EPFL

Thèse n° 10 192

Estimating and Improving the Robustness of Attributions in Text

Présentée le 30 mai 2023

Faculté des sciences et techniques de l'ingénieur
Laboratoire de traitement des signaux 4
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Ádám Dániel IVÁNKAY

Acceptée sur proposition du jury

Prof. F. G. Krzakala, président du jury
Prof. P. Frossard, directeur de thèse
Prof. S. Mohsen Moosavi-Dezfooli, rapporteur
Dr P.-Y. Chen, rapporteur
Dr J. Henderson, rapporteur

École polytechnique fédérale de Lausanne

2023

"There's no system foolproof enough to defeat
a sufficiently great fool."
— Edward Teller

To my beloved family...

Acknowledgements

First of all, I would like to thank my supervisor Prof. Pascal Frossard for his unconditional support throughout my thesis. I am especially grateful for his incredible guidance and patience in the past years, which helped me master even the hardest moments.
I would also like to express my gratitude towards the jury members Prof. Florent Krzakala, Prof. Seyed Mohsen Moosavi-Dezfooli, Dr. Pin-Yu Chen and Dr. James Henderson for the feedback and fruitful discussions.
Moreover, I am grateful for IBM Research Zurich for giving me the opportunity to pursue this PhD and providing me both with the necessary professional and financial support. Here, I would like to express my special appreciation in particular towards Dr. Douglas Dykeman, Dr. Chiara Marchiori, Dr. Ivan Girardi and Dr. Thomas Brunschwiler, who gave invaluable support and believed in me even in times where I doubted myself. I am not sure how much of this thesis would exists without them. Thank you!
A special thanks goes to the boi Kevin Thandiackal, Zoltan Arnold Nagy and Dr. Mattia Rigotti, who made coming to the office every day a real joy. I am glad to have met you and hope to stay friends with you for a long time. Moreover, a very big thank you to my current and former labmates at LTS4, Apostolos, Guillermo, Javier, Hermina, Ahmet, Sahar and all the others who made my visits to Lausanne a true pleasure.
I am sincerely grateful to Henry, Bálint, Lucas, Lorenz, Michael, Nelly, Tamás, Danuta and all my friends close to and far from home, for always being there, in good times and in bad times. I am glad I can call you friends. I would also like to wholeheartedly thank Domino for her unconditional support and understanding during these last months.
Finally, I am forever indebted to my family for their unimaginable love and support. I am grateful for all their sacrifices during the years, for never hesitating to offer help and seeing the best in me when I was at my worst. Thanks for making good times even better and bad times more bearable. I know I can never give back all that you have given me.

Lausanne, May 3, 2023

Abstract

End-to-end learning methods like deep neural networks have been the driving force in the remarkable progress of machine learning in recent years. However, despite their success, the deployment of such networks in safety-critical use cases, such as healthcare, has been lagging. This is due to the *black-box* nature of deep neural networks. Such networks rely on raw data as input and learn relevant features directly from the data, which makes understanding the inference process hard. To mitigate this, several explanation methods have been proposed, such as local linear proxy models, attribution maps, feature activation maps or attention mechanisms.

However, many of these explanation methods, attribution maps in particular, tend not to fulfill certain desiderata of faithful explanations, in particular robustness, i.e., explanations should be invariant towards imperceptible perturbations in the input that do not alter the inference outcome. The poor robustness of attribution maps to such input alterations is a key factor that hinders trust in explanations and the deployment of neural networks in high-stakes scenarios. While the robustness of attribution maps has been studied extensively in the image domain, it has not been researched in text domains at all. This is the focus of this thesis. First, we show that the existence of imperceptible, adversarial perturbations on attributions extends to text classifiers as well. We demonstrate this on five text classification datasets and a range of state-of-the-art classifier architectures. Moreover, we show that such perturbations transfer across model architectures and attribution methods, being effective in scenarios where the target model and explanation method are unknown.

Our initial findings demonstrate the need for a definition of attribution robustness that incorporates the extent to which the input sentences are altered, in order to differentiate between more perceptible adversarial perturbations. Thus, we establish a new definition of attribution robustness that reflects the perceptibility of such alterations. This allows for effectively quantifying and comparing the robustness of neural network attributions. As part of this effort, we propose a set of metrics that effectively capture the perceptibility of perturbations in text. Then, based on our new definition, we introduce a novel attack that yields perturbations that alter explanations to a greater extent while being less perceptible.

Lastly, in order to improve attribution robustness in text classifiers, we introduce a general framework for training robust classifiers, which is a generalized formulation of current robust training objectives. We propose instantiations of this framework and show, with experiments on three biomedical text datasets, that attributions in medical text classifiers lack robustness to small input perturbations as well. Then, we showcase that our instantiations successfully

train networks with improved robustness of attributions, outperforming baseline methods. Finally, we show that our framework performs better or comparably to current methods in image classification as well, while being more general.
In summary, our work significantly contributes to quantifying and improving the attribution robustness of text classifiers, taking a step towards enabling the safe deployment of state-of-the-art neural networks in real-life, safety-critical applications like healthcare ones.

Keywords: robustness, adversarial robustness, explainability, explainable deep learning, attribution maps, attribution robustness, text classification, robust attributions

Résumé

Les réseaux neuronaux profonds ont été le moteur des progrès remarquables de l'apprentissage automatique. Toutefois, le déploiement de ces réseaux dans les cas d'utilisation critique pour la sécurité a pris du retard. Cela est dû à la nature *boîte noire* des réseaux neuronaux profonds. Ils s'appuient sur des données brutes et apprennent des caractéristiques directement à partir des données, ce qui rend difficile la compréhension du processus d'inférence. Pour atténuer ce problème, plusieurs méthodes d'explication ont été proposées, telles que les modèles de substitution, les cartes d'attribution ou les mécanismes d'attention. Toutefois, bon nombre de ces méthodes, en particulier les cartes d'attribution, tendent à ne pas répondre à certains critères d'une explication précises, en particulier la robustesse, c'est-à-dire que les explications doivent être invariantes par rapport à des perturbations imperceptibles de l'entrée qui ne modifient pas la prédiction. La faible robustesse des cartes d'attribution est un facteur clé qui entrave la confiance dans les explications et le déploiement des réseaux neuronaux dans des scénarios critiques. Alors que la robustesse des attributions a été bien étudiée dans le domaine de l'image, elle n'a pas du tout été étudiée dans le cas du texte. C'est l'objet de cette thèse. Tout d'abord, nous montrons que l'existence de perturbations imperceptibles sur les attributions s'étend aux classificateurs de texte. Nous le démontrons sur cinq ensembles de données de classification de textes et sur plusieurs classificateurs de pointe. En outre, nous montrons que ces perturbations sont transférables entre les architectures de modèles et les méthodes d'attribution, et qu'elles sont efficaces dans les scénarios où le modèle cible et l'explication sont inconnus. Nos premiers résultats démontrent la nécessité d'une définition de la robustesse d'attribution qui intègre l'étendue des altérations d'entrée afin de différencier les perturbations plus perceptibles. Nous établissons donc une nouvelle définition de la robustesse de l'attribution qui reflète cela, permettant de quantifier et de comparer efficacement la robustesse des attributions. Dans ce cadre, nous proposons un ensemble de mesures qui rendent compte de la perceptibilité des perturbations dans le texte. Nous présentons ensuite une nouvelle attaque qui produit des perturbations altérant davantage les explications tout en étant moins perceptibles. Enfin, nous introduisons un framework général pour la formation de classificateurs robustes qui améliore la robustesse des attributions dans le texte. Nous proposons des instanciations de ce framework et montrons, sur trois ensembles de données biomédicales, que les attributions dans les classificateurs de textes médicaux manquent de robustesse face aux petites perturbations. Nous montrons ensuite que nos instanciations entraînent des réseaux avec une robustesse d'attribution améliorée, surpassant les méthodes de référence. Enfin, nous montrons que notre framework est plus performant ou

comparable aux méthodes actuelles de classification d'images, tout en étant plus général. En résumé, notre travail contribue de manière significative à la quantification et à l'amélioration de la robustesse d'attribution des classificateurs de texte, ce qui constitue une étape vers le déploiement sûr de réseaux neuronaux de pointe dans de réelles applications critiques pour la sécurité.

Mots clés : robustesse, robustesse contradictoire, explicabilité, apprentissage profond explicable, cartes d'attribution, robustesse de l'attribution, classification de textes, attributions robustes

Contents

1 Introduction

"To know what you know and what you do not know,
that is true knowledge."
— Confucius

The recent extraordinary success of deep neural networks (DNNs) has revolutionized the field of machine learning. Fueled by the ample amount of available data, deep end-to-end learning algorithms have set new standards in various artificial intelligence tasks like image and text classification, reinforcement learning, language modeling or machine translation. These networks rely on a large amount of raw data and optimize their parametrized architecture to efficiently extract the features relevant for the target task, without the need for domain specific knowledge.

This rapid success does not come without hurdles however. Due to these networks having a large amount of parameters combined in a highly non-linear fashion, the inference process becomes complex and almost impossible for humans to properly comprehend. In certain high stakes scenarios, such as autonomous driving or healthcare, this lack of understanding of the inference process results in the lack of trust in the models themselves. It prevents the safe real-life deployment of such powerful neural networks. Naturally, if human lives are at stake, car manufacturers or healthcare professionals do not trust systems whose decision process is not fully understood, as unexpected behaviours and faults can have severe consequences.

In order to mitigate this ever-growing need for transparency in DNNs, the field of explainable artificial intelligence (XAI) has seen a surge in recent years. Numerous methods have been introduced, both in text and image domains, to yield insights into several aspects of the inference process in DNNs. However, the explanations provided by these methods are often

subpar and inconsistent with the human reasoning of the given problem. For instance, they might highlight background as being important to recognize a foreground object in an image, or may not give a critical symptom in an EHR for an automated diagnosis any relevance. Even though the human reasoning process behind such problems might also be ambiguous, such behaviours are still highly unexpected and tend to strengthen the users reluctance towards using DNNs.

A particularly intricate aspect of explanation methods is their robustness to input alterations and potential outliers. When deploying the DNNs in real-life scenarios, the environment and nature of the input data can be highly diverse and noisy, potentially very different from what the networks were trained on. Therefore, in order to still deliver meaningful and sound insights into the inference process, it is crucial for explanation methods to be invariant towards these altered circumstances. However, this is not the case for most explanation methods [1], especially attribution maps. The outcome of such methods has been shown to be brittle and can be significantly altered in the presence of imperceptible, potentially targeted noise that does not affect the true causal chain. This serious defect is still far from fully understood and has only been researched in image classification so far. In fact, understanding the behaviour of explanations under worst-case, potentially adversarial scenarios remains largely incomplete, especially in textual domains. In this thesis, we aim to bridge this gap by proposing novel methods to study and quantify the robustness of DNN explanations to imperceptible input alterations in discrete input text domain, and propose methods to train DNNs that yield robust explanations.

1.1 Explainable AI, Plausibility and Faithfulness

Deep neural networks are often evaluated solely on their performance on a specific task. The higher the networks' accuracy on a given text classification task, the higher the BLEU score [2] of a translation given by a transformer [3] architecture or the lower the perplexity [4] of a language model on a dataset, the better the network is considered. Nevertheless, when deploying these networks in real-life scenarios, other aspects like inference runtime, hardware constraints as well as transparency, fairness and robustness of the networks become crucial. Even more in safety-critical scenarios like healthcare [5] or financial decision making [6], trust in the networks' prediction being robust and correct in unforeseen circumstances is arguably as important as their accuracy [7].

Unfortunately, the complex nature and large number of parameters of high-performing neural networks make it very hard for humans to understand the decision process. Experts can hardly answer the question why a network came to a certain decision by merely observing the network. Thus, numerous methods to provide transparency in the processes governing the DNN decisions have been proposed. Several local and global, pre- and post-hoc explanation algorithms aim to give insights into several aspects of neural networks. Amongst others, attribution maps [8, 9], local proxy models [10], input occlusion [11] provide explanations by highlighting

the most important features in the input that contribute to the decision process. Feature visualization maps [12] provide knowledge of the networks by extracting input patterns for specific layers and neurons that maximize their activations, aiming to extract a representation of the features learnt by DNNs. Intrinsic explainers or explanation producing architectures [13, 14] are specifically designed to output predictions and corresponding explanations at the same time, often claimed to be *inherently explainable*.

Even though most of these methods are principled approaches to explaining DNNs, it still remains unclear what the appropriate explanations for a given task are, as even human explanations are highly domain-specific. Moreover, it is rarely the case that only one explanation is correct, most interpretations of real-life phenomena are ambiguous. In addition, it is not guaranteed that these post-hoc methods reflect the true behaviour of DNNs to any extent. In fact, it has been shown that many explanation methods are independent of the underlying neural network outcomes [15], rendering them less useful to debug the network process. This reflects the difficult problem of defining *sound* explanations, as there is no unified view on what good explanations are, even for humans [16]. Therefore, a large body of research has focused on defining *desiderata* of explanations.

These desiderata capture many aspects of explanations, such as informativeness, fairness, fidelity or reliability, and their taxonomy in the literature is far from unified [17]. However, we follow current efforts to unify the notions describing evaluation criteria of explanations [18, 19], stating that the two main desiderata, comprising most properties of explanations, are the following.

Desideratum 1: Plausibility. Plausibility refers to the task specific measure of the extent to which the provided explanation is aligned with the human understanding and reasoning about the model and the task [20]. This encompasses aspects like interpretability (the explanation is presented in a form that is understandable to the user) or clarity (the explanations can be interpreted only in one way - unambiguity), both of which are essential for users to be able to actually understand the outcome of the explanation algorithm and for it to be convincing.

Desideratum 2: Faithfulness. Faithfulness [19] describes the accuracy of the explanation reflecting the true causal process of the neural network. The explanations are only useful if they truly describe how the network processes information and derives its output. Reasoning about explanations methods being faithful is hard and there is no uniformly agreed on way to measure faithfulness. However, the authors of [19] describe three assumptions that comprise most formulations of faithfulness. The first is the *model assumption*, stating that interpretation methods are faithful if and only if they provide equal explanations for various models that give the same outputs. This is connected to the completeness property by assuming that the explanations provide feedback on the complete reasoning process. The second assumption is *linearity*, formulating that certain parts in the input are more important than others, and that these parts are independent from each other. This is mainly assumed in heat map-based explanation methods like attributions [19]. The last assumption, the *prediction assumption*

is rooted in the robustness of explanations. It states that on similar inputs, interpretation algorithms should yield similar explanations if the predictions of the networks on those similar inputs are equal.

Overall, plausibility depends on the human understanding of the problem and the task at hand, thus is domain-specific by nature. It describes the interaction between user and explanation. Faithfulness, on the other hand, refers to the connection between the interpretation method and the model, which can be addressed in a domain-agnostic manner, assuming both the model and the explanation method are general. Even though these two desiderata are orthogonal to each other, they need to be assessed jointly. Naturally, it is possible for explanations to be faithful but not plausible, or plausible but not faithful. Yet, according to [18], plausible but unfaithful explanations (*convincing lies*) are particularly pernicious worst-case scenarios, as they are harder to spot than implausible unfaithful explanations (*unconvincing lies*), and can have severe consequences. Therefore, ensuring DNNs fulfill all aspects of faithful explanations, e.g., robustness, or being able to detect if they do not, for instance through the lens of plausibility, is crucial.

1.2 Robust Attribution Maps

An important building block of faithfulness is the robustness (or prediction) assumption defined in [19]. Explanations are expected to be close to identical for similar inputs if the network's prediction is identical on those inputs. This is especially important when deploying DNNs in safety-critical real world scenarios, where data is expected to be considerably diverse and noisy. Explanations that are resistant to imperceptible alterations of the data are crucial in such cases. For instance, a medical professional assessing electronic health records would neither understand nor trust a model that yields two significantly different attribution maps for seemingly identical input texts and predictions. Unfortunately, deep neural network models as well as current attribution methods severely lack robustness in conditions that are different from the training ones [1, 21].

When deploying networks with accompanying explanations in real life, this brittleness of attributions can be observed in multiple ways. While attributions tend to be robust against random noise in the input [22], semantic-preserving constant shifts have been shown to largely affect interpretations, even though the underlying neural network is invariant to such shifts [23]. Moreover, interpretation methods can be sensitive to common hyperparameter choices, such as baseline inputs, blur sizes or even random seeds [24]. On the other hand, certain interpretation methods lack sensitivity to model parameter and label randomization, shown by a randomization tests in [15], thus are necessarily not faithful. These are highly problematic shortcomings, as these aspects are hardly ever considered in research, let alone when neural networks are deployed in real life.

