAR dataset, the acc-reduced scores are similar between the random and the selected treatment, but for the label flip metric show a difference between, with the values closest for nine tokens.

Overall, the results of the Deletion and Preservation Check align with the expected behavior, indicating that the explanations are output complete.



(a) Acc-reduced for Preservation Check for all three datasets

(b) Label Flip for Preservation Check for all three datasets

Figure 17: Results for Preservation Check for all three datasets

5.2.4 Stability for Slight Variations

The Stability for Slight Variations evaluation addresses the robustness of the explanations by comparing the similarity of explanations between the original and a slightly altered sample. None of the papers categorized by Nauta et al. [100] evaluate Stability for Slight Variations of a text classification task, Liang et al. [83] do not report their sensitivity (SEN) metric for their text data evaluation and Camburu et al. [23] evaluate natural language explanations. However, a study conducted by Naylor et al. [101] addresses this gap with their methodology, available on their GitHub repository[50]. Naylor et al. evaluated SHAP and LIME on traditional models such as Logistic Regression, Random Forest, Explainable Boosting Machine. However, for the transformer-based method Bigbird [153], they chose different explainability methods - IntegratedGradients [133] and Saliency [125] - because of SHAP's and LIME's lower quality explanations for deep learning models [8].

For calculating the 'local Lipschitz values' function that provides the normalized difference between the SHAP values of the original and the perturbed input divided by the Euclidean distance of the original and perturbed embeddings normalized by the document length.

Implementation. To perform token perturbations, we initially attempted to use the replacement algorithm introduced by Naylor et al. [101], which employs a text perturbation method that replaces the token through next neighbor sampling. However, this method frequently resulted in replacing tokens with [unusedxxx] BERT tokens. The [unusedxxx][51] tokens are tokens designed to be replaced with domain specific words when doing further pretraining or fine-tuning and are otherwise randomly initialized, indicating an issue with the replacement procedure. We opted to utilize the TextAttack Python Package[52] Token Replacement Schema. Specifically, we employed the `Word Swap by BERT-Masked LM`[53] transformation, which generates possible word replacements for a MLM. As the method for the word swap we chose 'bert-attack' which is TextAttack's implementation of the data augmentation technique introduced by Li et al. in their paper 'BERT-ATTACK: Adversarial Attack Against BERT Using BERT' [82]. For single-token words, the MLM, in this case, our fine-tuned BERT model, provides a replacement to-

[50] https://github.com/mnaylor5/quantifying-explainability

[51] https://stackoverflow.com/questions/62452271/understanding-bert-vocab-unusedxxx-tokens

[52] https://github.com/QData/TextAttack

[53] https://textattack.readthedocs.io/en/latest/apidoc/textattack.transformations.word_swaps.html#word-swap-by-bert-masked-lm

ken. For multi-token words, the replacement is more difficult, since not all combinations of the sub-word tokens would make sense. The top-k possible replacements, all possible combinations of sub-words, are ranked based on the perplexity measure from the MLM. The `RepeatModification()`[54] constraint was set to prevent repeated modifications on words. Following Naylor et al., we set the percentage of words to be swapped at 10% of the tokens per sample and used 'high_yield' and 'fast_augment' for computational efficiency. The 'transformations_per_example' was set to 100, with samples being skipped if perturbations were not created within four minutes.

For the IMDB dataset, no perturbed texts could be created even when setting the 'transformations_per_example' to 30 and the length of time before skipping to the next document to up to 20 minutes led to no perturbed texts at all. We assume the reason for this failure is due to the much higher token length per document. Therefore, we only report results for the other two datasets. Before calculating the local Lipschitz values, we check if our perturbed texts include 15 instances with a Euclidean distance between the embedding matrices of less than 0.75. We adjusted the EPSILON values, representing the L-infinity ball's radius, from 0.25 (the value employed by Naylor et al. [101]) to 0.75. We arrived at this change after some trial and error because we couldn't find any perturbed texts with distances small enough at the original 0.25 value. Consequently, we only use documents with at least 15 perturbed texts having a radius of less than 0.75 to compute the local Lipschitz values. We use the function `local_lipschitz` provided in Naylor et al.'s repository to calculate the local Lipschitz values. It takes the difference of the SHAP values of the original and the perturbed texts and divides it with the Euclidean distance.

Results. Unfortunately, perturbed texts could not be generated for the IMDB dataset, even with adjustments in the 'transformations_per_example'. Thus, we only report results for the LIAR and Fake News datasets. Additionally, we set the epsilon value from 0.25, which is the value that Naylor et al. used, to 0.75 because we did not get any perturbed texts for which the distances were small enough. The local Lipschitz values for both datasets were calculated using the `local_lipschitz` function, provided in Naylor et al.'s repository. Specifically, for the LIAR dataset, the maximum local Lipschitz value is 4.8080, whereas for the Fake News dataset, it amounts to 0.7237, as illustrated in Table 16. Figures 18 and 19 display the local Lipschitz values in boxplot format for both datasets. For a figure showing both boxplots

[54] `https://textattack.readthedocs.io/en/latest/apidoc/textattack.constraints.pre_transformation.html#repeat-modification`

	Fake News	**LIAR**
Maximum Local Lipschitz	4.8080	0.7237

Table 16: Maximum Local Lipschitz values for the Stability for Slight Variation Test

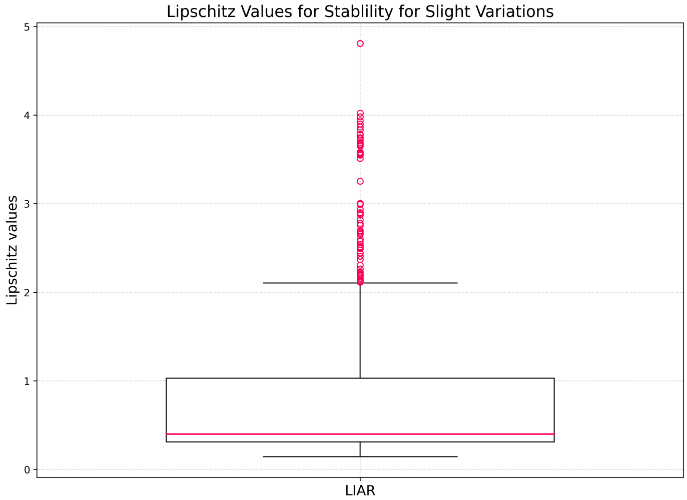


Figure 18: Results for the Stability for Slight Variation for the LIAR dataset

at the same scale, please refer to Appendix A.2 Figure 20. For the LIAR dataset, there are several outliers with values between 4.8 and 2, while the majority of values are closer to zero to two. The Fake News dataset shows similar results but on a smaller scale, with outliers between 0.72 and 0.54, and the majority of values ranging from 0.53 to 0.1. Comparing these results with Naylor et al.'s findings for traditional machine learning models (ranging from 0.1 to 2.1) and transformer-based models (maximum local Lipschitz values of 21 and 32), our results (0.1 to 4.8) appear promising. However, it is essential to consider the differences in radius used, and potential selection bias introduced by the perturbation method's limitations. In summary, the Stability for Slight Variations evaluation demonstrates promising results for both datasets, indicating a degree of robustness in the explanations provided by our approach.