A more nuanced notion of the robustness assumption of faithful explanations is adversarial robustness. While constant shifts or hyperparameter changes are easily detectable by the

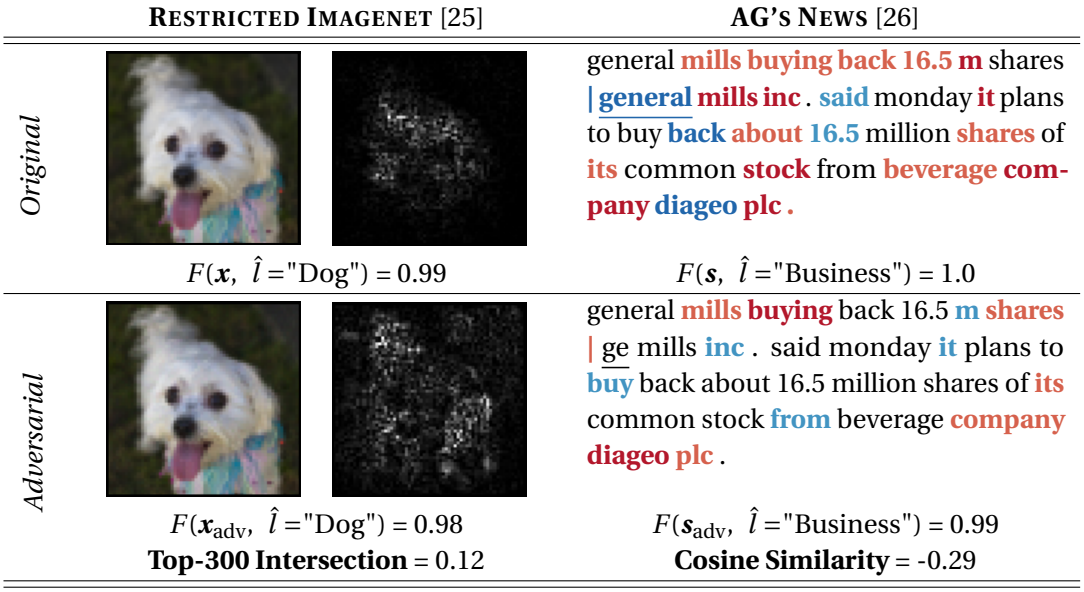


Figure 1.1: Attribution maps are susceptible to carefully crafted, imperceptible input perturbations in image and text classifiers. By modifying certain pixels (left) or substituting words (right) in the inputs, the heat map explanations of original (top) and adversarial (bottom) inputs are significantly altered, while the confidence of the correctly predicted classes ($\hat{l}$) are unchanged (denoted by F). The Top-300 intersection between the most important pixels in original and adversarial explanations drops to 0.12 and the cosine similarity between the word-wise attributions in the sentences drops to -0.29 after adversarially perturbing the inputs, even though the alterations to the inputs are close to imperceptible to the users.

user or professional interacting with the DNNs, it is crucial to address situations where the environment alterations are imperceptible and can hardly be detected. In fact, it has been shown that explanations of DNNs [1] are susceptible to adversarial input perturbations. These are carefully crafted, worst-case input alterations that are imperceptible to the human eye, but alter the explanations significantly, to an extent that they do not resemble the attributions of the original samples at all, while keeping the outcome prediction of the network unchanged. An example of the brittleness of such explanations can be observed in Figure 1.1.

The existence of such perturbations has first been shown in the image domain for image classifiers. The original formulation [1] for computing these imperceptible perturbations was defined as the following maximization problem:

$$\begin{aligned} \underset{\tilde{\boldsymbol{x}}}{\arg\max}\ & d\big[A(\tilde{\boldsymbol{x}}, F, \hat{l}), A(\boldsymbol{x}, F, \hat{l})\big] \\ \text{s.t. }\ & ||\tilde{\boldsymbol{x}} - \boldsymbol{x}||_{\text{p}} < \varepsilon \\ & \underset{i}{\arg\max}\ F_i(\tilde{\boldsymbol{x}}) = \underset{i}{\arg\max}\, F_i(\boldsymbol{x}) \end{aligned} \tag{1.1}$$

Here, $\boldsymbol{x}$ denotes a correctly classified, $\hat{l}$-labelled $\mathbb{R}^n$-dimensional input of the classifier $F : \mathbb{R}^n \to \mathbb{R}^{|\mathbb{L}|}$ with components F_i indicating the logits for class i and A denoting the attribution method. The parameter ε is the radius of the ℓ_{p}-ball around $\boldsymbol{x}$. Intuitively, this optimization problem solves the search for an input alteration that maximizes attribution change within

a small neighbourhood of the original input, while keeping predictions of the network unchanged. Such a formulation can directly be translated into the violation of the prediction assumption of faithfulness, as two very similar inputs yield significantly different interpretations for the same output class. We note that the constraints on the prediction of the output are crucial, as without it, the prediction assumption would not necessarily be violated. The imperceptibility of perturbations, i.e., the search in a small neighbourhood only, asserts that the inputs remain semantically equal, thus explanations might remain plausible. This opens up the possibility of convincing lies, i.e., plausible but unfaithful interpretations.

Therefore, understanding and mitigating this worst-case adversarial *fragility* of interpretations is critical. In the image domain, the lack of robustness has been connected to the high curvature of the decision boundary in neural network classifiers, drawing parallels between gradient-based attributions, traditional adversarial samples and adversarial training [27]. This is not surprising, as the exploration of adversarial explanations is tightly bound to the prediction loss in the local neighbourhood of the input [28]. Utilizing this connection, several methods have been proposed to train networks that yield more robust attributions than those of vanilla neural network image classifiers [29, 30].

Nevertheless, this phenomenon has not been studied in the textual domain nearly as widely as in the image domain. In fact, until recently [31, 32], it has not been proven that such perturbations exist for text classifiers as well. One reason is that adversarial search in a discrete input space is a harder problem than for the continuous input image space, as not every point in the input vector space is a valid input. Therefore, gradient-based search methods proposed in current literature can not be straightforwardly used with text data. Moreover, the widely used ℓ_p-norm-based input distance metrics are only limitedly useful for text, as the embedding space can behave very differently in terms of semantics than the discrete word space.

In this thesis, we aim to bridge the gap between images and text by providing an extensive study of the robustness of faithful explanations in textual modalities and propose solutions to make systems more robust. The next section provides an overview of the thesis and contains the main contributions.

1.3 Thesis Outline

The goal of this thesis is to estimate and improve the robustness of attribution maps as interpretation methods in deep neural networks for text classification. We organize the thesis as follows.

Chapter 2 offers an introduction into the definitions and common notations used in this thesis and gives an extensive review on related work in attribution robustness.

In Chapter 3, we analyze the robustness of attribution methods in the text domain. In particu-

lar, we are interested in the existence of adversarial perturbations that alter interpretations significantly while leaving the prediction unchanged. To this end, we introduce a novel black-box attack algorithm that aims to estimate the local robustness of common attribution methods in text classifiers. We find that the phenomenon of brittle interpretations poses an issue in text classification as it does for image applications. In our experiments we do not only show that imperceptible word substitution attacks exist for text attributions, but also that these transfer across both different models and attribution methods, posing a threat, even in scenarios where the exact model or attribution algorithm is unknown.

Motivated by the observations of Chapter 3, we further explore the brittleness of text attributions in Chapter 4. We give a rigorous mathematical formulation to effectively quantify the robustness of DNN explanation methods and allow accurate comparison of methods and models. This definition incorporates both the attribution change and the input distance between perturbed and unperturbed inputs into the formulation. To this end, we provide a set of metrics to effectively capture input distances in the text domain instead of the commonly used ℓ_p-distances for images, which are only limitedly useful for textual inputs. This further allows us to evaluate robustness in the light of plausibility. Utilizing our new definition of robustness, we improve state-of-the-art attribution robustness estimation by introducing a novel algorithm that computes perturbations which alter attributions more while resulting in more fluent perturbed text.

Then, in Chapter 5, we introduce a novel framework, FAR, to train networks with robust attributions. Our extensive experiments, on three biomedical datasets as a case study, show that training networks with this framework results in significantly higher attribution robustness than current baseline methods in text classifiers. We then show that FAR generalizes current robust training objectives in image classification and achieves better or comparable results to those in terms of AR, while having fewer assumptions.

Finally, Chapter 6 concludes the work and offers potential direction for future research in the field of robust interpretations.

1.4 Contributions

We summarize our main contributions as follows:

- We provide evidence that the of existence of adversarial perturbations targeting attribution maps extends to the textual domain, and does not merely affect image data applications.
- We develop new insights into the transfer capabilities of such perturbations to different model architectures and attribution methods.
- We propose a novel definition of attribution robustness in text, which captures both attribution change and input distance between perturbed and unperturbed text samples,

which permits to effectively quantify robustness of attributions in text.

- We present a generalized framework with new optimization objectives that allows to train DNNs with robust attribution maps both in the image and text modalities.
- We offer a first case study on attribution robustness in the biomedical domain, on three text classification datasets, showing how attribution robustness estimation can be adapted to different application domains and how robust networks can be built in critical healthcare applications.

2 Background

In this thesis, we examine the adversarial robustness of attribution maps in text classifiers. To introduce the topic, we first shortly describe the common notation of the chapters. Then, we describe existing methods to explain DNN decisions together with their desired properties and define the methods we use throughout this thesis. We then describe related work for robustness estimation, as well as the connection of robustness estimation of attributions to traditional adversarial robustness.

2.1 Notation

We first define the basic notation used in this thesis. First, let $\mathbb{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_N\}$ be a dataset of N text samples $\boldsymbol{s}_i$, each associated with a label from a predefined set of labels $\mathbb{L} = \{l_1, l_2, ..., l_{|\mathbb{L}|}\}$. Each sample $\boldsymbol{s}$ contains a sequence of tokens (or words) w_i taken from a discrete vocabulary set $\mathbb{W} = \{w_1, w_2, ..., w_{|\mathbb{W}|}\}$. A generic text classifier is then defined as a function $F : \mathbb{S} \to \mathbb{R}^{|\mathbb{L}|},\ F(\boldsymbol{s}) = (f \circ E)(\boldsymbol{s}) = \boldsymbol{o}$ mapping each sentence in $\mathbb{S}$ to the class logits $\boldsymbol{o} \in \mathbb{R}^{|\mathbb{L}|}$. Typically, a text classifier is a composition of an embedding function $E : \mathbb{S} \to \mathbb{R}^{d \times p},\ E(\boldsymbol{s}) = \boldsymbol{X}$, mapping each input sequence to a real valued matrix $\mathbb{R}^{d \times p}$, and a classifier function $f : \mathbb{R}^{d \times p} \to \mathbb{R}^{|\mathbb{L}|},\ f(\boldsymbol{X}) = \boldsymbol{o}$ mapping the embeddings to the output logits $\boldsymbol{o}$. Here, d denotes the embedding dimension and p the maximum sequence length. In general, the embedding function E is a non-differentiable lookup table mapping the tokens w_i in the input sentence $\boldsymbol{s}$ to a continuous vector in $\mathbb{R}^d$, while the classifier f can be any differentiable function, for instance a deep neural network. We denote $\hat{l} = \arg\max_{i \in \{1:|\mathbb{L}|\}} f_i(\boldsymbol{X}) = \arg\max_{i \in \{1:|\mathbb{L}|\}} F_i(\boldsymbol{s}) = \arg\max_{i \in \{1:|\mathbb{L}|\}} \boldsymbol{o}_i$ to be the maximum of the output logits, thus the DNNs predicted class of sample $\boldsymbol{s}$ – l denotes the true class.

The *perplexity* [4] of a text sample $\boldsymbol{s}$ with tokens w_i given a language model L is a function $PP : \mathbb{S} \to \mathbb{R}^+,\ PP(\boldsymbol{s}|L) = \boldsymbol{p}$ that measures how well the probability distribution on the language given by L predicts sample $\boldsymbol{s}$, as defined in Equation (2.1):

$$PP(\boldsymbol{s}|L) = 2^{-\sum_{w_i \in \boldsymbol{s}} p(w_i|L,\boldsymbol{s}) \log p(w_i|L,\boldsymbol{s})} \tag{2.1}$$

where PP denotes the perplexity of the text sample $\boldsymbol{s}$ and $p(w_i|L,\boldsymbol{s})$ the probability of token w_i given L and $\boldsymbol{s}$. Low perplexity values indicate that the model L has captured the true distribution of the text dataset $\mathbb{S}$ well.

Sentence encoders are embedding functions $E_s : \mathbb{S} \rightarrow \mathbb{R}^m,\ E_s(\boldsymbol{s}) = \boldsymbol{m}$ that assign a continuous embedding vector of dimension m to each text sample [33]. These embeddings are used to capture higher-level representations of the sentences, which can be used to train downstream tasks effectively. As they are jointly trained on a diverse set of multi-task problems, they are argued to capture the overall semantic meaning of the text well [33].

2.2 Methods of Explaining DNNs

The recent need for unraveling the black-box nature of DNNs has led to vibrant research in methods that provide insight into the inference process. Several surveys [34, 35, 36] focus on summarizing recent progress on these methods, we therefore refrain from giving an in-detail review of them, but shortly introduce the main concepts and taxonomies.

Interpretation methods are categorized based on their scope, methodology and usage [5]. Local scoped methods like attribution maps [37, 38] or local proxy models [10] provide explanations on individual data instances, highlighting important input features or characterizing inner model parameters for that specific instance. In contrast, global methods aim to provide insight into how DNNs work in general, what kind of features they learn or the patterns in the data they are especially sensitive to. Amongst others, layer visualization maps [12], Concept Activation Maps [39] or Automatic Concept-based Explanations [40] are examples of global methods. Based on their methodology, interpretation methods can be divided into three categories. Backpropagation-based methods [9, 41, 8, 38, 42, 37] compute the gradient of the last layer activations with respect to the input and aim to extract relevant input features using these gradients. Perturbation-based models [43, 11, 10, 44, 45, 46, 47] directly modify the inputs by occlusion, substitution or conditional sampling and observe how the model behaves in presence of such modifications. The third category encompasses all methods that are not perturbation or gradient-based, such as Anchors [47], Deep Taylor Expansion [48] or Automatic Concept-based Explanations [40]. Lastly, the usage of explanations can be two-fold. Intrinsic explanations [14, 13, 49, 50] embed the explanations into the architecture and extract them during the forward pass through the network. In contrast, post-hoc methods [11, 44, 43] compute their explanations after the inference process and treat the DNNs separately from their explanations. Post-hoc methods are extremely useful as they can be added to an already trained network, without requiring modifications in the network structure or retraining. Moreover, most post-hoc methods work in a model-agnostic fashion and can be used with almost any neural network architecture.

In this work, we focus on attribution maps [9, 51]. Attribution maps are local explanations that provide explanations in form of post-hoc heat maps. They assign a value to each input dimension that represents its importance and influence towards the prediction outcome of

DNNs. Positive values are associated with features contributing *towards* the class prediction, negative ones with features *against* it. These maps are popular because they provide explanations with no need for specific domain knowledge and can be applied to most architectures in a post-hoc fashion.

Mathematically, we define an attribution map as a function $A : \mathbb{S} \rightarrow \mathbb{R}^{\dim(\boldsymbol{X})}$, $A(\boldsymbol{X}, f, \hat{l}) = \boldsymbol{a}$ that assigns a scalar value to each element of $\boldsymbol{X}$ in sample $\boldsymbol{s}$, resulting in the attribution matrix $\boldsymbol{a} \in \mathbb{R}^{\dim(\boldsymbol{X})}$. This matrix represents each input elements influence towards the prediction outcome $\hat{l}$ of classifier f. This thesis considers four widely-used attribution methods in classification tasks, Saliency Maps (S) [9], DeepLIFT (DL) [41], Integrated Gradients (IG) [8] and the Self-Attention (A) [13], defined in the following Equations (2.2), (2.3), (2.4) and (2.5) respectively.

Saliency Maps [9] are defined as follows:

$$A^{\mathrm{S}}(\boldsymbol{s}, F, \hat{l}) = A^{\mathrm{S}}(\boldsymbol{X}, f, \hat{l}) = |\nabla_{\boldsymbol{X}} f_{\hat{l}}(\boldsymbol{X})| \tag{2.2}$$

where $\nabla_{\boldsymbol{X}}$ is the gradient of logit $f_{\hat{l}}$ with respect to $\boldsymbol{X}$.

DeepLIFT attributions [41] are computed with the following equation:

$$A^{\mathrm{DL}}(\boldsymbol{s}, F, \hat{l}) = A^{\mathrm{DL}}(\boldsymbol{X}, f, \hat{l}) = \sum_j \frac{\boldsymbol{z}_j - \overline{\boldsymbol{z}_j}}{\sum_i \boldsymbol{z}_{j,i} - \sum_i \overline{\boldsymbol{z}_{j,i}}} \cdot r_j^1 \tag{2.3}$$

Here, $\boldsymbol{z}$ denote the weighted activations, $\overline{\boldsymbol{z}}$ the weighted baseline activations and r the propagated relevance scores. The equation contains the relevance scores computed in vector form for the first layer, which are propagated through the network beginning from the last layer's logits. We refer to the original paper for a detailed description [41].

Our third attribution, *Integrated Gradients* [8] is computed utilizing Equation (2.4).

$$A^{\mathrm{IG}}(\boldsymbol{s}, F, \hat{l}) = A^{\mathrm{IG}}(\boldsymbol{X}, f, \hat{l}, \boldsymbol{B}) = (\boldsymbol{X} - \boldsymbol{B}) \cdot \int_{\alpha=0}^{1} \nabla_{\tilde{\boldsymbol{X}}} f_{\hat{l}}(\tilde{\boldsymbol{X}})|_{\tilde{\boldsymbol{X}} = \boldsymbol{B} + \alpha(\boldsymbol{X} - \boldsymbol{B})} \, d\alpha \tag{2.4}$$

where $\boldsymbol{B}$ denotes the null matrix $\mathbf{0}^{\dim(\boldsymbol{X})}$ and f is the classifier function. If not stated otherwise, we use the null-matrix as the baseline input for both DeepLIFT and Integrated Gradients.

Finally, *Self-Attention* [13] is defined as follows:

$$A^{\mathrm{A}}(\boldsymbol{s}, F, \hat{l}) = A^{\mathrm{A}}(\boldsymbol{X}, f, \hat{l}) = \frac{e^{\alpha_{[\mathrm{CLS}],-1}}}{\sum_{j \in \{1:p\}} e^{\alpha_{j,-1}}} \tag{2.5}$$

where α denotes the attention scores computed during forward propagation of the attention layer. We refer to the original paper for a detailed description of the attention mechanism [13]. For our attributions, we take the softmaxed attention weights of the `[CLS]` (or equivalent) token in the last attention layer of our transformer architectures, if not stated otherwise.

We apply basic post-processing to the attributions maps. We scale the values to be in the interval $[-1,1]$ by dividing them with the maximum of $|\mathbf{a}|$. Moreover, for text attributions, A contains values for each embedding dimension for each token in input sample $\boldsymbol{s}$, as $\dim(\boldsymbol{a}) = \dim(\boldsymbol{X})$. In order to extract per-token attributions, we sum the attribution values for each token along the embedding dimension, resulting in one value $A_i = \sum_j A_{j,i}(\boldsymbol{X}, f, \hat{l})$ per token, where i denotes the token index. If sentences are subword-tokenized, which is the case for most transformer architectures [3], we extract per-word attributions by summing up the attributions of the subword tokens.

2.3 Characterization of Robustness in Attribution Maps

2.3.1 Invariances and Sensitivity

The assessment of *faithfulness* [19] and in particular the robustness of attribution maps has been approached from many directions, though mainly in the image domain. The authors of [15] proposed *sanity checks* for many state-of-the-art attributions via model and data randomization tests. They found that some attribution methods are insensitive to these randomization tests, thus are unable to provide insight into tasks that depend on the data or model. According to the authors, these maps behave similarly to edge detectors. The work of [23] shows that some saliency methods lack reliability in the presence of constant input shifts, showing that they are significantly altered by the shifts, even though the underlying model and its predictions are unaffected. On a different note, the work of [24] concludes that attributions are sensitive to their hyperparameters — choice of baseline, noise radius or even the random seed.

In the textual domain, the current focus mainly lies on the attention mechanism. The authors of [52] show, by input reduction, that the iterative removal of least important words (deemed by the explanation methods) leads to inputs which are still predicted correctly, but make no sense and are highly implausible. In [53], it is shown that different attention values can be extracted for the same data and same predicted classes, thus contradicting the model assumption of faithful explanations. As stated in [54] however, this does not directly disprove the usefulness of attention as explanations, depending on how explanations are defined. The authors of [55] show that attention and gradient-based explanations might not correlate, and that altering attention values does not necessarily lead to changes in output predictions of transformers. Then, the utility of attention to determine which input features are most relevant is questioned in [56] and is found to be subpar to gradient-based saliency methods. The work of [57] investigates whether sparse attention is more interpretable, but finds it to yield even less correlation between its attention weights and influential inputs.

2.3.2 Adversarial Robustness of Attributions

One important aspect of DNNs and their explanations is how they behave in presence of small, imperceptible input alterations. Ideally, networks are invariant towards any input change that does not change the semantics of the input in any noticeable way. However, it has been shown that DNN predictions are highly susceptible to targeted input alterations — adversarial attacks, which change the output logits of the networks significantly [58]. This phenomenon is widely studied in the image domain [58, 21, 59, 60, 61, 62, 63] as well as the text domain [64, 65, 66, 67, 68, 69, 70]. Several methods have also been proposed to mitigate this sensitivity to adversarial attacks [62, 71, 27, 72] and make networks' predictions more robust.

The robustness estimation of attribution methods (attribution robustness – AR) is tightly connected to the aforementioned traditional adversarial robustness, but the goal of it is fundamentally different. While adversarial attacks aim to alter predictions with imperceptible input changes, AR examines the behaviour of explanations under such input alteration while keeping the predictions of the networks constant. The authors of [1] were the first to show that common explanations like DeepLIFT, Integrated Gradients or even train sample influence can significantly be altered with small input perturbations while keeping predictions intact. The works of [22] and [73] then show this phenomenon happening for other interpretation methods as well. In [74], small patches in the input images to a classifier are shown to alter heat map explanations significantly. On a different note, attributions have been found to be sensitive to perturbations of the model parameters [75]. These works highlight that most explanation methods do not fulfill the prediction assumption of faithful explanations.

The origin of the previously described fragility of attributions is far from well-understood. Most explanation methods investigated are based on the model gradients, propagated back from output logits to the input [9, 41, 8]. Therefore, it is hypothesized in [1] that the high curvature of the decision boundary in classifiers [27] causes high irregularity in gradients, which then affects the attributions. This is supported by the fact that the alignment between input gradients and inputs in robust image classifiers with low decision boundary curvature is found to be higher [28]. The alignment is the scalar product between input $\boldsymbol{X}$ and input gradient $\nabla_{\boldsymbol{X}} \max_i \boldsymbol{o}_i$ in images, written as $\langle \boldsymbol{X}; \nabla_{\boldsymbol{X}} \max_i \boldsymbol{o}_i \rangle$. Moreover, the authors of [76] establish a theoretical connection between the attribution robustness and the geometrical properties of the decision boundary, which is then further examined by [30]. The geometry of the boundary is closely tied to the network structure, which is then argued by [77] to be the main cause of fragile attributions.

These works highlight that attribution robustness needs to be investigated not only in perspective of the user but also of the model. As attribution methods in this thesis utilize direct model information, such as prediction loss gradients or attention values, their robustness will reflect both properties of the attribution methods as well as properties of the underlying classification model. In light of this, AR can be viewed as extension to prediction robustness, requiring stable explanations in addition to robust predictions. The aforementioned works

[28, 30, 76, 77], backed up by the prediction assumption of faithfulness, show that even when the predictions are robust and remain unchanged, the stability of explanations can still be affected by other model properties, such as the regularity of gradients.

The estimation of attribution robustness in DNNs has mostly been done in the image domain and seriously lags behind in the text input space. In fact, prior to this thesis, there has been no proof that such input perturbations exist for text classifiers as well. This thesis is the first, along with the parallel work of [32], to explore adversarial perturbations on attributions in text. Arguably, the text modality behaves differently than the continuous input image space. Attribution robustness estimation, in the previously described form, heavily relies on gradient optimization of the input space in a very small ε-ball around the original input $\boldsymbol{X}$, assuming every input within that ball bein valid. This is not the case for discrete inputs like text, finding imperceptible perturbations to discrete inputs is a much harder problem, which this thesis investigates thoroughly.

2.4 Improvement of Explanation Robustness

It is crucial for explanations to be robust to small input alterations in order to be faithful and build user trust in critical use cases [78, 79]. A medical professional for instance would neither believe nor trust a system that yields significantly different explanations for seemingly the same inputs and outputs. Therefore several methods have been introduced to train DNNs that have robust attributions in adversarial environments. It has been shown that traditional adversarial training, as defined in [62, 27], significantly enhances AR in image classification [29]. This is due to the fact that adversarial training regularizes the curvature of the decision boundary, thus making gradients more regular. This is in line with previous observations between gradient-input alignment and adversarially robust networks. The authors of [76] propose robust attributions by approximating the ReLU activations with SoftPlus functions, utilizing large β approximation values. In [30], smooth surface regulation is proposed to successfully increase the robustness of explanation methods. The alignment between input and gradients can also be directly increased by introducing a regularization term, similar to the one in [29], which penalizes the maximum misalignment of the two quantities. The work of [80] utilizes the axiomatic properties of Integrated Gradients attributions to minimize the worst-case attribution changes within an ε-bound of the input.

AR can also be enhanced by averaging the outcome of multiple attributions, as shown in [81]. This is closely related to another research direction, certified attribution robustness, that has recently been started to be explored. Contrary to previous approaches, these methods give (probabilistic or absolute) bounds on the robustness of explanations. The authors of [82] prove that a sparsified version of the SmoothGrad attribution [37], computing mean attributions over perturbed inputs, is certifiably robust against attacks. Moreover, certifiably robust attributions can also be achieved via Rényi differential privacy [83]. Recently, the work of [84] has extracted certified bounds on attribution changes within a small local neighbourhood of the input by

propagating symbolic intervals through the network, relying on relaxation techniques from non-convex optimization.

2.5 Summary

All of the aforementioned methods rely on the assumption that the adversarial inputs is close to the original one in its semantics, making sure the input perturbations do not alter the ground truth meaning of the data. An established proxy for this assumption in the image domain is the local ε-ball around the input. However, in discrete input spaces like text, enforcing semantic proximity is harder than in images. The behaviour of the embedding space can significantly differ from the word space. Therefore, previously described methods can not be utilized for text. Prior to this thesis, no method has been proposed that enhance attribution robustness in the textual domain. We aim to fill this gap.

The main points of this chapter are summarized as follows:

- Attributions methods in DNN classifiers are vulnerable to carefully crafted, imperceptible input perturbations that alter the outcome of the explanations significantly, while keeping the prediction of the DNN unchanged. This directly contradicts the robustness assumption of faithful explanations, preventing the deployment of state-of-the-art neural network architectures in real-life, safety-critical use cases. This phenomenon has been explored in the continuous input image space, but lacks assessment in discrete input modalities like text. One of the main focal points of this thesis is to investigate the robustness of attribution methods in text classification problems.

- Adversarial perturbations exploit the assumption that an ℓ_p-ball around the input accurately reflects semantic imperceptibility. This is not necessarily true for other modalities like text, where the word space can behave considerably differently than its embedding space. Especially since there is no unified definition of attribution robustness, this poses challenges in the estimation of attribution robustness, where the imperceptibility of perturbations needs to be assured. This thesis aims to accurately quantify AR in text, in light of perturbation size, and provide guidance on how to measure both attribution and input distances in the adversarial setup.

- Current methods to enhance attribution robustness operate only for continuous inputs. While they successfully train models that have more robust explanations, it is unclear how these training methods transfer to text. Moreover, they make heavy use of the fact that some widely-used attributions are gradient-based, thus exploit decision boundary curvature regularization techniques to achieve high robustness. This does not necessarily hold in the text domain. Thus, this thesis aims to introduce robust training objectives that have minimal assumptions on the nature of the data and are applicable in a wide variety of input modalities and domains.

3 Fragile Attributions in Text

"One of the basic rules of the universe is that nothing is perfect.
Perfection simply doesn't exist. Without imperfection,
neither you nor I would exist."
— Stephen Hawking

3.1 Introduction

Deep neural networks (DNNs) have become the state-of-the-art architectures for many existing machine learning tasks [85]. Yet, their *black-box* nature has raised the need for developing methods to mitigate the lack of interpretability caused by their increased complexity. A prominent method to unravel the black-box inference process of DNNs are attribution maps [9, 11, 86, 13]. However, the fragility of these methods towards imperceptible input perturbations that alter the interpretations without changing the prediction outcome of the DNNs, has damaged user trust in attribution methods. The imperceptible nature of such perturbations along with the unchanged prediction outcome is especially pernicious and directly contradicts the prediction assumption of faithfulness [19]. This prevents DNNs from being deployed on high-stakes, safety-critical applications, such as healthcare [87].

This fragility has mostly been studied in the continuous input image domain [1, 30, 29], and discrete input domains like text have seen no progress, even though this modality is arguably equally important. This is especially problematic given the increased reliance on attribu-

Part of this chapter has been published in
"Fooling Explanations in Text Classifiers". In *International Conference on Learning Representations (ICLR)*, 2022 [31]

tions as explanations and the attention mechanism as an inherently explainable method [88]. Therefore, in this chapter, we study the robustness of attributions in text classification problems. We first show that the existence of adversarial perturbations that alter attributions without changing the DNN predictions extends to text classifiers as well. To this end, we introduce a novel adversarial attack, TEXTEXPLANATIONFOOLER (TEF), that significantly alters the outcome of attribution maps with imperceptible word substitutions in the input text sequences, while keeping the predictions of the DNNs unchanged. Our experiments show that all attribution methods and classification models that we experiment on are susceptible to TEF perturbations. We then show that, similarly to traditional adversarial attacks on predictions, TEF perturbations transfer to and alter the outcome of attribution methods and models other than the ones they were computed for. This enables us to then take a step towards defining universal attacks on attributions, which require no knowledge of the underlying classification architecture or attribution method used at attack time. Specifically, we summarize our contributions as follows:

- We provide a novel baseline black-box adversarial attack, TEXTEXPLANATIONFOOLER (TEF) to estimate the local robustness of attribution maps in text classification problems.
- We evaluate attribution robustness (AR) on widely used, state-of-the-art text datasets and model architectures, showing that explanation methods' output can be significantly altered with our new attack. Figure 3.1 exemplifies the fragility of explanations in text.
- We provide insight into the transfer capability of TEF to different models and explanation methods, as well as introduce semi-universal adversarial perturbations to alter explanations without requiring access to the model at attack time.

The rest of this chapter is organized as follows: Section 3.2 describes the problem formulation of estimating attribution robustness in text classifiers. Section 3.3 describes our threat model and attack that computes the adversarial input samples for estimating AR. In Section 3.4, we describe our experimental setup as well as discuss the findings of these experiments on our evaluated models. Moreover, we provide ablation studies to assess which parts of out attack influence the AR estimation the most. Finally, in the same section, we provide our transfer attack setup and extraction of semi-universal perturbations.

3.2 Problem Formulation

Given an input text samples $\boldsymbol{s}$, predicted labels $\hat{l}$, a text classifier F with embedding function E and classifier function f, and attribution method A, we define *attribution robustness* (also *explanation robustness*, AR) as written in Equation (3.1).

$$r(\boldsymbol{s}) = \max_{\tilde{\boldsymbol{s}} \in N(\boldsymbol{s})} d\big[A(\tilde{\boldsymbol{s}}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big] \tag{3.1}$$

ORIGINAL SAMPLE	TEF PERTURBED SAMPLE
romanians pitch **rumsfeld** on base location | mihail **kogalniceanu** air base , **romania** - to entice the us military to make a home here , what better symbolic appeal could the romanian government make than to rename a street here quot;george washington boulevard ? $F(\boldsymbol{s},\ \hat{l} = \text{"World"}) = 0.99$	**romanians** pitch clinton on base places | **mihail kogalniceanu** air base , **rumania** - to entice the us military to make a home here , what better symbolic appeal could the romanian government make than to rename a street here quot;george washington **boulevard** ? $F(\boldsymbol{s}_{\text{adv}},\ \hat{l} = \text{"World"}) = 0.97$ **PCC**: -0.07
forgettable horror – more **gory** than psychological – with a highly **satisfying quotient of friday** - night **excitement and** milla **power .** $F(\boldsymbol{s},\ \hat{l} = \text{"Positive"}) = 0.99$	forgettable horror – more gory than psychological – with a **highly satisfying** quotient **of** friday - night **arousal and milla** wattage . $F(\boldsymbol{s}_{\text{adv}},\ \hat{l} = \text{"Positive"}) = 0.99$ **PCC**: 0.18

Figure 3.1: Example of fragile attributions. Highlighted red words are deemed most important *towards* the predicted class by the Integrated Gradients attribution method, blue ones *against* it. By substituting words in the original sample with our TEF attack, the *Pearson Correlation Coefficient* (PCC) of word importances drops to below 0.2 while maintaining the same confidence in the correctly predicted class (denoted by F).

with

$$\underset{i \in \{1 \ldots |\mathbb{L}|\}}{\arg\max} F_i(\tilde{\boldsymbol{s}}) = \underset{i \in \{1 \ldots |\mathbb{L}|\}}{\arg\max} F_i(\boldsymbol{s}) \tag{3.2}$$

where $r(\boldsymbol{s})$ denotes the estimated robustness constant on input sample $\boldsymbol{s}$ with label $\hat{l}$, d denotes a distance measure between the attribution functions $A(\tilde{\boldsymbol{s}}, F, \hat{l})$ and $A(\boldsymbol{s}, F, \hat{l})$. The rest of the notation is kept as in Section 2.1. Equation (3.1) quantifies how different the attributions of two input samples are, given the constraint in Equation (3.2) that enforces the inputs having the same prediction outcome.

The attribution robustness estimation for a given sample $\boldsymbol{s}$ is then solved utilizing the following Equation (3.3).

$$\boldsymbol{s}_{\text{adv}} = \underset{\tilde{\boldsymbol{s}} \in N(\boldsymbol{s})}{\arg\max}\ d\big[A(\tilde{\boldsymbol{s}}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big] \tag{3.3}$$

where $\boldsymbol{s}_{\text{adv}}$ denotes the solution to the estimation, i.e., the adversarial input. The vector $\boldsymbol{s}$ denotes the original, unperturbed input and $\tilde{\boldsymbol{s}}$ the perturbed input, optimized during estimation. The solution $\boldsymbol{s}_{\text{adv}}$ gives a robustness estimate by maximizing r defined in Equation (3.1) within a local neighbourhood N of $\boldsymbol{s}$. This neighbourhood is defined by the prediction constraint in Equation (3.2), i.e., the original and adversarial samples having the same predicted class. Moreover, we only allow word substitutions and draw substitution candidates from a pretrained synonym embedding, taking only the most relevant synonyms as candidates. We further enforce each word in the adversarial sample to have the same Part-of-Speech (POS) tag, computed by SpaCy [89] to enhance grammaticality. Moreover, we do not allow stop words to be substituted. These neighbourhood constraints encourage the adversarial

samples' semantic proximity to the original text. Our definition of attribution robustness and its neighbourhood constraints reflect other formulations in current research [1, 76, 90] and the prediction assumption of faithful explanations [19].

3.3 Threat Model and Robustness Estimator

In this Section, we describe our novel attack, TEXTEXPLANATIONFOOLER (TEF), to estimate the robustness of attribution maps in text.

We define our attack algorithm to estimate AR as a black-box attack. It only queries the model to obtain its output logits and the accompanied explanations of the inference process. The model might access its gradients to compute explanations, but the attack only utilizes the resulting explanations, no gradient or architectural information. We restrict the valid input perturbations to token substitutions, specifically insertions and deletions of tokens are forbidden, as they alter input lengths. Moreover, stop words are not permitted to be substituted. As such, TEF consists of the following two steps, with the schematic code written in Algorithm 1.

Step 1 - Word importance ranking (Lines 1-3 of Algorithm 1)
First, an importance ranking is extracted for each token of the input sample. Specifically, by iterating through each token in $\boldsymbol{s}$, we compute $I_{w_i} = d\big[A(\boldsymbol{s}_{w_i \to 0}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big]$ for each token w_i in $\boldsymbol{s}$, where $\boldsymbol{s}_{w_i \to 0}$ denotes the input sequence $\boldsymbol{s}$ with the i-th word masked to the zero embedding token. The input tokens are then sorted by the I_{w_i} values in a decreasing fashion, and high importance words are prioritized during substitution.