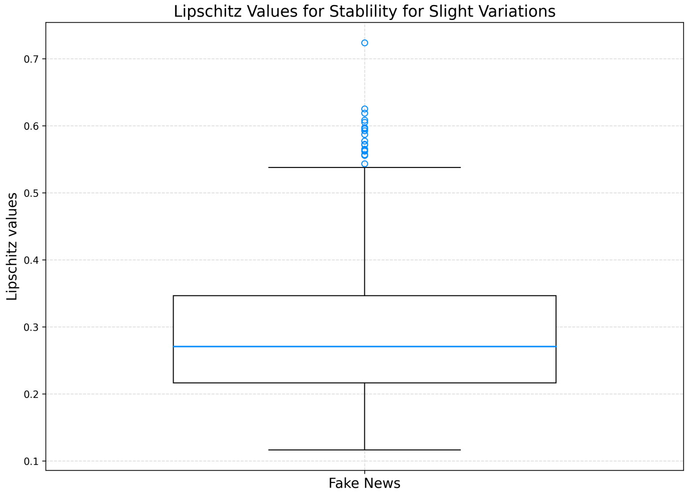


Figure 19: Results for the Stability for Slight Variation for the Fake News dataset

5.2.5 Summary of Results

This section provides an overview of the results of the previous section, specifically Table 17 shows a summary of the results presented in the previous sections.

For the Model Parameter Randomization Check, the Spearman Correlation Coefficients indicate that the explanations are highly dependent on the model.

In terms of the Incremental Deletion and Incremental Addition tests, we present the results for k=0.2, a value commonly used in the literature [27, 29]. It is worth noting that we are only presenting the result of one increment, this would, strictly speaking, be considered preforming the Single Deletion test. For the complete results, refer to Sections 5.2.2 and A.1. Overall, the results of the Incremental Deletion and Incremental Addition tests align with expectations. Deletions and additions based on SHAP values have a higher impact on the metrics compared to random deletions and additions, indicating the correctness of SHAP's outputs. It is important to note, that there were exceptions observed for the LIAR dataset, as well as some random behavior in the Fake News dataset. Notably, for k=0.2, the Incremental Deletion test showed the expected behavior, with the Fake News and IMDB datasets exhibiting higher AOPC values and lower log-odds scores compared to the random control, while the LIAR dataset did not follow the expected behavior consistently. The Incremental Addition test results are numerically closer, but they still follow the expected behavior of lower AOPC values and higher log-odds scores. In this case, the LIAR dataset also follows the expected behavior.

For the Deletion and Preservation Check, we present the results for k=10 for the LIAR and Fake News datasets and k=100 for the IMDB dataset, as k=10 is commonly used in the literature [28, 89, 83]. We discussed the reasons for the different k value for the IMDB dataset in Section 5.2.3. Overall, the results of the Deletion and Preservation Check align with the expected behavior, indicating that the explanations are output-complete. For the Deletion Check, the accuracy scores are lower for the selected treatment compared to the randomly perturbed control, as expected. This finding suggests that the tokens identified by SHAP have more significant impact on the accuracy as compared to randomly selected ones. The flip rate also follows expected behavior, with higher values for the random treatment, indicating that the selected treatment has a higher impact on the change in labels than the random treatment. Regarding the Preservation Check, we mostly observed the expected behavior, with higher acc-reduced and flip rate scores, except for the randomized control for the LIAR dataset, which showed a higher accuracy reduction than the selected treatment. As discussed, see Section 5.2.3, we assume this is due to the overall low accuracy rate for the LIAR dataset.

Due to implementation limitations, we only report the results for the Stability for Slight Variations evaluation for the LIAR and Fake News dataset. These results indicate a degree of robustness in the explanations, meaning slight changes in the input data do not lead to big changes in the model's predictions.

Evaluation Category	Metric	LIAR	Fake News	IMDB
Model Parameter Randomization Check	absolute	-0.0063	0.0135	0.1216
	original	-0.0126	-0.1186	0.0017
Incremental Deletion	AOPC	0.16469	0.40587	0.28863
	AOPC rand	0.24356	0.16970	0.09627
	log-odds	-1.86845	-8.10128	-4.60414
	log-odds rand	-2.91056	-3.25118	-1.48369
Incremental Addition	AOPC	0.49440	0.43110	0.23720
	AOPC rand	0.50440	0.49943	0.45535
	log-odds	-6.56619	-8.70773	-4.18947
	log-odds rand	-6.71837	-10.0963	-6.70994
Deletion Check	acc-red	0.514	0.512	0.579
	acc-red rand	0.537	0.532	0.682
	flip	0.493	0.506	0.581
	flip rand	0.528	0.526	0.694
Preservation Check	acc-red	0.548	0.599	0.835
	acc-red rand	0.552	0.575	0.675
	flip	0.585	0.595	0.876
	flip rand	0.541	0.569	0.683
Stability for Slight Variation	max local Lipschitz	4.808	0.7237	-

Table 17: Overview of the results of the evaluations

6 Conclusion

In this chapter, we give a short summary of the work we conducted, restate, and answer our main research question: "*How can we effectively evaluate the suitability of SHAP for interpreting text data, particularly in the context of fake news detection using BERT, with a focus on assessing correctness, output completeness, and continuity?*" as well as our sub-questions, discuss the limitations of our thesis and explore possible directions for future work.

6.1 Summary and Findings

In order to achieve our research objectives, we began by introducing the field of XAI. First, we define the intended audience, list some of the purposes XAI aims to achieve, clarify the terminology associated with explainability and present the specific definition of explainability used within this thesis. Subsequently, we present a taxonomy designed to categorize XAI methods and apply it to 23 notable explainability methods. Following this, we provide a theoretical overview of our fake news detection model, based on BERT and introduce the datasets we utilized. Furthermore, we provide a short background on SHAP our chosen explainability method and apply it to our fake news detection model.