Step 2 - Candidate selection (Lines 4-19 of Algorithm 1)
For each word w_i in $\boldsymbol{s}$ sequentially, iterated over by decreasing I_{w_i}, a set of substitution candidates $\mathbb{C}$ of $|\mathbb{C}| = N$ elements is extracted. This candidate set is constructed from the counter-fitted GloVe [91] synonym embeddings by the authors of [92]. The candidates are then filtered by their Part-Of-Speech (POS) in the sentence, tagged by SpaCy [89], only allowing replacements with equal POS. Stop words are also discarded from $\mathbb{C}$. Subsequently, the word w_i is then separately replaced by each candidate c_k in $|\mathbb{C}|$, resulting in the candidate sentence $\tilde{\boldsymbol{s}}_{w_i \to c_k}$. If $\tilde{\boldsymbol{s}}_{w_i \to c_k}$ passes the prediction filter, it is considered as final candidate for replacing w_i. The *final selection* as replacement for w_i is then made to be the $c_k \in \mathbb{C}$ that maximizes $d\big[A(\tilde{\boldsymbol{s}}_{w_i \to c_k}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big]$. The algorithm is aborted when the ratio of number of replacements n to sentence length exceeds the maximum value ρ_{max}.

3.4 Experiments and Results

In this section, we present an extensive evaluation of our attribution robustness (AR) estimation attack, TEF, for text sequence classification problems. We examine the performance of TEF and study the impact of different factors on its robustness evaluation performance. We

Algorithm 1 `TextExplanationFooler (TEF)`

Input: Input sentence $\boldsymbol{s}$ with predicted class $\hat{l}$, classifier F, attribution A, attribution distance d, number of synonyms N, maximum perturbation ratio ρ_{max}
Output: Adversarial sentence $\boldsymbol{s}_{\text{adv}}$

$\boldsymbol{s}_{\text{adv}} \leftarrow \boldsymbol{s}$, $d_{max} \leftarrow 0$, $r \leftarrow 0$
for $w_i \in \boldsymbol{s}$ **do**
    $I_{w_i} = d\big[A(\boldsymbol{s}_{w_i \to 0}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big]$
for $w_j \in \langle w_1, ..., w_{|\boldsymbol{s}|} \rangle \mid I_{w_{m-1}} \geq I_{w_m}\ \forall m \in \{2, ..., |\boldsymbol{s}_{\text{adv}}|\}$ **do**
    if $w_j \in \mathbb{S}_{\texttt{Stopwords}}$ **then**
        continue
    $\mathbb{C}_j \leftarrow \texttt{SynonymEmbeddings}(w_j, N)$
    $\mathbb{C}_j \leftarrow \texttt{POSFilter}(w_j, \mathbb{C}_j, \boldsymbol{s}_{\text{adv}})$
    for $c_k \in \mathbb{C}_j$ **do**
        $\tilde{\boldsymbol{s}}_{w_j \to c_k} \leftarrow$ Replace token w_j in $\boldsymbol{s}_{\text{adv}}$ with c_k
        if $\underset{i \in \{1:|\mathbb{L}|\}}{\arg\max} F(\tilde{\boldsymbol{s}}_{w_j \to c_k}) = \hat{l}$ **then**
            $\tilde{d} \leftarrow d\big[A(\tilde{\boldsymbol{s}}_{w_i \to c_k}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big]$
            if $\tilde{d} > d_{max}$ **then**
                $\boldsymbol{s}_{\text{adv}} \leftarrow \tilde{\boldsymbol{s}}_{w_i \to c_k}$
                $d_{max} \leftarrow \tilde{d}$
                $n \leftarrow n + 1$
    if $\rho = \frac{n+1}{|\boldsymbol{s}|} > \rho_{max}$ **then**
        return $\boldsymbol{s}_{\text{adv}}$
return $\boldsymbol{s}_{\text{adv}}$

find that our attack effectively reduces the correlation of original and attacked attributions on all datasets and models. Moreover, we describe our transfer and semi-universal attacks and examine their robustness estimation performance, showing that even under circumstances where the model and explainer are unknown to the attacker, TEF perturbations transferred from other models decrease attribution robustness effectively.

3.4.1 Evaluation Setup

Our TEF attack is evaluated on five commonly used public text classification datasets, AG's News [26], MR reviews [26], IMDB Movie Reviews [93], Fake News Dataset [94] and Yelp [95]. We train six different word embedding-based architectures for each dataset, namely a CNN, an LSTM, an LSTM containing a single attention layer with one head (LSTMAtt) and three state-of-the-art finetuned transformer-based architectures, BERT [65], RoBERTa [96] and XLNet [97]. Table 3.1 contains a summary of our model performances as well as details on the datasets. The text samples are tokenized with the default English SpaCy [89] tokenizer for the CNN, LSTM and LSTMAtt models and embedded with the pretrained GloVe 6B 300-dimensional word vectors [91]. The transformer-based models use their own pretrained tokenizers and embeddings. We use PyTorch [98] with Captum [99] to implement our models and explainers

DATASET	CNN	LSTM	LSTMATT	BERT	ROBERTA	XLNET	Mean $\|\boldsymbol{s}\|$	$\|\mathbb{L}\|$
AG'S NEWS	89.7%	90.8%	91.4%	94.2%	94.0%	93.8%	45	4
IMDB	82.0%	87.2%	87.3%	89.4%	93.3%	93.7%	270	2
FAKE NEWS	98.9%	99.6%	99.6%	99.8%	100.0%	100.0%	919	2
MR	73.0%	76.4%	78.0%	82.2%	87.7%	86.3%	22	2
YELP	49.0%	54.8%	60.0%	62.6%	67.6%	-	159	5

Table 3.1: Accuracies, average text length and number of classes of our models trained on the five datasets.

and the Hugging Face Transformers library [100] to finetune the transformer architectures on our datasets.

We evaluate the robustness of three commonly used explanation methods in natural language processing. These are Saliency Maps (S), Integrated Gradients (IG) and the Attention mechanism (A), defined in Section 2.2. We use S and IG in combination with all our architectures, Attention only with LSTMAtt, BERT, RoBERTa and XLNet. During the attack, we set the attribution distance d of Equation (3.1) to be $d(\tilde{\boldsymbol{a}}, \boldsymbol{a}) = 1 - \frac{\text{PCC}(\tilde{\boldsymbol{a}}, \boldsymbol{a})+1}{2}$, with PCC denoting the Pearson Correlation Coefficient [101] of original and adversarial attributions $\tilde{\boldsymbol{a}}$ and $\boldsymbol{a}$. We then report the standard Pearson Correlation Coefficient (PCC), Kendall's Rank Order Correlation (ROC) [102], Spearman's Correlation Coefficient (SCC) [103] and the Top-10%, Top-30% and Top-50% intersections of original and adversarial attribution to measure AR in Equation (3.1). These are common metrics that correspond to human measures of AR [1, 76, 90]. All of these metrics can be used as distance measure in Equation (3.1), however, they correlate, thus would lead to similar findings. Additionally, in order to quantify imperceptibility of perturbations, the semantic similarity of adversarially perturbed and unchanged sentences is reported, along with the relative increase of average perplexity of the perturbed samples, given by the GPT-2 [104] pretrained language model. Semantic similarity (SemS) is measured by the cosine similarity between the sentence embeddings produced by the Universal Sentence Encoder (USE) [105]. This is a sentence embedding widely used in adversarial attacks on text [68, 67]. Perplexity increase (Δ_{PP}) indicates how much the *likelihood* of the perturbed data has decreased, given a language model, and is often used to validate language models [4].

Due to the lack of related work in this field, we compare the AR estimation performance of TEF to our RANDOMATTACK (RA) baseline. RA serves as an agnostic attack, utilizes a random word importance ranking in Step 1 of TEF and selects a random synonym in the final selection in Step 2. POS and stop word filters are still utilized in RA to keep linguistic constraints intact.

3.4.2 Robustness of Explanations

In order to assess the attribution robustness (AR) of the aforementioned models and explainers, we vary the parameter ρ_{max} of TEF, which denotes the maximum ratio of perturbed tokens in the input sample. A larger ρ_{max} value leads to lower attribution correlation, as potentially more words are substituted in the input. We then capture the aforementioned metrics PCC,

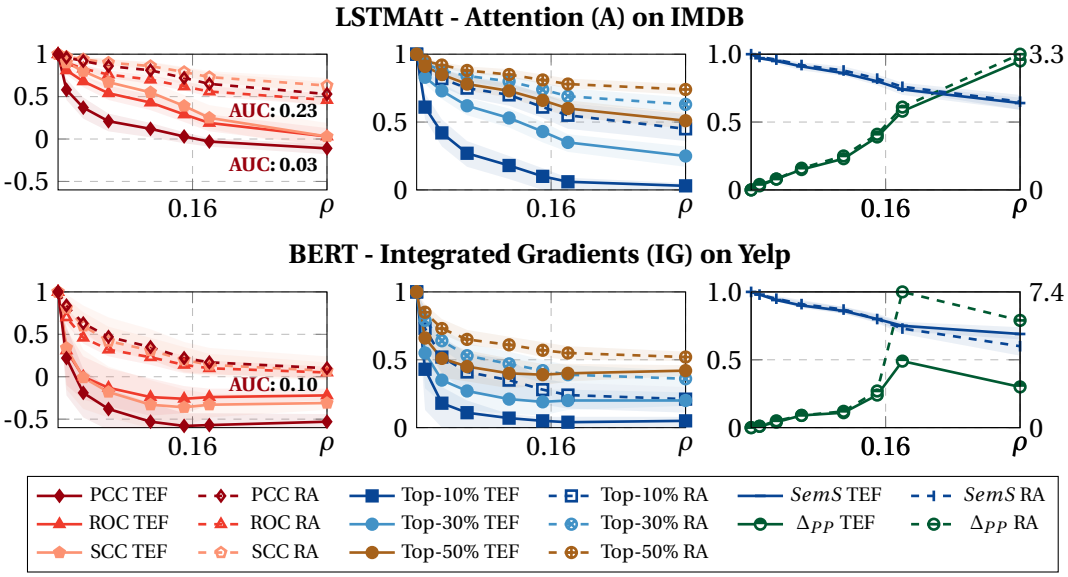


Figure 3.2: Robustness of attribution maps on several architectures and explainers. We plot the average correlations (PCC, ROC, SCC) (left), the Top-10%, Top-30% and Top-50% intersections (middle), the semantic similarity and increase of average perplexity (right) as functions of the perturbed ratio ρ. Dashed lines indicate the metrics for our RANDOMATTACK (RA). The AUC indicates the area under the PCC curve, lower values correspond to overall lower feature attribution correlations in the operation interval of ρ. The perplexity increase values are indicated on the right axis of the right column, all other metrics on the left.

ROC, SCC, SemS, Top-10%/30%/50% intersections – which indicate the tokens with highest attribution values – and Δ_{PP} to evaluate AR. Lower correlation and Top-K intersection values indicate lower robustness of the attribution methods, as adversarial attribution values do not correlate with the original ones – are thus *dissimilar*. Additionally, in order to quantify performance of our attack over the whole operation interval of $0 \leq \rho_{max} \leq 0.4$, we compute the area under the Pearson Correlation Curve (AUC). A lower value of AUC corresponds to lower robustness overall, as correlation values are lower. We note that a particular value of ρ_{max} does not guarantee that all input samples have exactly ρ_{max} ratio of perturbed tokens. Therefore, we quantize our samples based on their actual, resulting perturbed ratio ρ such that samples with similar ρ are grouped together. These bins are computed per dataset, ensuring the comparability of resulting curves and AUCs for each plot. Moreover, we choose the number of candidates in Step 2 of TEF to be $N = |\mathbb{C}| = 15$, as it is a good trade-off between TEF estimation performance and attack run time. As expected, we find that TEF is able to significantly outperform the baseline provided by RA in terms of all AR metrics, on all datasets, models and explanation methods considered in this work. The semantic similarity decreases with increasing ρ and stays above 0.7 in most cases. A subset of these results is shown in Figure 3.2, the rest can be found in Appendix A. The observation that TEF perturbations significantly outperform RA and yield lower correlation and top-k intersection values, together with the fact that resulting samples share predictions with the non-perturbed ones effectively highlights that the explanations given by these models and attribution methods lack faithfulness.

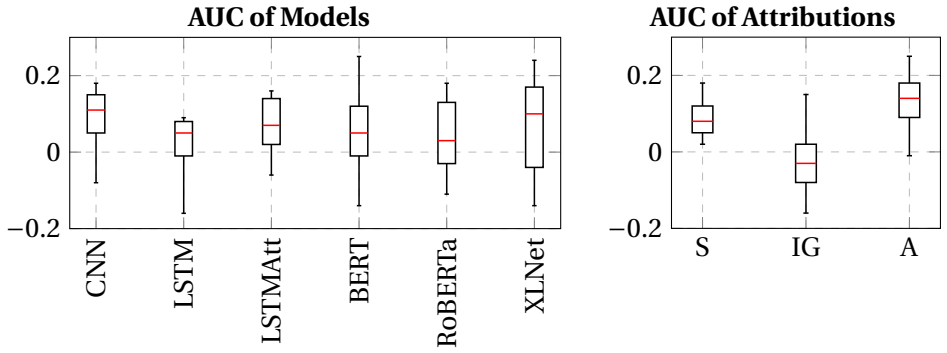


Figure 3.3: Box plot of the TEF Area Under the PCC Curves (AUC) from Figure 3.2 aggregated over attribution methods and datasets (left), as well as aggregated over classifier models and datasets (right). Lower values mean lower Pearson correlation over the evaluated interval $0 \leq \rho \leq 0.32$, thus less robust architectures or attribution methods. No particular classifier model tends to be significantly more robust than others, however, we find that the explanation method Attention (A) is more robust than S and IG, with IG being the least stable method.

In addition to these, in Figure 3.3, we aggregate the AUC values over the datasets and plot the marginal AUCs for each model and attribution method we evaluate on. We do not find that any architecture is significantly more robust to TEF perturbations, as AUC values are similar for each model (Figure 3.3 left). However, the attention mechanism of transformers seems to be more robust to perturbations than other explanation methods, as attention tends to result in higher AUC values (Figure 3.3 right). The least robust attribution is IG, yielding lower AUC values than A and S.

3.4.3 Ablation Study

In addition to the fully random attack described in the previous section, we compare TEF to our semi-random attacks RANDOMIMPORTANCE (RI) and RANDOMSYNONYM (RS). We randomize the word importance ranking of TEF (RI) but keep the selection of best final synonym, and we randomize the final synonym selection of TEF (RS) but keep the word importance ranking respectively. Figure 3.4 shows our findings for these experiments, along with comparisons to RANDOMATTACK (RA). The PCC curves and the AUC values show that RI consistently outperforms RS in terms of PCC over the whole operation interval of ρ. Moreover, the impact of word importance ranking diminishes with increasing ρ, especially for shorter datasets like MR. This can be observed by RS performing closer to RA for high ρ values.

3.4.4 BERT's Attention Layers and Heads

BERT's attention weights can be used to help gain insight into a models prediction by understanding which parts of the input are most attended to [106]. Our BERT models have 12 layers with 12 attention heads (144 heads in total), each producing a distribution of attention weights over its inputs. Estimating the AR of all heads together is not useful, as effects would average out. Therefore, we run TEF to estimate the robustness of each head separately. Figure

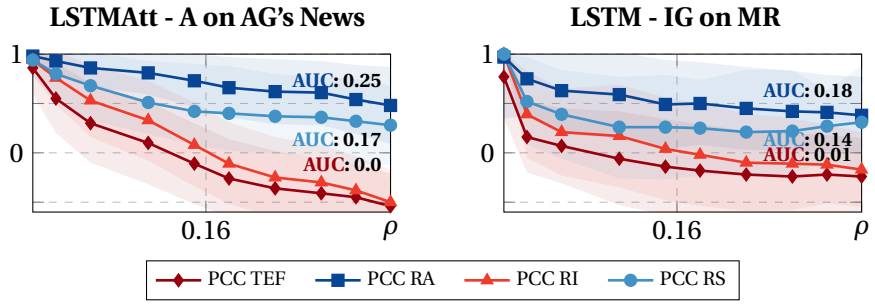


Figure 3.4: Ablation study of TEF. We compare the PCC of TEF, RA, the RANDOMIMPORTANCE (RI) attack and the RANDOMSYNONYM (RS) attack. We find that RI behaves slightly worse than TEF, while RS behaves slightly better than RA in terms of reducing attribution correlation over all ρ values.

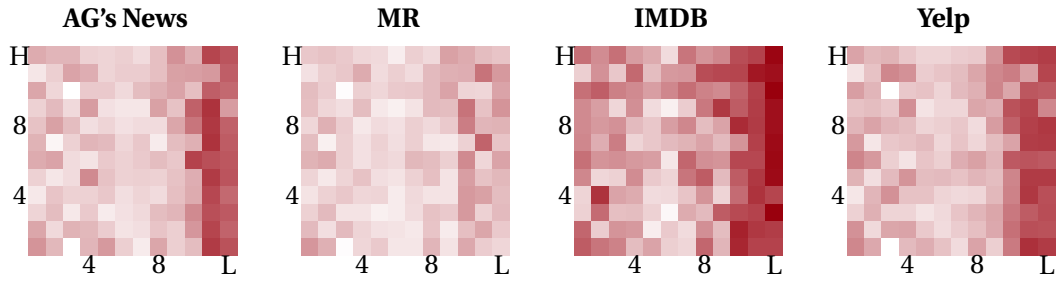


Figure 3.5: Estimated robustness of BERT attention weights on different layers (X-axis) and heads (Y-axis) for $\rho_{max} = 0.2$. Red cells indicate average PCC values close to 0, hence less robust attention head weights, while white cells have average PCCs close to 1. Attention heads in later layers tend to be less robust, while heads within a layer are equally robust in most layers.

3.5 contains the average PCCs of the attention weights before and after perturbing the inputs with TEF. We find that attention weights in later layers tend to be more susceptible to input perturbations than in earlier layers. Moreover, heads within a layer tend to be comparably robust. We leave a thorough, theoretical analysis of this phenomenon to future work. We conclude that the increasing reliance on attention weights to provide inherent interpretations to BERT predictions needs careful investigation, especially in safety-critical applications.

3.4.5 Transferability of Perturbations to Models and Explanation Methods

The adversary does not necessarily possess information about the deployed model or the exact method to produce the accompanying explanations. Therefore, it is crucial for systems to be as resistant to transfer attacks as possible in order to evade perturbations constructed on similar models and explanations.

Thus, we examine how our classifiers and attribution methods react to transfer attacks computed with TEF. We alter the input samples for a given model and explanation method with TEF, then evaluate the PCC of attributions on the same samples but different architectures and explainers. The results are found in Figure 3.6. We observe that transfer attacks perform

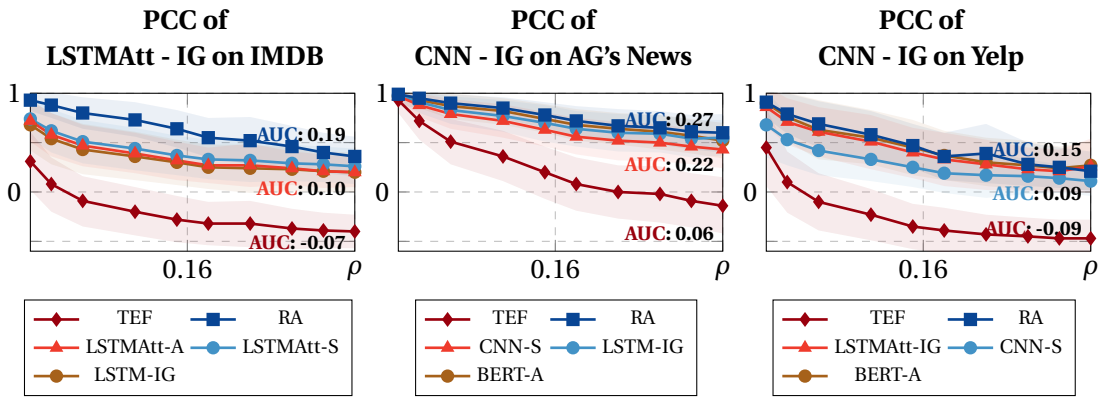


Figure 3.6: Transfer capabilities of TEF to other models and explanation methods. The lines indicate the estimated PCC of TEF perturbations transferred from the indicated models and explanations. TEF and RA curves indicate the PCC curve of optimal TEF and RA perturbations respectively, without transfer.

better than RA, some even by approx. 0.4 in terms of average PCC decrease in the operation area of $\rho \approx 0.1$. However, as expected, they significantly fall short of the performance of TEF. Therefore, we conclude that transferring TEF perturbations across models and explainers effectively highlights fragility of explanations, but TEF provides tighter AR bounds without transfer.

3.4.6 Semi-Universal Perturbations

In this section, we take a step towards defining universal perturbations, similarly to the work of the authors [61] and [69]. These provide fast and computationally cheap perturbations during attack time that are able to mislead classifiers with pre-computed perturbations. However, we attack the explanations of text classifiers, instead of their predictions and call our perturbations *semi-universal attack policies*.

We split the test dataset into two equally sized parts, the attack set and the evaluation set. We utilize the former for constructing our semi-universal attack policies and the latter to evaluate how effectively our semi-universal attack alters the attributions maps of our models.

First, for each sample in the attack set, we compute the optimal TEF perturbation for all our models and explainers. We then extract statistics of these perturbations, which are the most common replacement and the replacement frequency for each replaced token and sort these by decreasing frequency. These are our *semi-universal attack policies*, seen in Table 3.2. During this phase, we query the model for predictions and explanations, as we compute optimal TEF perturbations.

Second, we evaluate our semi-universal attack that utilizes the aforementioned policies to alter explanations of classifiers. The inputs to this attack are a text sample, a semi-universal policy and a maximum perturbed ratio ρ_{max}. The attack iterates over the policy, starting with

AG's News

Token	# repl.	Replacement
reuters	146k	goldman
said	131k	avowed
new	130k	nouvelle
ap	107k	ha
oil	72.8k	tar
...	...	...
workers	10.9k	labourers
...	...	...
zone	2.9k	field

IMDB

Token	# repl.	Replacement
movie	430k	cinematographic
film	338k	cine
good	122k	decent
great	103k	whopping
bad	102k	wicked
...	...	...
amazing	17.1k	staggering
...	...	...
scary	6.8k	fearful

MR

Token	# repl.	Replacement
movie	8.2k	cinematographic
film	8.0k	cinematographic
story	2.6k	conte
good	2.5k	decent
comedy	2.4k	humorist
...	...	...
triumph	139	victory
...	...	...
shines	69	glows

Table 3.2: Semi-universal attack policies for different datasets.

the token in the first row and finishing with the last. Whenever the current token is found in the input text sample, it is replaced with the replacement token in the list. If the perturbed ratio exceeds ρ_{max}, the attack is aborted. In such a way, perturbed inputs are created without querying the model during attack time. The actual perturbation for each text sample depends on the sample, hence the name semi-universal attack policy. The resulting samples are then evaluated on a given model and explanation method. Representative results are given in Figure 3.7. We conclude that our semi-universal policies are effective in reducing attribution correlation when the adversary has no access to the target model and explanation method, as indicated by the lower AUC values of the semi-universal PCC curves.

3.5 Conclusion

In this chapter, we introduced a novel black-box attack called TEXTEXPLANATIONFOOLER (TEF), that successfully perturbs input data such that the outcome of popular explanation methods in text classification is significantly altered, but not the prediction of the classifier. This attack highlights the lack of robustness of current text attribution methods and provides a baseline estimator for attribution robustness. We compared it to the random attack, showing its superior performance to it on five different, widely used datasets. Moreover, our exper-

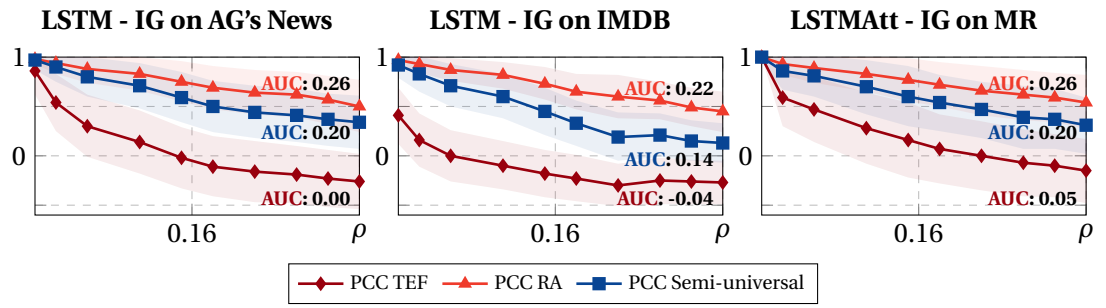


Figure 3.7: Average PCC of the indicated architectures and explainers after applying the semi-universal perturbations (Semi-universal), compared to TEF and RA attacks. The semi-universal attack successfully decreases the correlation of original and attacked attribution maps.

iments show perturbations computed with TEF transfer across classifier architectures and attribution methods, a similar behaviour to traditional adversarial attacks. Finally, we showed the existence of semi-universal perturbation policies that are capable of altering explanations without querying the model during attack-time, even without having access to perturbations for those models.

We provided first evidence of the lack of attribution robustness in DNN text classifiers, which prompts us to consider the following shortcomings of our methods and study this phenomenon further in the next chapter. First, TEF uses the counter-fitted synonym embeddings [92] to extract word substitution candidates. These embeddings are finetuned GloVe-embeddings [91] on a handcrafted synonym-antonym dataset and contain synonyms for only single words, including no contextualized information. Moreover, such a synonym embedding space is not readily available for most use cases and can be complex to extract for a specific domain. Second, the definition of attribution robustness in Equation (3.1) does not take input perturbation size into account, thus does not control how perceptible the alterations are. Arguably, perturbations that alter attributions equally but are more perceptible should result in higher robustness estimates. Hence, in the next chapter, we provide a better definition of AR in text classification problems and introduce a novel estimator that gives tighter bounds on the true robustness of attribution methods.

4 Attribution Robustness Estimation through Semantic Awareness

"Dream in a pragmatic way."
— Aldous Huxley

4.1 Introduction

In the previous chapter, we showed the existence of adversarial perturbations of text attributions by providing a baseline attack that alters attributions in DNN classifiers while maintaining the correct prediction outcome. The proposed black-box attack, TEXTEXPLANATIONFOOLER (TEF), maximizes the distance between attributions of original and perturbed inputs with word substitutions drawn from a synonym embedding space. This effectively alters the resulting explanations, highlighting the brittleness of attribution maps and contradicting the robustness assumption of faithful explanations [31, 19].

In this chapter, we study the phenomenon of fragile attributions in text further. First, we point out that the definition of attribution robustness (AR) in the previous chapter does not take the semantic distance of original and adversarial samples into account. Therefore, certain adversarial samples produced by TEF can be out-of-context and lead to the same estimated robustness as other, smoother and more fluent adversarial inputs, which are arguably much more pernicious. Therefore, in light of this aspect of robustness, we introduce a novel definition of AR that takes the (semantic) distance between original and adversarial input texts into account, besides the attribution distance. This allows for differentiating perturbations based

Part of this chapter has been published in
"Estimating the Adversarial Robustness of Attributions in Text with Transformers", *Preprint.* Under Review. 2022 [31].

on perceptibility as well.

Second, the candidate extraction of TEF with pretrained synonym embeddings [92], introduced in the previous Chapter 3, only takes single words into account and does not consider their context. This can result in quite perceptible perturbations and non-fluent adversarial input samples. To mitigate this, we introduce a novel, context-aware AR estimator in this chapter, whose candidate extraction utilizes transformer-based masked language models (MLMs), and therefore considers the context of the replacement words. Our results show that this indeed leads to smoother and more fluent adversarial samples. Moreover, MLMs are readily available in many different text domains where vocabularies contain very specific expressions, such as healthcare, while synonyms embeddings might be hard to construct or train in these domains.

Hence, it is fundamental to develop general methods that can effectively estimate the behaviour and robustness of the networks and attributions in the presence of input perturbations, and accurately quantify the perceptibility of those perturbations. In this chapter, we aim to provide solutions to these problems. We summarize the contributions as follows:

- We introduce a definition of attribution robustness (AR) in text classification that takes into account both the attribution distance and perceptibility of perturbations.
- We propose a benchmark set of metrics to effectively capture aspects like semantic distance to original, smoothness and grammaticality of perturbed inputs. This is key to understand the perceptibility of small adversarial input perturbations in text.
- We introduce a novel and powerful attack algorithm, CONTEXT-AWAREEXPLANATION-ATTACK (CEA), which is shown to consistently outperform state-of-the-art adversaries and therefore allows us to provide tighter estimates of attribution robustness in text classifiers. CEA utilizes masked language models (MLMs) for context-aware candidate extraction. This is crucial, as domain-specific MLMs are becoming increasingly available, making them a progressively attractive alternative to less effective, custom synonym embeddings on which current methods have to rely. Figure 4.1 exemplifies adversarial perturbations of our novel CEA attack compared to TEF, introduced in Chapter 3.
- We speed up robustness estimation with the usage of distilled language models and batch masking.

This chapter is organized as follows. First, in Section 4.2, we give our novel definition of attribution robustness that incorporates the perceptibility of perturbations. Then, in Section 4.3, our methods to characterize distances in text are described. Building on these two sections, our novel, context-aware AR estimator is introduced in Section 4.4. Finally, Section 4.5 contains our experiments and results on the attack described in the previous section.

ORIGINAL SAMPLE	CEA PERTURBED SAMPLE	TEF PERTURBED SAMPLE
peek at the week : ben vs. the streak | yet another risky game for that patriots winning streak , now at 21 . pittsburgh hasn # 39;t lost at home , and rookie quarterback ben roethlisberger hasn # 39;t lost , period . $F(\boldsymbol{s},\ \hat{l}$ = "Sports") = 0.99	peek at the playoffs : ben vs. the steelers | yet another risky game for that patriots winning streak , now at 21 . pittsburgh hasn # 34 lost at home , and rookie quarterback ben roethlisberger hasn # 39;t lost , period ≥ $F(\boldsymbol{s}_{\text{adv}},\ \hat{l}$ = "Sports") = 0.95 **PCC**: 0.02 *SemS*: 0.97, $r(\boldsymbol{s})$: 14.9	hoodwink at the zou : suis vs. the wave | yet another risky game for that patriots winning streak , now at 21 . pittsburgh hasn # 39;t lost at home , and rookie quarterback ben roethlisberger hasn # 39;t lost , period . $F(\boldsymbol{s}_{\text{adv}},\ \hat{l}$ = "Sports") = 1.0 **PCC**: 0.22 *SemS*: 0.9, $r(\boldsymbol{s})$: 3.4
press the delete key . $F(\boldsymbol{s},\ \hat{l}$ = "Negative") = 0.99	hit the delete key . $F(\boldsymbol{s}_{\text{adv}},\ \hat{l}$ = "Negative") = 0.95 **PCC**: -0.05 *SemS*: 0.98, $r(\boldsymbol{s})$: 30	newspaper the delete key . $F(\boldsymbol{s}_{\text{adv}},\ \hat{l}$ = "Negative") = 0.95 **PCC**: 0.6 *SemS*: 0.8, $r(\boldsymbol{s})$: 1.1
intel seen readying new wi - fi chips | intel corp . this week isexpected to introduce a chip that adds support for a relativelyobscure version of wi - fi , analysts said on monday , in a movethat could help ease congestion on wireless networks . $F(\boldsymbol{s},\ \hat{l}$ = "Sci/Tech") = 0.78	intel seen readying wireless wi - fi chips | intel corp . this week isexpected to launch a specification that added support for a relativelyobscure version of wi - fi , analysts said on monday , in a movethat could help ease congestion on wireless networks . $F(\boldsymbol{s}_{\text{adv}},\ \hat{l}$ = "Sci/Tech") = 0.95 **PCC**: 0.27 *SemS*: 0.98, $r(\boldsymbol{s})$: 20	intel seen readying nouveau wi - fi chips | intel corp . this week isexpected to insert a dies that summing support for a relativelyobscure version of wi - fi , analysts said on monday , in a movethat could help ease congestion on wireless networks . $F(\boldsymbol{s}_{\text{adv}},\ \hat{l}$ = "Sci/Tech") = 0.95 **PCC**: 0.28 *SemS*: 0.91, $r(\boldsymbol{s})$: 4

Figure 4.1: Three examples of fragile attribution maps in text sequence classifiers. In each row, careful alteration of the original sample results in significantly different attribution maps while maintaining the prediction confidence F in the correctly predicted class. Red words have positive attribution values, i.e. contribute towards the true class, while blue words with negative attributions against it. Our novel CEA attack yields perturbed samples that have lower Pearson Correlation Coefficient (PCC) values between the words highlighted by the attribution method in the original and perturbed inputs, as well as higher semantic similarity values (*SemS*) of the original and adversarial sentences, compared to the baseline TEF attack. This results in higher estimated robustness constants $r(\boldsymbol{s})$ (see Section 4.2), thus lower robustness of the classifiers against attacks.

4.2 Semantic-Aware Attribution Robustness in Text

In the previous chapter, we defined AR as the maximal attribution distance with a given locality constraint in the search space. Here we take this notion further and argue that the extent of the input perturbation is important to take into account. Two adversarial samples with similarly altered attributions might in fact strongly differ in terms of how well they maintain semantic similarity to the original sample (see 3$^{\text{rd}}$ example in Figure 4.1). This suggests that a proper measure of attribution robustness should ascribe higher robustness to methods that are only vulnerable to larger perturbations while being impervious to imperceptible ones. Thus, we define attribution robustness for a given text sample $\boldsymbol{s}$ with true and predicted label

$\hat{l}$ as functions of both resulting attribution distance and input perturbation size, written in Equation (4.1).

$$r(\boldsymbol{s}) = \max_{\tilde{\boldsymbol{s}} \in N(\boldsymbol{s})} \frac{d\big[A(\tilde{\boldsymbol{s}}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big]}{d_s(\tilde{\boldsymbol{s}}, \boldsymbol{s})} \tag{4.1}$$

with the constraint that the predicted classes of $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$ are equal, written in Equation (4.2).

$$\underset{i \in \{1 \dots |\mathbb{L}|\}}{\arg\max} F_i(\tilde{\boldsymbol{s}}) = \underset{i \in \{1 \dots |\mathbb{L}|\}}{\arg\max} F_i(\boldsymbol{s}) \tag{4.2}$$

Here, d denotes the distance between attribution maps $A(\tilde{\boldsymbol{s}}, F, \hat{l})$ and $A(\boldsymbol{s}, F, \hat{l})$, F the sequence classifier with output probability F_i for class i, and d_s the distance of input text samples $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$. $N(\boldsymbol{s})$ indicates a neighbourhood of $\boldsymbol{s}$: $\{N(\boldsymbol{s}) = \tilde{\boldsymbol{s}} \mid d_s(\tilde{\boldsymbol{s}}, \boldsymbol{s}) < \varepsilon\}$ for a small ε. This definition is inspired by the robustness assumption of faithful explanations [19]. Contrary to the definition of AR in the previous chapter, Equation (4.1) incorporates the distance of original and adversarial input samples, in the denominator. This allows for differentiating AR with perturbations that alter attributions equally, but differ strongly in terms of perceptibility. The estimated robustness of an attribution method A on a model F then becomes the expected per-sample $r(\boldsymbol{s})$ on dataset $\mathbb{S}$, see Equation (4.3).

$$r(A, F) = \mathbb{E}_{\boldsymbol{s} \in \mathbb{S}}\big[r(\boldsymbol{s})\big] \tag{4.3}$$

We call this r the estimated attribution robustness (AR) constant. The robustness of attribution method A on the model F is inversely proportional to $r(A, F)$, as high values correlate with large attribution distances and small input perturbations, which indicates low robustness.