We provided an overview of evaluation methods for assessing correctness, output completeness and continuity relevant to text data, referencing the original papers and authors. We also indicated, the type of evaluation, whether there is a GitHub repository available, whether SHAP or LIME had been evaluated using these methods, along with the datasets and text data tasks used. Additionally, we describe each evaluation category and evaluation test and list what the evaluations exactly entail. This provides a simple way to get an overview of the available evaluation methods and answers the question of how to effectively evaluate the suitability of XAI methods on text data.

In order to answer sub-question one (SQ1): "*To what extent does SHAP demonstrate correctness, align with the original model's behavior in terms of faithfulness, when applied to BERT's fake news detection?*" we performed the Model Parameter Check, which serves as a sanity check for the findings. The visualizations and the Spearman Correlation Coefficients collectively indicate a strong reliance on the model for the provided explanations. Furthermore, the results from the Incremental Deletion and Incremental Addition Tests, another evaluation we conducted to assess the correctness criterion, align with our expectations. In comparison to the random changes, the modifications made based on the SHAP values show a greater impact on the metrics, confirming the reliability of SHAP's outputs. However, it's essential to note that there were some exceptions observed for the LIAR dataset, and instances of random behavior were noted within the Fake News dataset.

Sub-question two (SQ2) which asks: "*How does SHAP perform in terms of output completeness, providing sufficient information to explain the output of the model when explaining fake news detection by BERT?*" was addressed by conducting the Deletion and Preservation Check. Overall, the results of the Deletion and Preservation Check are consistent with the expected behavior, suggesting that the explanations provided by SHAP are indeed output complete. The deletions and preservations conducted based on SHAP values show a more pronounced effect as compared to the randomly created alterations.

The final sub-question three (SQ3) with the following text: "*How well does SHAP demonstrate continuity, indicating the generalizability of its findings, for BERT's fake news detection?*" was answered through the execution of the Stability for Slight Variation test. It is important to note that, due to certain implementation limitations, we are exclusively presenting the results for the Stability for Slight Variations evaluation for the LIAR and Fake News datasets. The results collectively imply that the explanations generated by SHAP maintain a certain level of stability and consistency, even when subjected to minor variations in the input data. Essentially, this indicates that slight changes in the input data do not result in significant alterations in the model's predictions, highlighting the generalizability of SHAP's findings for BERT's fake news detection.

In summary, the analysis conducted on fake news datasets shows that SHAP satisfies the three selected Co-12 properties proposed by Nauta et al. [100]: correctness, output completeness, and continuity for the conducted evaluations. However, certain limitations were identified, especially when dealing with black box models that exhibit relatively low accuracy, as observed with the LIAR dataset.

6.2 Limitations

As with any research project, this thesis has several limitations that warrant careful consideration.

Scope Limitations. Firstly, the scope of a master thesis inherently imposes limitations on the depth and breadth of the study. Notably, we focused on examining only one black box model, BERT, which has a strict limit on the length of input text. This choice might not fully capture the interpretability challenges posed by more complex models like BigBird [153], which accommodate longer input sequences. Additionally, the black box model used for the LIAR dataset exhibited a relatively low accuracy rate of 0.56, potentially affecting the explainability of the results (Table 8). We were also limited in the number of evaluation methods we could implement, leading us to focus on only three Cs of the Co-12. Ideally, a comprehensive evaluation would encompass all 12 criteria or, at the very least, include all criteria related to the content part of the Co-12, including consistency, contrastivity, and covariant complexity [100].

Dataset Limitations. Another limitation lies in the selection of datasets. Broadly speaking, the classification of fake news and, consequently, the development of datasets for fake news classification proves to be difficult and, to a degree, subjective. The choice of well-known and high-quality datasets is, therefore, important. While the LIAR dataset is widely recognized and satisfies nine out of ten criteria listed by D'Ulizia et al. [44], the Kaggle Fake News dataset is less established. In addition, the use of a dataset which not only includes the titles, such as the LIAR and Fake News dataset, but also the whole text of the (fake) news articles, alongside any pictures, would be a more realistic representation. In connection to this change in dataset, a black box model that accepts longer and mixed data type input, such as BigBird [153], as previously mentioned, would be necessary. Additionally, the choice of not only including two fake news datasets, but also the IMDB dataset which is a sentiment detection dataset, might lead to some confusion to the reader and the inadvertent comparison between two unrelated classification tasks. While this choice was motivated by the intention to facilitate the comparison of the results of this work with previous works in the realm of XAI evaluation, we acknowledge that is can create some confusion.

XAI Limitations. Regarding the choice of XAI method, we only evaluated one method, SHAP. Although it is a popular method and frequently cited in the literature (i.e. it was mentioned in eleven out of 15 surveys we investigated, see Table 1), it has some shortcomings, particularly for Deep Neural Networks like BERT. This limitation is evident in Naylor et al.'s [101] decision not to evaluate SHAP on transformer-based explainability methods.

Computational Limitations. Computational constraints further limited our ability to explore the full potential of some evaluation methods. For instance, we could only explain 1000 instances per dataset due to computational limitations, and a larger sample size might have yielded more robust results. This limitation also stopped us from performing the Model Parameter Randomization Check multiple times as recommended. The limited resources also prevented us from performing evaluations that required retraining the model, such as the methods proposed by Han et al. [61] and Cheng et al. [31]. Moreover, we could not perform certain evaluations, like Luo et al.'s [89] preservation check over 50 rounds to plot average trends, or create an approximator as Liang et al. [83] did, for comparison purposes.

Suitability Limitations. Furthermore, certain evaluation methods were omitted due to unavailable resources or incompatibility with our approach. For instance, the Explanation Randomization Check [94, 127] and the White Box Check [32, 70, 159] were not implemented as they required explanations built into the model and a white box model, respectively. Similarly, the Controlled Synthetic Data Check [5, 52, 68, 81, 103, 109] was omitted, as it exceeded the scope of this thesis to create a synthetic text dataset. The Predictive Performance method [30, 32, 114, 155] from the output completeness group was dropped because we do not have a white box model or predictive explanations. In the continuity evaluation methods, we did not include the evaluation of Fidelity to Slight Variations [78, 108], as our focus was not on evaluating a decision rule or a white box model. Additionally, the Connectedness [147, 80, 105] measure was left out of the analysis due to the absence of a counterfactual explanation model.