4.3 Distances in Text Data

In order to compute the attribution robustness constant r from Equation (4.1), the distance measures in the numerator and denominator of Equation (4.1) need to be defined. In explainable AI, it is often argued that only the relative rank between input features or tokens is important when explaining the outcome of a classifier, or even only the top few features. Users frequently focus on the features deemed most important to explain a decision and disregard the less important ones [90, 1, 76]. Therefore, it is common practice [31, 32] to use correlation coefficients and top-k intersections as distance measures between attributions, as these tend to reflect the human understanding of altered explanations better. For this reason, we utilize the Pearson Correlation Coefficient (PCC) [101] as attribution distance $d\big[A(\tilde{\boldsymbol{s}}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big] = 1 - \dfrac{1 + \mathrm{PCC}\big[A(\tilde{\boldsymbol{s}}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big]}{2}$ of Equation (4.1).

Measuring distance between text inputs in the adversarial setting is not as straightforward as in the image domain, where ℓ_p-norm induced distances are common. String distance metrics [107] can only be used limitedly, as two words can have similar characters but entirely different semantics. For this reason, we propose the following set of measures to effectively capture smoothness, semantic distance to original, and correctness of grammar of adversarial text

inputs in the denominator of Equation (4.1).

First, we utilize pretrained sentence encoders to measure the semantic textual similarity between the original and adversarial text samples. This can be computed by the cosine similarity between the sentence embeddings of the two text samples, given as

$$d_{\mathrm{s}}(\tilde{\boldsymbol{s}}, \boldsymbol{s}) = 1 - \frac{s_{cos}[E_s(\tilde{\boldsymbol{s}}), E_s(\boldsymbol{s})] + 1}{2} \tag{4.4}$$

where d_{s} denotes the semantic distance between samples $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$, s_{cos} the cosine similarity, and $E_s(\tilde{\boldsymbol{s}})$ and $E_s(\boldsymbol{s})$ the sentence embeddings of the two input samples. The semantic textual similarity provides a measure how close the two inputs are in their semantic meaning. To this end, the Universal Sentence Encoder [105] is widely-used in adversarial text setups [68, 31]. However, this architecture is not state-of-the-art on the STSBenchmark dataset [108], a benchmark used to evaluate semantic textual similarity. Therefore, we utilize a second sentence encoder architecture trained by the authors of [109], MiniLM. This model achieves close to state-of-the-art performance on the STSBenchmark while maintaining a low computational cost. Evaluating with a top-performing embedding model on the STSBenchmark dataset acts as an automated proxy to human evaluation of the semantic distances of original and adversarial samples, which would require manual effort and would not scale well.

Our second input distance is derived from the perplexity of original and adversarial inputs $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$. We capture the relative increase of perplexity when perturbing the original sentence $\boldsymbol{s}$, given the pretrained GPT-2 language model [104] (Equation 4.5).

$$d_s(\tilde{\boldsymbol{s}}, \boldsymbol{s}) = \frac{PP(\tilde{\boldsymbol{s}}|L) - PP(\boldsymbol{s}|L)}{PP(\boldsymbol{s}|L) + \varepsilon} \tag{4.5}$$

where d_{s} denotes the distance between inputs $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$, PP the perplexity of the text sample given the GPT-2 language model L and ε is a small constant. Intuitively, this measure indicates how natural the resulting adversarial inputs are. Positive values indicate higher perplexity of adversarial samples, thus less fluent text than the original, while negative values correspond to lower perplexities of altered inputs, thus even more fluent inputs than the original text samples.

Lastly, we capture the increase of grammatical errors in the input samples using the LanguageTool API [1]. As grammatical errors are easily perceived by the human observer, they significantly contribute to the perceptibility of adversarial perturbations [66].

4.4 Context-Aware Robustness Estimation

Given our AR definition in Equation (4.1), in order to estimate the true robustness of an attribution method for a given model, all possible input sequences $\tilde{\boldsymbol{s}}$ within the neighbourhood

[1] https://languagetool.org

N of $\boldsymbol{s}$ would have to be searched. This is a computationally intractable problem. Therefore we restrict the search space (the neighbourhood N) to sequences $\tilde{\boldsymbol{s}}$ that only contain token substitutions from the predefined vocabulary set $\mathbb{W}$. Moreover, we restrict the ratio of substituted tokens in the original sequence to ρ_{max}, considering only $|\mathbb{C}|$ number of possible substitutions for each token in $\boldsymbol{s}$. The number $|\mathbb{C}|$ is chosen to yield high attribution distance while keeping the computation cost low, detailed in Section 4.5.2. This way, we reduce the total perturbation set from $|\mathbb{W}|^{|\boldsymbol{s}|}$ to $|\mathbb{C}|^{|\boldsymbol{s}|\cdot\rho_{max}}$ samples. The adversarial sequence $\boldsymbol{s}_{\text{adv}}$ then becomes the perturbed sequence that maximizes $r(\boldsymbol{s})$ from Equation (4.1)

We estimate AR with our novel CONTEXT-AWAREEXPLANATIONATTACK (CEA). CEA is a black-box attack, only having access to the model's prediction and the accompanying attributions, not the intermediate representations, architectural information or gradients. Algorithm 2 contains the pseudocode for our CEA attack. Similarly to TEF from the previous Chapter, CEA consists of the following two steps.

Step 1: Word importance ranking (Lines 1-3 of Algorithm 2)
The first step extracts a priority ranking of tokens in the input text sample $\boldsymbol{s}$. For each word w_i in $\boldsymbol{s}$, CEA computes $I_{w_i} = d\big[A(\boldsymbol{s}_{w_i\to 0}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big]$, where $\boldsymbol{s}_{w_i\to 0}$ denotes the token w_i in $\boldsymbol{s}$ set to the mask token and d denotes the attribution distance measure in Equation (4.1), described in the previous section. The tokens in $\boldsymbol{s}$ are then sorted by descending values of I_{w_i}. Thus, we estimate words that are *likely* to result in large attribution distances and prioritize those for substitutions towards building explanation attacks.

Step 2: Candidate selection and substitution (Lines 4-21 of Algorithm 2)
The second step substitutes each highest ranked token in $\boldsymbol{s}$, denoted by decreasing I_{w_i}, with a token from a candidate set $\mathbb{C}$, in descending importance order. Each highest ranked token has its separate candidate set $\mathbb{C}$. CEA extracts these sets by masking the words and querying a transformer-based masked language model (MLM). In order to keep the computational costs low, we utilize the DistilBERT pretrained masked language model [110], a BERT-MLM with significantly fewer parameters and more computationally efficient. In each candidate set, stop words are filtered out. Then, similarly to TEF, the tokens w_i are replaced by each candidate c_k in $\mathbb{C}_j$ and the prediction constraint is applied. If this constraint is fulfilled, the candidate that maximizes $d\big[A(\tilde{\boldsymbol{s}}_{w_i\to c_k}, F, \hat{l}), A(\boldsymbol{s}, F, \hat{l})\big]$ is selected as final substitution for w_i. The algorithm aborts if $n_{\text{max}} = |\boldsymbol{s}|\cdot\rho_{max}$ words have been substituted.

In order to further reduce computational cost, CEA uses batch masking. Thus, instead of masking each token separately and querying the MLM for candidates, the first $n_b = |\boldsymbol{s}|\cdot\rho_b$ most important tokens are masked at once and the language model is queried for candidates for all of these masked tokens. Here, n_b denotes the number, ρ_b the ratio of words in $\boldsymbol{s}$ to be masked at once. For instance, during AR estimation of a 100 word text sample, given $\rho_{max} = 0.15$ and $\rho_b = 0.05$, the MLM is queried only $(100\cdot 0.15)/(100\cdot 0.05) = 3$ times with batch masking instead of $100\cdot 0.15 = 15$ times without it. We compared the runtime of CEA using non-distilled [111] and distilled [110] BERT MLMs, with and without batch masking, and

Algorithm 2 `Context-AwareExplanationAttack (CEA)`

Input: Input sentence $\boldsymbol{s}$ with predicted class $\hat{l}$, classifier F, attribution A, attribution distance d, DistilBERT-MLM L, number of candidates N, maximum perturbation ratio ρ_{max}, batch masking ratio ρ_b
Output: Adversarial sentence $\boldsymbol{s}_{\text{adv}}$

$\boldsymbol{s}_{\text{adv}} \leftarrow \boldsymbol{s}$, $d_{max} \leftarrow 0$, $n \leftarrow 0$
for $w_i \in \boldsymbol{s}$ **do**
    $I_{w_i} = d\big[A(\boldsymbol{s}_{w_i \to 0}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big]$
$\boldsymbol{s}_B \leftarrow \langle \boldsymbol{s}_{1\ldots b}, \boldsymbol{s}_{b+1\ldots 2b}, \ldots, \boldsymbol{s}_{|\boldsymbol{s}|-b+1\ldots|\boldsymbol{s}|} \rangle$ with $I_{w_{b-1}} \geq I_{w_b}\ \forall j \in \{2, \ldots, |\boldsymbol{s}_B|\}$ and $\forall b \in \{1, .., |\boldsymbol{s}_j|\}$
for $\boldsymbol{s}_b \in \boldsymbol{s}_B$ **do**
    $\mathbb{C}_{\mathbf{b}} \leftarrow L(\boldsymbol{s}_{b \to [MASK]}, \boldsymbol{s}_{\text{adv}})$
    for $w_j \in \boldsymbol{s}_b$ **do**
        if $w_j \in \mathbb{S}_{\texttt{Stopwords}}$ **then**
            continue
        for $c_k \in \mathbb{C}_j$ **do**
            $\tilde{\boldsymbol{s}}_{w_j \to c_k} \leftarrow$ Replace w_j in $\boldsymbol{s}_{\text{adv}}$ with c_k
            if $\underset{i \in \{1:|\mathbb{L}|\}}{\arg\max} F(\tilde{\boldsymbol{s}}_{w_j \to c_k}) \neq \hat{l}$ **then**
                continue
            $\tilde{d} = d\big[A(\tilde{\boldsymbol{s}}_{w_i \to c_k}, F, \hat{l}), A(\boldsymbol{s}, F, \hat{l})\big]$
            if $\tilde{d} > d_{max}$ **then**
                $\boldsymbol{s}_{\text{adv}} \leftarrow \tilde{\boldsymbol{s}}_{w_i \to c_k}$
                $d_{max} \leftarrow \tilde{d}$
                $n \leftarrow n + 1$
        if $\rho = \frac{n+1}{|\boldsymbol{s}|} > \rho_{max}$ **then**
            return $\boldsymbol{s}_{\text{adv}}$
return $\boldsymbol{s}_{\text{adv}}$

found considerable performance increase with batch masking and distillation. The results are reported in Section 4.5.2.

4.5 Experimental Results

In this section, we present our AR estimation experiments. Specifically, we describe the evaluation setup and results with our novel robustness definition. We show that CEA consistently outperforms our baseline method, TEXTEXPLANATIONFOOLER (TEF) in terms of the attribution robustness constant r described in Section 4.2. Thus, we convey that CEA extracts smoother adversarial samples that are able to alter attributions more significantly than TEF. Finally, we compare the runtime of CEA to TEF and show that CEA achieves comparable runtimes, while still outperforming TEF in the previously mentioned aspects.

DATASET	CNN	LSTM	LSTMATT	BERT	ROBERTA	XLNET
AG'S NEWS	89.7%	90.8%	91.4%	94.2%	94.0%	93.8%
MR	73.0%	76.4%	78.0%	82.2%	87.7%	86.3%
IMDB	82.0%	87.2%	87.3%	89.4%	93.3%	93.7%
YELP	49.0%	54.8%	60.0%	62.6%	67.6%	-
FAKE NEWS	98.9%	99.6%	99.6%	99.8%	100.0%	100.0%

Table 4.1: Accuracies of each trained classifier trained. Our models achieve comparable results to state-of-the-art performance for each dataset.

4.5.1 Setup

We evaluate the robustness constant r estimated by CEA on the AG's News [26], MR Movie Reviews [26], IMDB [93], Yelp [95] and the Fake News datasets [94]. We train a CNN, an LSTM, an LSTM with an attention layer (LSTMAtt), a finetuned BERT [111], RoBERTa [96] and XLNet [97] classifier for each dataset. The accuracies of the models can be found in Table 4.1, along with a detailed description of the models in Appendix B. We estimate the robustness of the Saliency Maps (S), Integrated Gradients (IG) and Self-Attention (A) attribution methods. The CNN and LSTM architectures are used in combination with S and IG, the remaining LSTMAtt, BERT, RoBERTA and XLNet are used with all three attributions. Thus, we evaluate 16 combinations of models and attributions for each dataset.

We vary the ρ_{max} parameter of CEA between 0.01 and 0.4. A value of ρ_{max} does not necessarily lead to the actual perturbed ratio of tokens ρ to be $\rho = \rho_{max}$ due to the prediction constraint. We set the batch masking size $\rho_b = \min(\rho_{max}, 0.15)$, as the MLM was trained by masking approx. 15% of the tokens [110]. We set $|\mathbb{C}| = 15$, as larger values do not tend to result in better estimation in terms of r, but in significantly higher attack runtimes. This makes our experiments comparable to TEF from the previous chapter.

Our attack and experiments are implemented in PyTorch [98], utilizing the Hugging Face Transformer library [100], Captum [99] and SpaCy [89]. We run each experiment on an NVIDIA A100 GPU with three different seeds and report the average results.

4.5.2 Results

We report the following metrics as functions of the true perturbed ratio ρ. The average PCC values of original and adversarial attribution maps indicate the amount of change in explanations. Lower values correspond to larger attribution changes. The input distance between text samples is captured by the semantic textual similarity values of the original and adversarial samples, measured by the cosine similarity between the USE [105] and MiniLM [109] sentence embeddings ($SemS_{USE}$ and $SemS_{MiniLM}$), as well as the relative perplexity increase (Δ_{PP}). The average increase in number of grammatical errors (GE) after perturbation is also reported. Using these values, we report the estimated robustness constants r_{USE}, r_{MiniLM} and r_{PP}, according to Equation (4.3). In each of these, the scaled PCC (introduced

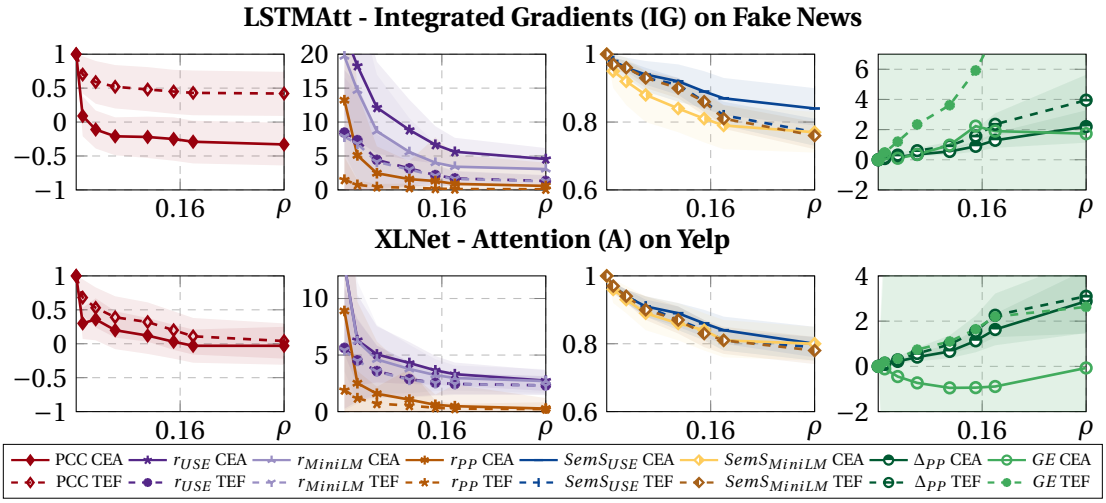


Figure 4.2: AR metrics as functions of the ratio of perturbed tokens ρ. We plot the mean and standard deviation of the Pearson correlations (PCC) between original and adversarial attributions, semantic similarities ($SemS$), relative perplexity increase (Δ_{PP}), increase of number of grammatical errors (GE) in original and adversarial text inputs and the estimated AR robustness constants (r). We compare these values for our novel CONTEXT-AWAREEXPLANATIONATTACK (CEA - continuous lines) and the baseline TEXTEXPLANATIONFOOLER (TEF - dashed lines). We observe consistent improvement of robustness estimation with CEA compared to TEF, reflected in higher r-values in the second column. This is attributed to both lower PCC values, higher semantic similarities of perturbed sentences to the original ones and lower adversarial perplexity of CEA perturbations.

in Section 4.3) is used as attribution distance. We compare these metrics for our novel CEA algorithm and the baseline TEF from Chapter 3. Figure 4.2 reports these metrics as a function of the true perturbed word ratio ρ. The continuous lines contain the reported metrics for our CEA attack, the dashed lines for the baseline TEF. The figure shows that CEA perturbations alter explanations more (lower average PCC values) while yielding adversarial samples equally or more semantically similar to the original inputs than TEF (higher average $SemS$, lower average PP and GE values). Moreover, the perplexity increase is consistently lower for CEA perturbations, leading to more fluent adversarial samples. This is well-captured by resulting robustness constants r, which are consistently higher for CEA than TEF, showing both that our AR definition of Equation (4.1) is a suitable indicator for AR in text classifiers, and that CEA estimates this robustness better than the state-of-the-art TEF attack. The rest of the results is reported in Appendix B.

In order to quantify the overall performance of CEA over the whole operation interval of ρ, we compute the area under the estimated r curves (2$^{\text{nd}}$ column in Figure 4.2). These are calculated as the integral $\text{AUC}_r = \int_\rho r(A,F)d\rho$. High AUC_r values correspond to high r-values, thus low overall attribution robustness. We then compare the resulting AUC_r estimated with our CEA algorithm to the competitor method TEF. Figure 4.3 shows the relative increase of AUC when estimating with CEA rather than TEF, for each of the 16 combinations of models and attribution methods for a given dataset. For instance, a value of 0.5 indicates a relative increase of 50% in estimated AUC_r, i.e., if TEF results in $\text{AUC}_r = 1.0$, CEA yields $\text{AUC}_r = 1.5$. We plot the AUC_r increase estimated with the semantic textual similarities from USE (AUC_r^{USE}), MiniLM

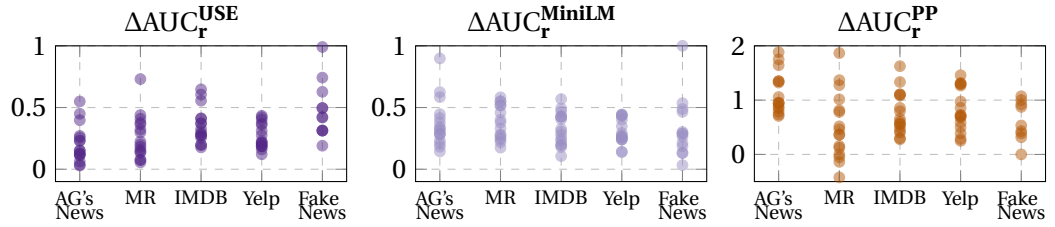


Figure 4.3: Relative increase Δ of AUC$_r$ when estimating the robustness constants r with CEA compared to TEF. Each point corresponds to one of the 16 combinations of model and attribution method, on the indicated dataset. The r-values are estimated with the PCC as attribution similarity, varying the input distance measures d_s as described in Section 4.5.2. We observe a relative increase of 0.3 – 1.5 for almost all models, attribution maps and datasets evaluated on. This shows that CEA consistently provides better perturbations that alter attributions more while being more fluent and semantically similar to the unperturbed input.

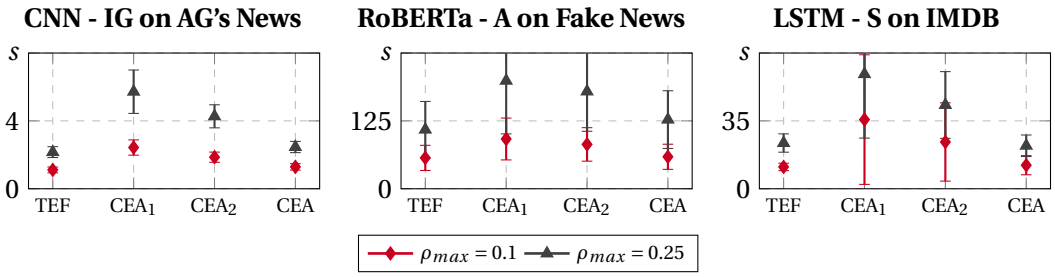


Figure 4.4: Per-sample runtime (s) of our AR estimator algorithm versions. CEA, with a distilled MLM and batch masking, achieves comparably fast estimation as TEF, while CEA with a non-distilled BERT MLM (CEA$_1$) is the slowest estimator, with a relative increase in runtime of approx. 1.5-2.5 compared to TEF. Distillation of the MLM (CEA$_2$) improves the runtime by around 25-35% compared to CEA$_1$.

(AUC$_r^{MiniLM}$) and with the relative perplexity increase (AUC$_r^{PP}$). The attribution distance in the numerator of r is set to the PCC described in Section 4.3. We observe an increase in AUC$_r$ of 0.3 – 0.5 with USE and TSE, and 0.5 – 1.5 with PP for most models, attribution maps and datasets. This further shows that CEA consistently yields higher robustness constants r than TEF, providing better perturbations that alter attributions more while being less perceptible.

Finally, we note that querying transformer-based masked language models (MLMs) is computationally expensive. Naively substituting the synonym extraction from TEF with an MLM-based candidate extraction results in a significant increase in estimation time. Therefore, we use the methods described in Section 4.4 to achieve comparable estimation time in our CEA algorithm and TEF. Figure 4.4 contains the per-sample attack time for TEF, CEA with the non-distilled BERT MLM (CEA$_1$), CEA with DistilBERT MLM (CEA$_2$) and our final CEA algorithm with DistilBERT MLM and batch masking, for $\rho_{max} \in \{0.1, 0.25\}$. We observe that CEA$_1$ results in a significant increase in mean estimation time by a factor of around 2 compared to TEF on both a smaller, medium and a larger datasets. Using CEA$_2$ for estimating AR decreases the runtime by a large margin compared to CEA$_1$. Finally, when applying both a distilled MLM and batch masking — CEA, the per-sample attack time is comparable to the baseline TEF, while maintaining better AR estimation performance.

4.6 Conclusion

This chapter introduced a novel notion of attribution robustness in text classifiers. Crucially, it does not only take the attribution distance into account, but also incorporates the size of the perturbations, which contributes significantly to the perceptibility of attacks. To this end, we introduced semantic textual similarity measures, the relative perplexity increase and the number of grammatical errors as ways to effectively quantify perturbation size in text. This allows for accurately quantifying the robustness aspect of faithful explanations and the comparison of different models and attribution methods in this regard. Given our robustness definition, we introduced CONTEXT-AWAREEXPLANATIONATTACK (CEA), a novel state-of-the-art attack method that results in a tighter estimator for attribution robustness in text classification problems. It is a black-box estimator that utilizes a distilled MLM with batch masking to extract adversarial perturbations with small computational overhead. Our experiments show that CEA outperforms the baseline TEF by altering DNN attributions more significantly with less perceptible perturbations.

Accurately quantifying the robustness of explanations is a crucial first step towards training robust, safely applicable DNNs in many critical areas, such as medicine, law or finance. Often, the focus of assessing faithfulness lies on disproving its assumptions and thus disproving its faithfulness. Arguably, this approach is suboptimal [19] and a more nuanced understanding of how robust attributions are in a specific configuration of input space search is more useful. The methods introduced in this chapter allow to do exactly that.

Equipped with the methods to estimate AR, the next step for safely deploying DNNs in real-life scenarios is to train networks that yield robust attributions. It is crucial to not only quantify the robustness, but also achieve robust enough DNN explanations that give reliable insight into the network's decision process. The next chapter discusses a novel, general framework, usable both in text and image input spaces, which allows for training DNNs with state-of-the-art robust attribution maps.

5 Training Robust Attributions

"I suppose it is tempting, if the only tool you have is a hammer,
to treat everything as if it were a nail..."
— Abraham Maslow

5.1 Introduction

In the previous two chapters of this thesis, we established a thorough study on the estimation of attribution robustness in text classification problems. It is important to understand under which circumstances current attribution methods deliver faithfully robust explanations and in what environments they might fail to do so. Fragile explanations are highly problematic in critical scenarios where the decision outcome needs to be accompanied by a sound explanation, such as healthcare, law or finance.

Understanding the limitations of interpretation methods in terms of robustness is essential. However, it is equally if not more important to explore methods that enhance their robustness in order to increase user trust in DNN architectures. All current methods [112, 29, 30, 76] that train networks with robust attributions focus on the image modality. In fact, the only attempt to enhance attribution robustness in text was made by the authors of [32], who show that

Part of this chapter has been published in
"FAR: A General Framework for Attributional Robustness". In *The 32nd British Machine Vision Conference (BMVC)*, 2021 [90]

Part of this chapter has been accepted for publication in
"DARE: Towards Robust Text Explanations in Biomedical and Healthcare Applications". In *The 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023

traditional adversarial training increases AR in text as well. However, since current methods exploit gradient optimization techniques to enhance AR in networks, which rely on every point in the input space being a valid input, it is not straightforward to use these for discrete input modalities like text. In this chapter, we aim to solve this problem and introduce methods to successfully enhance AR in text classifiers as well.

To this end, we develop a general framework, FAR, to train robust attributions, which has minimal assumptions about the input modality at hand. FAR relies on a general optimization objective that adapts traditional adversarial training to attributions, even allowing for providing attribution targets. To solve the optimization problems of our FAR framework, we make use of the AR estimator in the previous chapter, along with its domain-agnostic nature to introduce an attack that provides domain-plausible perturbations.

As robust attributions are especially important in safety-critical domains, we utilize FAR to provide a case study on AR in biomedical healthcare text classification. Specifically, we show how our estimators can be adapted to biomedical data and study the robustness of attributions, as well as how training with our FAR framework enhances AR in this domain. Our experiments show that attributions are fragile in healthcare text classifiers as well, and that our FAR objective successfully improves AR in transformer-based classifiers, setting a new state-of-the-art baseline.

In order to gain a better understanding how FAR performs in comparison to other robust training methods, we instantiate FAR to train robust networks in image classification problems as well. Our novel training objectives, AAT and AdvAAT – derived from FAR, directly optimize for high correlation of original and adversarial attribution maps in a small ε-bound neighbourhood of the input images, while being able to train robust predictions as well. Our experiments show that these two objectives outperform or perform comparably to current methods in the image domain, while being more generally applicable.

In light of the above aspects, we summarize the contributions of this chapter as follows:

- We define a general *framework for attributional robustness* (FAR) as general problem formulation for training robust attributions. Key aspects of this framework are:
  - It allows for separate optimization for robust predictions and explanations,
  - It generalizes to more explanation methods and attribution distances than current methods,
  - It allows for providing ground truth explanations.
- We show how to solve the optimization problems of the FAR objectives for text by reinterpreting the methods of previous chapters on AR estimation. This leads us to introduce DOMAINADAPTIVEARESTIMATOR (DARE), a novel AR estimator based on domain-plausible attacks that can be used to solve the adversarial sample search in a domain-specific way.

- Our experiments show that attributions are fragile in the critical domain of biomedical text as well, and that FAR successfully trains networks with state-of-the-art attribution robustness.
- We provide two instantiations of FAR for image classifiers, our AAT and AdvAAT methods that directly optimize for maximal correlation between original and adversarial feature importance within a small neighbourhood of the input, to compare the performance of FAR to other state-of-the-art methods in terms of AR. We find that FAR outperforms these on two image datasets experimented on and performs comparably on three others.

We organize this chapter as follows. In Section 5.2, we describe our novel framework, FAR, its assumptions, optimization problem, the solvers for the optimization and we show how FAR is a general formulation of other, existing robust training objectives. Next, in Section 5.3, we introduce how FAR operates in text classification problems, on a biomedical domain use case. To this end, we describe our datasets and how our AR formulation can be adapted to multilabel classification problems. Lastly, we discuss our findings of the experiments on the aforementioned datasets. Finally, in Section 5.4, we show how FAR can be instantiated for the image space and compare its performance in terms of AR to current existing methods.

5.2 Framework for Attribution Robustness (FAR)

In this section, we introduce FAR, our general problem formulation of improving AR in DNNs. We then show how the optimization problem of FAR can be solved with slightly modifying previously described algorithms. Lastly, we derive existing robust attribution training methods in image classification problems from this framework, showcasing the general nature of our formulation.

Our framework builds around the prediction assumption of faithful explanations [19]. For this, we extend the classical notion of adversarial training [112] to attribution maps, considering the following assumptions:

- *Similar attribution maps.* The first assumption of our framework is that similar inputs should give near-identical explanations, if the predictions are equal. It is often argued in explainable AI [1] that only the relative ranks of input features matter, or even only a subset of them. Therefore, in this chapter, we utilize the cosine similarity to measure attribution similarity, as well as correlation metrics like Kendall's Rank Order Correlation [102] and the Top-K Intersection [1].
- *Perceptually identical inputs.* The second assumption of FAR is that adversarial perturbations do not change the underlying semantics of the data and its ground truth labels. In textual input classifiers, we make use of the neighbourhood assumptions in previous chapters, namely the synonym embeddings and MLM candidate extractions, linguistic constraints along with semantic textual similarity measures, to keep ground

truth labels of inputs intact. In addition to these, we require the same predicted class $\arg\max_{i\in\{1\dots|\mathbb{L}|\}} F_i(\tilde{\boldsymbol{s}}) = \arg\max_{i\in\{1\dots|\mathbb{L}|\}} F_i(\boldsymbol{s})$ of perturbed and original inputs. The latter constraint motivates the assumption that similar inputs should have similar explanations. In image classification, in addition to the prediction constraint, we utilize ℓ_p-ball restrictions of size ε around the input to ensure unchanged ground truth labels of the data.

- *Target attributions.* We observe that the second term in the attribution distance of Equation (4.1) in our AR definition of the previous chapter, which is the attribution map of the unperturbed input, can act as target (ground truth) attribution. The robustness of attributions can thus be interpreted as how robustly the networks attribute to these targets within a small neighbourhood of the input. Generally, ground truth for attribution maps is not available, therefore, we use the attributions of the unperturbed inputs as targets. However, allowing to provide these targets could provide useful for datasets in which they are given.

5.2.1 Optimization Problem

Given the above points, we define our framework as a *robust training loss* which formulates generic objectives for robustifying any smooth attribution method and dissimilarity. The objectives can be used to enhance robustness of the attributions separately from the inference outcome, or jointly encouraging adversarial and attributional robustness. This is important because most of the times, robustness is required both in attributions as well as predictions of the network to achieve trustworthiness [19]. Thus, the extraction of adversarial samples becomes the following maximization problem, written in Equation (5.1).

$$\boldsymbol{s}_{\text{adv}} = \underset{\tilde{\boldsymbol{s}}\in N(\boldsymbol{s})}{\arg\max}\Big\{(1-\gamma)\cdot l_c(\tilde{\boldsymbol{s}}, F, \hat{l}) + \gamma\cdot d\big[A(\tilde{\boldsymbol{s}}, F, \hat{l}),\ A^T(\boldsymbol{s}, F, \hat{l})\big]\Big\} \tag{5.1}$$

with $\boldsymbol{s}_{\text{adv}}$ denoting the adversarial sample, N the neighbourhood space of the original sample $\boldsymbol{s}$, l_c the classification loss of classifier F on $\boldsymbol{s}$ with label $\hat{l}$. Moreover, d denotes a distance between attribution maps A, γ a constant with $0 \leq \gamma \leq 1$, controlling the trade-off between maximizing prediction and attribution loss. We use A^T as the target attributions which can be provided during the robustness estimation. However, if those are not utilized, we set $A^T(\boldsymbol{s}, F, \hat{l}) = A(\boldsymbol{s}, F, \hat{l})$, i.e., target attributions are the attributions of the unperturbed input samples. Optionally, the following prediction constraint P can be utilized to keep the classification outcome unchanged.

$$\mathrm{P}\big[F(\boldsymbol{s}_{\text{adv}}), \hat{l}\big] = \underset{i\in\{1\dots|\mathbb{L}|\}}{\arg\max} F_i(\boldsymbol{s}_{\text{adv}}) = \underset{i\in\{1\dots|\mathbb{L}|\}}{\arg\max} F_i(\boldsymbol{s}) = \hat{l} \tag{5.2}$$

Given the above extraction of adversarial samples, robust networks can be trained by solving the following optimization problem in Equation (5.3).

$$\theta^* = \underset{\theta}{\arg\min} \sum_{\boldsymbol{s}\in\mathbb{S}} \Big\{(1-\delta)\cdot l_c(\boldsymbol{s}_{\text{adv}}, F, \hat{l}) + \delta\cdot d\big[\boldsymbol{s}_{\text{adv}}, F, \hat{l}),\ A^T(\boldsymbol{s}, F, \hat{l})\big]\Big\} \tag{5.3}$$

Algorithm 3 `Adversarial DomainAdaptiveAREstimator` (DARE)

Input: Input $\boldsymbol{s}$ with label set $\hat{l}$, classifier F, attribution A, distance metric d, prediction constraint P, flag to apply prediction constraint `keep_pred`, language model MLM, number of candidates $|\mathbb{C}|$, maximum perturbation word ratio ρ_{max}, regularization constant γ
Output: Adversarial sentence $\boldsymbol{s}_{\text{adv}}$

$\boldsymbol{s}_{\text{adv}} \leftarrow \boldsymbol{s}$, $d_{max} \leftarrow 0$, $n \leftarrow 0$
$I_{\boldsymbol{s}} = \nabla_{\boldsymbol{X}} \Big\{ (1-\gamma) \cdot l_c(\boldsymbol{s}, F, \hat{l}) + \gamma \cdot d\big[A(\boldsymbol{s}+\varepsilon, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big] \Big\}$
for $w_i \in \langle w_1, ..., w_{|\boldsymbol{s}|} \rangle | I_{m-1} \geq I_m \forall\, m \in \{2, ..., |\boldsymbol{s}|\}$ **do**
    if $w_i \in \mathbb{S}_{\texttt{Stopwords}}$ **then**
        continue
    $\mathbb{C}_i \leftarrow \text{MLM}(w_i, \boldsymbol{s}, |\mathbb{C}|)$
    for $c_k \in \mathbb{C}_i$ **do**
        $\tilde{\boldsymbol{s}}_{w_{ik}} \leftarrow$ Replace w_i in $\boldsymbol{s}_{\text{adv}}$ with c_k
        if not $\text{P}\big[F(\tilde{\boldsymbol{s}}_{w_{ik}}), \hat{l}\big]$ **and** `keep_pred` **then**
            continue
        $\tilde{d} = (1-\gamma) \cdot l_c(\tilde{\boldsymbol{s}}_{w_{ik}}, F, \hat{l}) + \gamma \cdot d\big[A(\tilde{\boldsymbol{s}}_{w_{ik}}, F, \hat{l}),\ A(\boldsymbol{s}, F, \hat{l})\big]$
        if $\tilde{d} > d_{max}$ **then**
            $\boldsymbol{s}_{\text{adv}} \leftarrow \tilde{\boldsymbol{s}}_{w_{ik}}$
            $d_{max} \leftarrow \tilde{d}$
            $n \leftarrow n+1$
        if $\rho = \frac{n+1}{|\boldsymbol{s}|} > \rho_{max}$ **then**
            return $\boldsymbol{s}_{\text{adv}}$
return $\boldsymbol{s}_{\text{adv}}$

with θ^* denoting the optimal network parameters and δ a constant with $0 \leq \delta \leq 1$, controlling the robustness regularization and the notation kept from the previous sections.

Formulating the AR problem as above has the following advantages. First, the choice of A is not fixed - the framework can be used to robustify any attribution method. Second, the domain of explanations and input data is not coupled, hence the shortcomings of current methods in images, such as input-alignment training [29] do not exist for our framework. Third, the choice of A^T is not fixed, therefore target (ground truth) explanations can be provided if present, and robust explanations can be trained with respect to these. Fourth, the attribution distance metric d can be chosen to any smooth d, depending on the use case. Lastly, by varying the regularization parameters γ and δ, we can adjust the trade-off between robust attributions and predictions. Note that setting $\gamma = 1$ and using the prediction constraint from Equation (5.2) allows for solely optimizing for robust attributions, while setting $\gamma \neq 1$ without the prediction constraint encourages both robust attributions and predictions.