Implementation Limitations. In the Preservation Check, we only calculated the decision flip and not the Jensen-Shannon divergence, which was performed by Serrano and Smith [121]. Moreover, we did not compute the comprehensiveness and sufficiency scores, as noted by Kumar and Talukdar [76] based on DeYoung [36] (based on Yu [150]), as it would have exceeded the scope of this thesis. Furthermore, investigating different replacement options for tokens other than the [PAD] token could have been valuable to assess the impact of the choice of replacement token on evaluation results or whether simply deleting the tokens would yield better outcomes.

In the Deletion Check [28, 76, 83, 89], the values and increment steps for k could have been chosen differently. While we selected different values for k based on the higher token count in the IMDB dataset, making the results comparable, it is not ideal to compare results for two different k values.

In conclusion, this thesis has identified several limitations that should be carefully considered when interpreting the results. In order to build upon this work and address these constraints, the next section discusses possible directions of future research.

6.3 Future Work

Moving forward, future research in XAI should build upon the identified limitations and explore various avenues to enhance the understanding and application of XAI methods. By addressing these aspects, we can pave the way for more interpretable and trustworthy AI systems. In order to advance the field of XAI evaluation, it is essential to assess multiple datasets from various application areas with different objectives. Moreover, evaluating multiple models and explainability methods will provide a more comprehensive understanding of their strengths and weaknesses. A promising direction is the creation of an evaluation framework that systematically assesses XAI methods based on the Co-12 criteria. Such a standardized framework would enable researchers to compare and select the most suitable XAI method for their specific tasks, thereby achieving optimal explainability. Quantus, an evaluation framework for neural network explanations introduced by Hedström et al. [62], takes a significant step towards fulfilling this goal. It allows for evaluation across various data types and provides a range of evaluation metrics, including faithfulness, robustness, localization, complexity, and randomization, or axiomatic metrics. NLP evaluations are in development for future versions of Quantus, but are as of the writing of this thesis not yet available. Such benchmarking and evaluation tools should be widely adopted to establish a standardized practice for introducing and comparing XAI methods.

While functional evaluations using proxy tasks are crucial initial steps, future research should go beyond and consider human-grounded and application-grounded evaluation approaches, as proposed by Doshi-Velez and Kim [41]. Involving real humans in simplified or real tasks can provide valuable insights into the practicality and usefulness of XAI methods in real-world scenarios. For instance, understanding how XAI methods impact the decision-making process of human when detecting fake news can be a compelling avenue to explore. In summary, future work in XAI should encompass diverse datasets, multiple models, and various evaluation frameworks to foster the development of reliable and applicable XAI techniques. By addressing the identified limitations and embracing a multi-faceted evaluation approach, we can move towards more interpretable and trustworthy AI systems.

References

[1] Munich security conference. Accessed: 09.08.2023. Available online: https://www.who.int/director-general/speeches/detail/munich-security-conference.

[2] Disinformation and Russia's war of aggression against Ukraine. *OECD*, August 2023. Accessed: 09.09.2023. Available online: https://www.oecd.org/ukraine-hub/policy-responses/disinformation-and-russia-s-war-of-aggression-against-ukraine-37186bde.

[3] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, PP:1–1, 09 2018.

[4] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[5] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

[6] IPN SIG AI. Dutch artificial intelligence manifesto. TU Delft, 2018. Accessed: 02.07.2023. Available online: http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf.

[7] Mustafa A Al-Asadi and Sakir Tasdemir. Using artificial intelligence against the phenomenon of fake news: a systematic literature review. *Combating Fake News with Computational Intelligence Techniques*, pages 39–54, 2022.

[8] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

[9] Gina Cherelus Amy Tennery. Microsoft's AI Twitter bot goes dark after racist, sexist tweets. *U.S*, March 2016.

[10] Sule Anjomshoae, Timotheus Kampik, and Kary Främling. Py-ciu: A python library for explaining machine learning predictions using contextual importance and utility. In *IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI)*, 2020.

[11] Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020.

[12] Alejandro Barredo Arrieta, 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.

[13] AB Athira, SD Madhu Kumar, and Anu Mary Chacko. A systematic survey on explainable ai applied to fake news detection. *Engineering Applications of Artificial Intelligence*, 122:106087, 2023.

[14] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[15] Richard A Berk and Justin Bleich. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Pub. Pol'y*, 12:513, 2013.

[16] Umang Bhatt, 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.

[17] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

[18] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076*, 2021.

[19] Jan vom Brocke, Alan Hevner, and Alexander Maedche. Introduction to Design Science Research. In *Design Science Research. Cases*, pages 1–13. Springer, Cham, Switzerland, September 2020.

[20] Ceren Budak. What happened? the spread of fake news publisher content during the 2016 us presidential election. In *The World Wide Web Conference*, pages 139–150, 2019.

[21] Nadia Burkart and Marco F Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

[22] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. A survey on xai and natural language explanations. *Information Processing & Management*, 60(1):103111, 2023.

[23] Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. Make up your mind! adversarial generation of inconsistent natural language explanations. *arXiv preprint arXiv:1910.03065*, 2019.

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

[25] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*, 2018.

[26] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[27] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. *arXiv preprint arXiv:2004.02015*, 2020.

[28] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR, 2018.

[29] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data, 2018.

[30] Runjin Chen, Hao Chen, Jie Ren, Ge Huang, and Quanshi Zhang. Explaining neural networks semantically and quantitatively. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9187–9196, 2019.

[31] Weiyu Cheng, Yanyan Shen, Linpeng Huang, and Yanmin Zhu. Incorporating interpretability into latent factor models via fast influence analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 885–893, 2019.

[32] Jonathan Crabbe, Yao Zhang, William Zame, and Mihaela van der Schaar. Learning outside the black-box: The pursuit of interpretable models. *Advances in Neural Information Processing Systems*, 33:17838–17849, 2020.

[33] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[34] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *U.S*, October 2018.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[36] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

[37] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31, 2018.

[38] Jürgen Dieber and Sabrina Kirrane. A novel model usability evaluation framework (muse) for explainable artificial intelligence. *Information Fusion*, 81:143–153, 2022.

[39] Jia Ding, Yongjun Hu, and Huiyou Chang. Bert-based mental model, a better fake news detector. In *Proceedings of the 2020 6th international conference on computing and artificial intelligence*, pages 396–400, 2020.

[40] Weiping Ding, Mohamed Abdel-Basset, Hossam Hawash, and Ahmed M Ali. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, 2022.

[41] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

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

[43] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Omer Rana, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): core ideas, techniques and solutions. *ACM Computing Surveys (CSUR)*, 2022.

[44] Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518, 2021.

[45] Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of animal ecology*, 77(4):802–813, 2008.

[46] Dumitru Erhan, Aaron Courville, and Yoshua Bengio. Understanding representations learned in deep architectures. *Départment d'Informatique et Recherche Opérationnelle, Université de Montréal, Tech. Rep*, 1355:1, 2010.

[47] Kevin Fauvel, Véronique Masson, and Elisa Fromont. A performance-explainability framework to benchmark machine learning methods: application to multivariate time series classifiers. *arXiv preprint arXiv:2005.14501*, 2020.

[48] Rubén R Fernández, Isaac Martín De Diego, Víctor Aceña, Alberto Fernández-Isabel, and Javier M Moguerza. Random forest explainability using counterfactual sets. *Information Fusion*, 63:196–207, 2020.

[49] Lorenzo Jaime Yu Flores and Yiding Hao. An adversarial benchmark for fake news detection models. In *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*, 2022.

[50] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[51] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The annals of applied statistics*, pages 916–954, 2008.

[52] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020.

[53] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[54] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

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

[56] Donald P Green and Holger L Kern. Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees. In *The annual summer meeting of the society of political methodology*, pages 100–110, 2010.

[57] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

[58] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[59] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

[60] Umm-E Habiba, Justus Bogner, and Stefan Wagner. Can requirements engineering support explainable artificial intelligence? towards a user-centric approach for explainability requirements. In *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, pages 162–165. IEEE, 2022.

[61] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*, 2020.

[62] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[63] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *Management Information Systems Quarterly*, 28(1):6, 2008.

[64] Sara Hooker, 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.

[65] National Digital Skills Initiative. Ai portugal 2030, 2019. Accessed: 02.07.2023. Available online: https://www.portugal.gov.pt/download-ficheiros/ficheiro.aspx?v=3D%3DBAAAAB%2BLCAAAAAAABACzMDQxAQC3h%2ByrBAAAAA%3D%3D.

[66] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3):1353, 2022.

[67] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*, 2021.

[68] Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time

series predictions. *Advances in neural information processing systems*, 33:6441–6452, 2020.

[69] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33:4211–4222, 2020.

[70] Yunzhe Jia, James Bailey, Kotagiri Ramamohanarao, Christopher Leckie, and Michael E Houle. Improving the quality of explanations with local embedding perturbations. In *Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery & Data Mining*, pages 875–884, 2019.

[71] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788, 2021.

[72] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905. PMLR, 2020.

[73] Junaed Younus Khan, Md Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:100032, 2021.

[74] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

[75] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[76] Sawan Kumar and Partha Talukdar. Nile: Natural language inference with faithful natural language explanations. *arXiv preprint arXiv:2005.12116*, 2020.

[77] Christophe Labreuche and Simon Fossier. Explaining multi-criteria decision aiding models with an extended shapley value. In *IJCAI*, pages 331–339, 2018.

[78] Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. Robust and stable black box explanations. In *International Conference on Machine Learning*, pages 5628–5638. PMLR, 2020.

[79] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

[80] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294*, 2019.

[81] Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi. Evaluating explanation methods for neural machine translation. *arXiv preprint arXiv:2005.01672*, 2020.

[82] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*, 2020.

[83] Jian Liang, Bing Bai, Yuren Cao, Kun Bai, and Fei Wang. Adversarial infidelity learning for model interpretation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 286–296, 2020.

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

[85] Helena Löfström, Karl Hammar, and Ulf Johansson. A meta survey of quality evaluation criteria in explanation methods. In *International Conference on Advanced Information Systems Engineering*, pages 55–63. Springer, 2022.

[86] Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. Xai systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19):9423, 2022.

[87] Scott Lundberg and Su-In Lee. An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478*, 2016.

[88] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[89] Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention. In *IJCAI*, pages 4244–4250, 2018.

[90] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655, 2021.

[91] John A McDermid, Yan Jia, Zoe Porter, and Ibrahim Habli. Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A*, 379(2207):20200363, 2021.

[92] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[93] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

[94] Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. Towards transparent and explainable attention models. *arXiv preprint arXiv:2004.14243*, 2020.

[95] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems, 2020.

[96] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

[97] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. Shap-based explanation methods: A review for nlp interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, 2022.

[98] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

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

[100] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schl ötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv: 2201.08164*, 2022.

[101] Mitchell Naylor, Christi French, Samantha Terker, and Uday Kamath. Quantifying explainability in nlp and analyzing algorithms for performance-explainability tradeoff. *arXiv preprint arXiv:2107.05693*, 2021.

[102] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, 2018.

[103] Jose Oramas, Kaili Wang, and Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. *arXiv preprint arXiv:1712.06302*, 2017.

[104] Sebastian Palacio, Adriano Lucieri, Mohsin Munir, Sheraz Ahmed, Jörn Hees, and Andreas Dengel. Xai handbook: Towards a unified framework for explainable ai. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3766–3775, 2021.

[105] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of the web conference 2020*, pages 3126–3132, 2020.

[106] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[107] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[108] Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. Model agnostic supervised local explanations. *Advances in neural information processing systems*, 31, 2018.

[109] Gregory Plumb, Jonathan Terhorst, Sriram Sankararaman, and Ameet Talwalkar. Explaining groups of points in low-dimensional representations. In *International Conference on Machine Learning*, pages 7762–7771. PMLR, 2020.

[110] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[111] Nikaash Puri, Sukriti Verma, Piyush Gupta, Dhruv Kayastha, Shripad Deshmukh, Balaji Krishnamurthy, and Sameer Singh. Explain your move: Understanding agent actions using specific and relevant feature attribution. *arXiv preprint arXiv:1912.12191*, 2019.

[112] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020.

[113] Atul Rawal, James Mccoy, Danda B Rawat, Brian Sadler, and Robert Amant. Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives. *IEEE Transactions on Artificial Intelligence*, 1(01):1–1, 2021.

[114] Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems*, 33:12187–12198, 2020.

[115] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Nothing else matters: Model-agnostic explanations by identifying prediction invariance, 2016.

[116] Marco Tulio Ribeiro, 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.

[117] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[118] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[119] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700, pages 5–22. Springer Cham, 2019.

[120] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *arXiv e-prints*, pages arXiv–2105, 2021.

[121] Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

[122] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Glocalx - from local to global explanations of black box ai models. *Artificial Intelligence*, 294:103457, 2021.

[123] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

[124] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[125] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[126] Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*, 2018.

[127] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning*, pages 9046–9057. PMLR, 2020.

[128] Royal Society. Machine learning: The power and promise of computers that learn by example, 2017. Accessed: 02.07.2023. Available online: https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf.

[129] Kacper Sokol and Peter Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 56–67, New York, NY, USA, 2020. Association for Computing Machinery.

[130] Timo Speith. A review of taxonomies of explainable artificial intelligence (xai) methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, 2022.

[131] Mateusz Staniak and Przemyslaw Biecek. Explanations of model predictions with live and breakdown packages. *arXiv preprint arXiv:1804.01955*, 2018.

[132] Kyung Ho Sun, Hyunsuk Huh, Bayu Adhi Tama, Soo Young Lee, Joon Ha Jung, and Seungchul Lee. Vision-based fault diagnostics using explainable deep learning with class activation maps. *IEEE Access*, 8:129169–129179, 2020.

[133] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[134] Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. New explainability method for bert-based model in fake news detection. *Scientific Reports*, 11(1):1–13, 2021.

[135] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018.

[136] Michael Van Lent, William Fisher, and Michael Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, pages

900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[137] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[138] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

[139] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[140] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[141] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[142] Lilo Wagner. Shap's partition explainer for language models. Towards Data Science, 2020. Accessed: 29.03.2023. Available online: https://towardsdatascience.com/shaps-partition-explainer-for-language-models-ec2e7a6c1b77.

[143] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[144] Krzysztof Wecel, Marcin Sawiński, Milena Stróżyna, Włodzimierz Lewoniewski, Ewelina Księżniak, Piotr Stolarski, and Witold Abramowicz. Artificial intelligence—friend or foe in fake news campaigns. *Economics and Business Review*, 9(2), 2023.

[145] Rebecca Wexler. Opinion | When a Computer Program Keeps You in Jail. *N.Y. Times*, June 2017.

[146] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019.

[147] Prateek Yadav, Peter Hase, and Mohit Bansal. Low-cost algorithmic recourse for users with uncertain cost functions. *arXiv preprint arXiv:2111.01235*, 2021.

[148] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

[149] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020.

[150] Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*, 2019.

[151] Hao Yuan, Yongjun Chen, Xia Hu, and Shuiwang Ji. Interpreting deep models for text analysis via optimization and regularization methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5717–5724, 2019.

[152] Muhammad Rehman Zafar and Naimul Mefraz Khan. Dlime: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv preprint arXiv:1906.10263*, 2019.

[153] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

[154] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[155] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6261–6270, 2019.

[156] Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, 2020.

[157] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[158] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

[159] Zihan Zhou, Mingxuan Sun, and Jianhua Chen. A model-agnostic approach for explaining the predictions on clustered data. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1528–1533. IEEE, 2019.

A Appendix Implementation

A.1 Additional Data

A.1.1 Incremental Deletion

k	selected		random	
	log-odds	AOPC	log-odds	AOPC
0.0	-0.00392612	0.00002936	-0.00392612	0.00002936
0.05	-0.38184656	0.03474851	-0.38113147	0.04160373
0.10	-0.71909668	0.06784426	-1.22568188	0.12816179
0.15	-1.37464343	0.12209127	-1.91875781	0.18084680
0.20	-1.86844653	0.16469155	-2.91055615	0.24356367
0.25	-2.76645361	0.23355931	-3.33437085	0.27545204
0.30	-4.00546216	0.32518498	-3.60013647	0.29890098
0.35	-5.01510889	0.39570039	-5.17522217	0.40593384
0.40	-5.92034424	0.46159685	-4.83987842	0.38507960
0.45	-6.32884766	0.48154449	-6.07063037	0.46200573
0.50	-6.43659766	0.49377825	-5.67181445	0.43262911
0.55	-6.59184863	0.50036948	-6.49045166	0.49143379
0.60	-6.62948486	0.50246265	-6.30734375	0.47956479
0.65	-6.66912598	0.50356588	-6.58052100	0.49730723
0.70	-6.66120850	0.50340183	-6.63290527	0.49898166
0.75	-6.66051123	0.50351593	-6.61929590	0.50056069
0.80	-6.68966943	0.50547725	-6.72957617	0.50481270
0.85	-6.69789111	0.50622098	-6.72217334	0.50514642
0.90	-6.70391797	0.50540638	-6.74807129	0.50780969
0.95	-6.69646338	0.50537874	-6.73931543	0.50785232
1.00	-6.72904834	0.50541636	-6.72904834	0.50541636

Table 18: Data from Incremental Deletion for LIAR dataset

k	selected		random	
	log-odds	AOPC	log-odds	AOPC
0.0	-0.00011695	0	-0.00011695	0
0.05	-2.59317603	0.12298645	-0.08601118	0.00403207
0.1	-2.90963062	0.13883336	-1.37482593	0.06735068
0.15	-3.28298096	0.15628299	-0.5896723	0.03872784
0.2	-8.10128271	0.40587032	-3.25117822	0.16969807
0.25	-8.25443652	0.40983014	-2.60413599	0.1404355
0.30	-8.87174023	0.44495834	-9.73171387	0.48557283
0.35	-9.09339258	0.45485212	-9.77090234	0.48472909
0.4	-9.40825098	0.47215044	-4.28186182	0.20971341
0.45	-9.8574873	0.49216706	-5.22720996	0.27417812
0.5	-9.95612988	0.4963744	-8.3664834	0.43014187
0.55	-10.02598242	0.49939972	-9.90566895	0.49379274
0.60	-10.06341113	0.49942831	-10.12072363	0.49943232
0.65	-10.07434961	0.4994307	-10.06013574	0.49942808
0.70	-10.08742969	0.49943114	-10.08327344	0.49843298
0.75	-10.10134863	0.49943171	-10.10318652	0.49942996
0.8	-10.1079668	0.49943195	-10.13158008	0.4994327
0.85	-10.1145957	0.49943221	-10.17019727	0.49943397
0.9	-10.11984961	0.49943244	-10.11674609	0.4994323
0.95	-10.12298633	0.49943256	-10.11262109	0.49943216
1.0	-10.13699121	0.49943312	-10.13699121	0.49943312

Table 19: Data from Incremental Deletion for Fake News dataset

k	selected		random	
	log-odds	AOPC	log-odds	AOPC
0.0	0.00000016	0	0.00000016	0
0.05	-1.21153247	0.07751493	-0.24892204	0.0184306
0.1	-2.27622827	0.1319528	-0.4815405	0.04372915
0.15	-3.52476465	0.21843115	-0.88011719	0.06575675
0.2	-4.60413574	0.28863297	-1.48369006	0.09627091
0.25	-5.96924512	0.41308942	-1.17333765	0.07767533
0.30	-6.52472998	0.45286035	-1.70680994	0.11296995
0.35	-6.78939453	0.46613326	-3.39323364	0.21926708
0.4	-6.91854004	0.46950745	-3.97044092	0.25257533
0.45	-6.9839668	0.47391463	-4.91249316	0.30751829
0.5	-7.01066162	0.47445619	-4.90115137	0.33726841
0.55	-7.01675439	0.47458008	-5.68863721	0.39342666
0.60	-7.0384707	0.47518192	-6.07084863	0.41282088
0.65	-7.04859961	0.47538576	-5.88862109	0.4076257
0.70	-7.04692725	0.47525054	-6.17170605	0.42394694
0.75	-7.04653857	0.47538795	-6.34142676	0.43617425
0.8	-7.04658838	0.47544286	-6.68873193	0.45963706
0.85	-7.04331592	0.47520218	-6.69846582	0.45752609
0.9	-7.04551123	0.47529746	-6.91135596	0.47201402
0.95	-7.04116016	0.4751471	-6.99322998	0.4735903
1.0	-7.03997119	0.47511425	-7.03997119	0.47511425

Table 20: Data from Incremental Deletion for IMDB dataset

k	selected		random	
	log-odds	AOPC	log-odds	AOPC
0.0	-6.72904834	0.50541636	-6.72904834	0.50541636
0.05	-6.74285693	0.5054566	-6.73333447	0.5062301
0.10	-6.74552344	0.50461021	-6.71044092	0.50441691
0.15	-6.68350146	0.50152783	-6.74761035	0.50537812
0.20	-6.56618555	0.49439865	-6.71836523	0.50440368
0.25	-6.5620835	0.49416114	-6.69850879	0.50338774
0.30	-6.44684326	0.48935354	-6.70920801	0.50418305
0.35	-6.06865381	0.46456978	-6.68115527	0.50210989
0.40	-5.56011377	0.43220391	-6.41517627	0.48547151
0.45	-4.59760449	0.36779271	-6.16967188	0.47103553
0.50	-3.93877246	0.32752876	-5.2676709	0.41176346
0.55	-2.80002124	0.24388274	-5.00971631	0.39600393
0.60	-2.23705933	0.2030432	-4.93826514	0.38649039
0.65	-1.89127295	0.17428288	-5.12337012	0.40325192
0.70	-1.56424731	0.15007216	-4.80088086	0.38092769
0.75	-1.28331055	0.12923559	-3.93632275	0.3197923
0.80	-1.19082507	0.12147481	-2.17817188	0.19936835
0.85	-1.02842737	0.10849409	-1.43866321	0.1319553
0.90	-0.97408319	0.10394937	-0.84304797	0.08448845
0.95	-0.86212524	0.09414612	-0.55519904	0.05847695
1.00	-0.00392612	0.00002936	-0.00392612	0.00002936

Table 21: Data from Incremental Addition for LIAR dataset

A.1.2 Incremental Addition

k	selected		random	
	log-odds	AOPC	log-odds	AOPC
0.0	-10.13699121	0.49943312	-10.13699121	0.49943312
0.05	-9.58837207	0.48846993	-10.12370996	0.4994326
0.1	-9.54696484	0.4846503	-10.12979004	0.49943278
0.15	-9.55930957	0.48473863	-10.10643066	0.49942484
0.2	-8.70772656	0.43109882	-10.09626563	0.4994315
0.25	-8.63099707	0.42805946	-10.15069727	0.49943467
0.30	-8.56350098	0.42374021	-10.07733691	0.49943096
0.35	-8.4379873	0.41533822	-10.10302148	0.49943155
0.4	-8.34674609	0.40908066	-9.98133594	0.49699947
0.45	-7.95800439	0.39567853	-9.88981641	0.49205203
0.5	-7.70826953	0.38646807	-3.36901074	0.18330394
0.55	-7.47499121	0.37425467	-4.70510889	0.22505065
0.60	-7.21185352	0.36296912	-7.54922314	0.39379894
0.65	-7.0335166	0.35347065	-9.70007324	0.48873277
0.70	-0.79801715	0.05736699	-5.35102148	0.27625097
0.75	-0.62280042	0.04561079	-6.66542969	0.34418335
0.8	-0.43254053	0.03284928	-0.85533405	0.05712198
0.85	-0.37304483	0.02991163	-7.07728662	0.35766413
0.9	-0.19233707	0.01885029	-0.25411478	0.01597884
0.95	-0.12606393	0.01356455	-0.05434944	0.00392775
1.0	-0.00011695	0	-0.00011695	0

Table 22: Data from Incremental Addition for Fake News dataset

k	selected		random	
	log-odds	AOPC	log-odds	AOPC
0.0	-7.03997119	0.47511425	-7.03997119	0.47511425
0.05	-6.54343506	0.44069602	-7.01018896	0.47368897
0.10	-5.59774414	0.34273006	-6.81277441	0.46109075
0.15	-4.84169482	0.27441068	-6.84432422	0.45981271
0.20	-4.18946973	0.23720367	-6.70994385	0.45535181
0.25	-3.46526099	0.18884906	-6.46656641	0.43251884
0.30	-2.81693677	0.17280242	-6.09783203	0.4241792
0.35	-1.81115881	0.1214187	-5.84776318	0.38036545
0.40	-1.17119031	0.08938811	-5.82611963	0.40684606
0.45	-0.73162653	0.06233537	-5.19995264	0.34760192
0.50	-0.49827414	0.04671369	-4.86967627	0.34741578
0.55	-0.25450563	0.0229084	-4.47940039	0.30632564
0.60	-0.16030234	0.01612601	-3.92589673	0.26001632
0.65	-0.13265717	0.01276443	-2.65981104	0.17483781
0.70	-0.1031486	0.009991	-2.48974609	0.16252126
0.75	-0.0714341	0.00692863	-2.01095435	0.13597678
0.80	-0.05920336	0.0048718	-1.18649817	0.08469561
0.85	-0.03124971	0.00395329	-1.792474	0.11972604
0.90	-0.01829287	0.00288804	-0.74350012	0.05755134
0.95	-0.00078859	-0.00010067	-0.18057053	0.01334382
1.00	0.00000016	0	0.00000016	0

Table 23: Data from Incremental Addition for IMDB dataset

k	selected		random	
	acc	flip	acc	flip
0	0.561	1.0	0.561	1.0
1	0.566	0.869	0.568	0.919
2	0.565	0.81	0.564	0.859
3	0.558	0.753	0.566	0.821
4	0.551	0.706	0.574	0.763
5	0.547	0.652	0.566	0.735
6	0.535	0.59	0.575	0.684
7	0.533	0.548	0.545	0.614
8	0.53	0.531	0.525	0.57
9	0.516	0.513	0.525	0.54
10	0.514	0.493	0.537	0.528
11	0.51	0.485	0.517	0.508
12	0.518	0.491	0.514	0.507
13	0.521	0.484	0.52	0.489
14	0.513	0.478	0.509	0.476

Table 24: Data from Deletion Check for LIAR dataset

A.1.3 Deletion Check

A.1.4 Preservation Check

A.2 Additional Figures

k	selected		random	
	acc	flip	acc	flip
0	0.984	1.0	0.984	1.0
1	0.953	0.957	0.963	0.971
2	0.839	0.843	0.915	0.915
3	0.763	0.765	0.864	0.866
4	0.691	0.687	0.793	0.791
5	0.635	0.631	0.73	0.724
6	0.585	0.579	0.671	0.665
7	0.557	0.551	0.613	0.607
8	0.535	0.529	0.567	0.561
9	0.521	0.515	0.545	0.539
10	0.512	0.506	0.532	0.526
11	0.506	0.5	0.519	0.513
12	0.506	0.5	0.522	0.516
13	0.505	0.499	0.509	0.503
14	0.505	0.499	0.507	0.501

Table 25: Data from Deletion Check for Fake News dataset

k	selected		random	
	acc	flip	acc	flip
0	0.884	1.0	0.884	1.0
10	0.819	0.885	0.882	0.97
20	0.777	0.831	0.877	0.955
30	0.724	0.764	0.861	0.919
40	0.704	0.726	0.841	0.899
50	0.662	0.684	0.792	0.85
60	0.636	0.656	0.785	0.835
70	0.623	0.637	0.748	0.782
80	0.607	0.615	0.715	0.747
90	0.594	0.6	0.699	0.715
100	0.579	0.581	0.682	0.694
110	0.566	0.572	0.658	0.674
120	0.57	0.57	0.633	0.645
130	0.555	0.557	0.618	0.63
140	0.547	0.555	0.588	0.606

Table 26: Data from Deletion Check for IMDB dataset

k	selected		random	
	acc	flip	acc	flip
0	0.515	0.47	0.515	0.47
1	0.515	0.47	0.515	0.47
2	0.52	0.475	0.515	0.472
3	0.522	0.481	0.517	0.472
4	0.519	0.486	0.517	0.472
5	0.517	0.488	0.522	0.481
6	0.534	0.509	0.52	0.489
7	0.53	0.521	0.542	0.493
8	0.533	0.54	0.533	0.506
9	0.546	0.551	0.543	0.538
10	0.548	0.585	0.552	0.541
11	0.56	0.621	0.544	0.585
12	0.563	0.642	0.546	0.593
13	0.557	0.66	0.56	0.627
14	0.564	0.689	0.553	0.648

Table 27: Data from Preservation Check for LIAR dataset

k	selected		random	
	acc	flip	acc	flip
0	0.505	0.499	0.505	0.499
1	0.505	0.499	0.505	0.499
2	0.505	0.499	0.505	0.499
3	0.505	0.499	0.505	0.499
4	0.506	0.5	0.505	0.499
5	0.512	0.506	0.505	0.499
6	0.526	0.52	0.506	0.5
7	0.542	0.536	0.516	0.51
8	0.562	0.556	0.514	0.508
9	0.583	0.577	0.552	0.546
10	0.599	0.595	0.575	0.569
11	0.642	0.638	0.606	0.602
12	0.676	0.674	0.643	0.645
13	0.722	0.724	0.678	0.678
14	0.777	0.785	0.733	0.735

Table 28: Data from Preservation Check for Fake News dataset

k	selected		random	
	acc	flip	acc	flip
0	0.512	0.506	0.512	0.506
10	0.692	0.702	0.512	0.506
20	0.752	0.768	0.523	0.517
30	0.773	0.797	0.532	0.526
40	0.794	0.814	0.546	0.54
50	0.799	0.821	0.569	0.567
60	0.802	0.832	0.586	0.586
70	0.809	0.837	0.603	0.601
80	0.818	0.852	0.627	0.625
90	0.829	0.869	0.665	0.665
100	0.835	0.871	0.675	0.683
110	0.838	0.876	0.7	0.714
120	0.841	0.877	0.725	0.729
130	0.842	0.886	0.733	0.743
140	0.852	0.896	0.753	0.771

Table 29: Data from Preservation Check for IMDB dataset

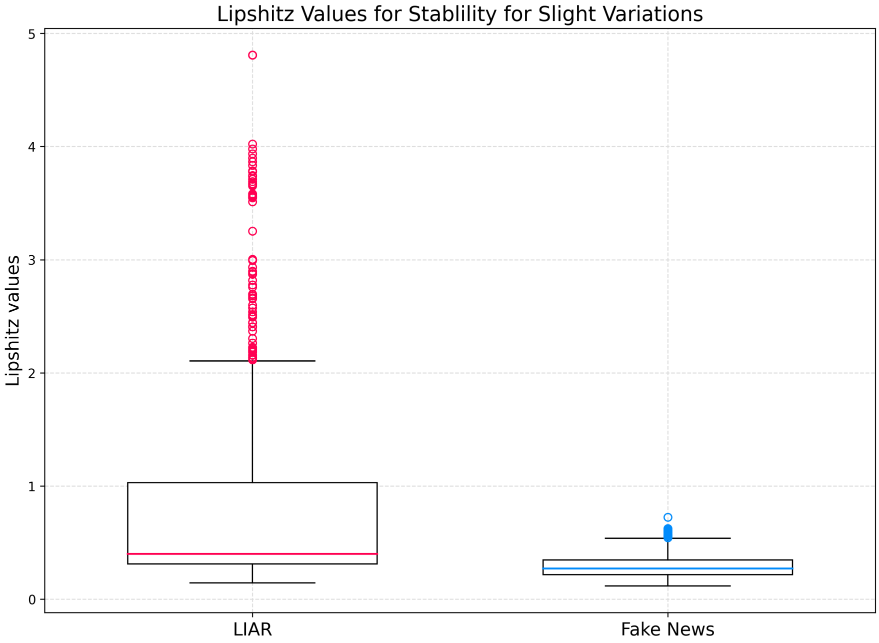


Figure 20: Local Lipschitz Values for the LIAR and Fake News Dataset