5.2.2 Solving the Optimizations of FAR

The training of the networks with the robust objective in Equation (5.3) can be solved with standard gradient optimization techniques. In order to solve the inner maximization of the

adversarial search in Equation (5.1), we make use of the notions and algorithms developed in the previous chapters of this thesis. We use CEA from Section 4.4 with the following aspects and modifications. First, we note that masked language models are in fact effective candidate extractors for word substitutions, as not only do they take context of the words into account, but can be trained on unlabelled data in an unsupervised fashion, thus pretrained models are available for many domains and use cases. This is important, because the search space for word substitutions in constrained by the candidate extractors, which ensures domain-specific imperceptibility of perturbations. Second, in order to increase the computational efficiency of the AR estimators, we substitute the word deletion-based importance ranking of CEA with a gradient-based ranking, as it only requires one forward and backward pass. Lastly, we omit the batch masking. This results in our novel AR estimator, which we call adversarial domain-agnostic AR estimator (DARE), written in Algorithm 3.

5.2.3 Recovering Existing Objectives

In this section, we show that current robust training methods for image classifiers can be derived from our general FAR framework by utilizing specific parameters, attribution distances and methods. Thus, we showcase the general nature of our problem formulation. Here, we strictly apply our notions to continuous input image classification problems, therefore the inputs $\boldsymbol{s}$ are images $\boldsymbol{x}$ and we utilize the ε-balls around the inputs $\boldsymbol{x}$ and neighbourhood functions N, i.e., $N(\boldsymbol{x}) = \{\tilde{\boldsymbol{x}} | \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_\mathrm{p} < \varepsilon\}$.

Madry's Robust Prediction [112] can be recovered by utilizing the adversarial search in Equation (5.1) with $\gamma = 0$, the training objective in Equation (5.3) with $\delta = 0$ and omitting the prediction constraint of Equation (5.2). The objective then becomes as follows:

$$\theta^* = \arg\min_\theta \sum_{\boldsymbol{x}\in\mathbb{S}} \max_{\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|_\mathrm{p}<\varepsilon} l_c(\tilde{\boldsymbol{x}}, F, \hat{l}) \tag{5.4}$$

The *Axiomatic Attribution Regularization* terms (IG-NORM and IG-SUM-NORM) in [80] can be recovered using the IG attribution map $A(\boldsymbol{x}, F, \hat{l}) = \mathrm{IG}(\tilde{\boldsymbol{x}}, \boldsymbol{x})$, where the baseline of IG is set to $\boldsymbol{B} = \boldsymbol{X}$ and the attribution distance function to $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_1$. Moreover, $\gamma = 1$, $0 < \delta < 1$ and the prediction constraint is applied. As such, the training optimization in Equation (5.3) becomes as follows.

$$\theta^* = \arg\min_\theta \sum_{\boldsymbol{x}\in\mathbb{S}} \left\{ l_c(\boldsymbol{x}, F, \hat{l}) + \delta \cdot \max_{\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|_\mathrm{p}<\varepsilon} \|\mathrm{IG}(\tilde{\boldsymbol{x}}, \boldsymbol{x})\|_1 \right\} \tag{5.5}$$

The IG-SUM-NORM training objective in [80] can be analogously derived from Equation (5.3) and (5.1) by omitting the prediction constraint and setting $0 < \gamma < 1$ and $0 < \delta < 1$. Note that $\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}) = 0$ holds due to the completeness axiom of IG [8].

The input-gradient *Spatial Alignment* regularization term introduced in [29] corresponds to utilizing the sum of positive spatial alignment of true class logits $\hat{l}$ input gradients and the negative spatial alignment of the second largest logits $\bar{l}$ input gradient as attribution map A,

written in the following Equation (5.6):

$$A(\boldsymbol{x}, F, \hat{l}, \bar{l}) = d_{\cos}\left[\nabla_{\boldsymbol{x}} F_{\bar{l}}(\boldsymbol{x}), \boldsymbol{x}\right] - d_{\cos}\left[\nabla_{\boldsymbol{x}} F_{\hat{l}}(\boldsymbol{x}), \boldsymbol{x}\right] \tag{5.6}$$

where $d_{\cos}$ denotes the cosine distance between the input gradient and the image. $\bar{l}$ is the second largest class' logit, the rest of the notation is kept from previous sections. By applying $f(a) = \log\left[1 + \exp(a)\right]$to A, using $d(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_1$ and omitting the use of target attribution maps, the regularization term in [29] can be recovered from Equation (5.3).

5.3 FAR in Text Classification

In this section, we describe our methods and experiments to utilize FAR to train robust classifiers in text. To showcase its efficacy for text classifiers, we adapt DARE and FAR to work in the biomedical and healthcare domain, utilizing the domain-agnostic nature of our AR estimator DARE (Algorithm 3) from Section 5.2.2. We chose this domain because training robust networks in healthcare is especially critical, as unfaithful prediction outcomes and explanations can have severe consequences. Moreover, labelled biomedical data is much less available than in general text domains, which showcases the advantages of utilizing pretrained MLMs as candidate extractors in DARE. Our experiments show that attributions of DNNs in this critical domain are susceptible to adversarial perturbations as well, and that our FAR robust objectives outperform the baseline adversarial training in terms of AR, establishing a new baseline for state-of-the-art robustness of attributions in text.

5.3.1 Medical Text Datasets

In healthcare, text can appear in many different forms with very heterogeneous vocabularies. Therefore, we choose three text datasets that cover different aspects of relevant use cases in the medical domain. Often, the datasets are not big enough to train models with large numbers of parameters, such as transformers. Therefore, we make heavy use of transfer learning by utilizing pretrained language models and finetuning them on our datasets.

Our first dataset, Drug Reviews (DR) [113], consists of patient reviews of different medical drugs, classified into a rating of 1 to 10 for patient satisfaction. The dataset contains 215063 samples, written in mostly layman's terms along with the names of the drugs and symptom descriptions. Given the dataset's nature, the classification model we choose is a finetuned RoBERTa model, with pretrained weights from Hugging Face [100].

The Hallmarks of Cancer [114] dataset (HoC) consists of 1852 biomedical publication abstract associated with 0 or more hallmarks of cancer [115]. The samples are peer-reviewed publication texts, containing few to no misspellings with scientific biomedical vocabulary. As the dataset contains only a small amount of samples, we finetune a pretrained BioLinkBERT [116] model from Hugging Face to achieve state-of-the-art classification accuracy on this dataset.

MLM	HoC	Drug Reviews	MIMIC-III
BERT	0.786	0.702	0.677
DistilBERT	0.733	0.599	0.580
DistilRoBERTa	0.768	**0.745**	0.604
PubMedBERT	**0.908**	0.704	0.781
BioClinicalBERT	0.775	0.629	0.847
ClinicalBigBird	-	-	0.372
Clinical-Longformer	-	-	**0.867**

Table 5.1: Top-5 accuracies of the masked language models (MLMs) on our datasets Hallmarks of Cancer (HoC), Drug Reviews and MIMIC-III. Each word in each sample of the dataset is masked and the sample is then propagated through the MLM. If the original masked word is in the top-5 predictions of the MLM, the sample counts as positive.

Lastly, we evaluate the MIMIC-III [117] Discharge Summary dataset (MIMIC). This is a set of extremely long, de-identified, free text ICU discharge summaries from patient admitted to critical care, written by medical professionals. The corresponding ICD-9 codes [118] are associated with each sample in a multilabel fashion. This dataset contains in average 2500 words per sample [117], thus traditional BERT-based models are not feasible as their runtime scales quadratically with the sequence length. Therefore, we finetune a pretrained Clinical-Longformer model [119], a Longformer MLM [120] trained on the MIMIC-III discharge summaries. For an in-depth, more detailed description of our datasets and models, we refer to Appendix C.

5.3.2 AR in Multilabel Healthcare Datasets

Many text classification datasets in healthcare do not only have one label per sample. In HoC, multiple hallmarks can be associated with an abstract, and MIMIC contains hardly any discharge summary with only one associated ICD-9 code. In these cases, the label $\hat{l}$ of an input sample $\boldsymbol{s}$ becomes a set of predicted labels, and the prediction constraint of attribution robustness in Equation (5.2) holds as long as the predicted set of labels from the original sample is equal to the one from the adversarial sample. We denote this constraint as P in our estimation algorithm DARE. Moreover, attribution methods compute maps on a per-class basis, where the overall attribution $\mathbf{A} = A(\boldsymbol{s}, F, \hat{l})$ equals the attribution of the single predicted class $\hat{l}$. In a multilabel case, we extend this notion to the sum of attributions for each predicted class, thus the overall attribution map becomes $\mathbf{A} = \sum_{l_i \in \hat{l}} A(\boldsymbol{s}, F, l_i)$.

In order to use DARE to estimate AR in the biomedical domain, we make use of the domain-agnostic nature of the candidate extractor MLM. It is crucial to extract in context, smooth word substitution candidates, whose vocabularies overlap with the vocabulary of the domain at hand. For this reason, as our substitution candidate extractors, we choose a pretrained MLM that maximizes the top-5 accuracy of predicting the words in dataset, when each is masked separately, averaged over the dataset. Consequently, we use the MLMs DistilRoBERTa [121] for Drug Reviews, PubMedBERT [122] for HoC and Clinical-Longformer [119] for MIMIC-III.

Table 5.1 summarizes the accuracies of the MLMs that we have tested.

5.3.3 Adversarial Training in Text

It has been shown in image classification [29, 76, 80] and hinted in textual domains [32] that adversarial training, as traditionally defined on the predictions of DNNs [112], not only enhances prediction robustness in classifiers, but also improves attribution robustness. Therefore, we utilize this method as a baseline to compare our robust FAR objectives to.

In an untargeted setup, adversarial training [21, 62] augments the training data with samples $\boldsymbol{s}_{\text{adv}}$ specifically computed as a function of the input to maximize the classification loss l_c, written in Equation (5.7).

$$\boldsymbol{s}_{\text{adv}} = \underset{\tilde{\boldsymbol{s}} \in N(\boldsymbol{s})}{\arg\max}\, l_c(\tilde{\boldsymbol{s}}, F, \hat{l}) \tag{5.7}$$

where N denotes the search neighbourhood of original sample $\boldsymbol{s}$, F the classifier and $\hat{l}$ the true label of sample $\boldsymbol{s}$. The classifiers then are trained following the optimization objective in Equation (5.8).

$$\theta^* = \underset{\theta}{\arg\min} \sum_{\boldsymbol{s} \in \mathbb{S}} l_c(\boldsymbol{s}_{\text{adv}}, F, \hat{l}) \tag{5.8}$$

where θ^* denotes the optimal model parameters. In order to solve the inner optimization problem in Equation (5.7), we choose the A2T [123] attack framework, as it provides flexibility in terms of candidate extraction methods and is optimized for adversarial training runtime. By adapting A2T to use our the MLMs described in Section 5.2.2, we successfully extract in-context and imperceptible adversarial samples for training.

5.3.4 Experimental Setup

For each dataset described in Section 5.3.1, we compare the attribution robustness of a classification model trained with three different training objectives: i) a vanilla natural model trained with the cross-entropy loss; ii) a model trained with adversarial training as described in Section 5.3.3 and iii) a model trained with robust FAR objectives from Section 5.2. The attribution methods evaluated are Saliency Maps (S) [9], DeepLIFT (DL) [41], Integrated Gradients (S) [8] and the models' self-attention weights (A) [13]. We choose these as they are popular methods to provide explanations for DNNs in healthcare [34]. We use DARE from Section 5.2.2, with the corresponding MLMs from Table 5.1 to extract adversarial samples and analyze the cosine similarity of original and adversarial attributions, the semantic similarity between original and adversarial input text samples and combining these two metrics, the resulting attribution robustness constants $r(\boldsymbol{s})$, described in Section 4.2. A complete set of estimation parameters is given in Table 5.2. An extensive parameter search is very time-consuming, thus left for future work.

To evaluate the semantic similarity between original and perturbed inputs, methods in the

PARAMETER	HALLMARKS OF CANCER	DRUG REVIEWS	MIMIC-III
CANDIDATE EXTRACTOR	PubMedBERT	DistilRoBERTa	Clinical-Longformer
$\rho_{\max}$	0.05	0.05	0.005
$\lvert\mathbb{C}\rvert$	5	5	3
$d(A_{\text{adv}}, A)$	cosine	cosine	cosine
$d_s(\boldsymbol{s}_{\text{adv}}, \boldsymbol{s})$	MedSTS semantic embeddings	MedSTS semantic embeddings	MedSTS semantic embeddings

Table 5.2: Hyperparameters used for estimating attribution robustness for our three datasets Hallmarks of Cancer, Drug Reviews and MIMIC-III. Candidate extractor denotes the MLM used for extracting the replacement candidates in DARE, $\rho_{\max}$ the maximum ratio of perturbed words in each sample, $\lvert\mathbb{C}\rvert$ the number of replacement candidates extracted for each word, $d(A_{\text{adv}}, A)$ the attribution distance metric and $d_s(\boldsymbol{s}_{\text{adv}}, \boldsymbol{s})$ the text input distance.

previous chapters utilize state-of-the-art sentence embeddings on the STSBenchmark dataset [108]. We argue that this is suboptimal, as it is not clear whether it captures perturbation perceptibility in the biomedical domain as well. Therefore, here we utilize the model made public by [124] to evaluate semantic distance between texts. This model is the top performing RoBERTa model on the MedSTS dataset [125], a state-of-the-art dataset for semantic similarity in the biomedical domain. We denote this similarity $MedSemS$.

Our vanilla (Van.) models are trained with the standard cross-entropy classification loss, the adversarially trained models with the A2T adversarial training framework [123], utilizing the MLMs from Table 5.1 as candidate extractors. To train our FAR robust models (FAR-IG), we use the FAR training framework described in Section 5.2.1, using DARE in Algorithm 3 to solve the inner maximization of Equation (5.3), the cosine distance as attribution distance and Integrated Gradients (IG) as attribution distance. For reproducibility, we report the full set of training parameters in Appendix C. The estimation is reported with a three-fold cross validation, averaging the results. The models and datasets are implemented in PyTorch [98] and PyTorch Lightning [126], the pretrained weights are taken from the Hugging Face library [100], with the attributions implemented with Captum [99]. The models are finetuned on the datasets using 4 NVIDIA A100 GPUs.

5.3.5 Results

Table 5.3 summarizes the results of our experiments. We observe that the non-robust vanilla models (Van.) perform poorly in terms of cosine similarity between original and adversarial attribution maps compared to their robust counterparts (Adv. and FAR-IG). Especially the attributions DeepLIFT (DL) and Integrated Gradients (IG) are significantly altered by the attacks. This is reflected in the higher estimated robustness constants $r(\boldsymbol{s})$ for the vanilla models. Thus, we conclude that training networks with no robustness objective is largely suboptimal if faithful and robust explanations are needed.

However, both the baseline adversarial training and our adapted FAR objectives are able

		$cos(A_{adv}, A)$				*MedSemS*				$r(\boldsymbol{s})$			
	MODEL	*S*	*DL*	*IG*	*A*	*S*	*DL*	*IG*	*A*	*S*	*DL*	*IG*	*A*
HoC	VAN.	0.67 ±0.22	-0.09 ±0.22	0.06 ±0.27	0.66 ±0.14	0.79 ±0.12	0.79 ±0.13	0.79 ±0.09	0.78 ±0.1	0.76 ±0.11	2.6 ±0.11	2.2 ±0.22	0.77 ±0.11
	ADV.	0.81 ±0.09	0.09 ±0.22	0.46 ±0.23	0.74 ±0.14	0.79 ±0.1	0.79 ±0.13	0.79 ±0.09	0.78 ±0.1	0.45 ±0.11	2.2 ±0.25	1.3 ±0.16	0.59 ±0.09
	FAR-IG	**0.84** ±0.08	**0.24** ±0.2	**0.65** ±0.26	**0.86** ±0.08	0.77 ±0.14	0.77 ±0.14	0.78 ±0.11	0.77 ±0.14	**0.35** ±0.12	**1.6** ±0.31	**0.8** ±0.31	**0.3** ±0.05
DRUG REV.	VAN.	0.89 ±0.12	0.25 ±0.32	0.48 ±0.35	0.72 ±0.18	0.92 ±0.08	0.92 ±0.09	0.92 ±0.09	0.91 ±0.09	0.69 ±0.07	4.1 ±0.19	3.3 ±0.22	2.1 ±0.1
	ADV.	0.91 ±0.12	0.36 ±0.3	0.49 ±0.34	0.78 ±0.17	0.91 ±0.09	0.9 ±0.1	0.91 ±0.09	0.9 ±0.09	0.45 ±0.06	3.7 ±0.17	2.8 ±0.14	1.1 ±0.09
	FAR-IG	**0.93** ±0.11	**0.77** ±0.28	**0.86** ±0.21	**0.86** ±0.12	0.9 ±0.09	0.9 ±0.09	0.9 ±0.09	0.89 ±0.1	**0.35** ±0.05	**1.2** ±0.14	**0.8** ±0.14	**0.73** ±0.07
MIMIC-III	VAN.	0.35 ±0.27	0.08 ±0.33	0.0 ±0.37	0.7 ±0.26	0.88 ±0.07	0.84 ±0.07	0.82 ±0.11	0.84 ±0.07	3.1 ±	2.9 ±0.18	2.8 ±0.15	0.94 ±0.2
	ADV.	**0.44** ±0.32	**0.12** ±0.26	0.0 ±0.45	**0.76** ±0.21	0.85 ±0.07	0.77 ±0.19	0.8 ±0.03	0.81 ±0.13	**1.9** ±0.21	**1.9** ±0.47	**2.5** ±0.27	**0.63** ±0.12
	FAR-IG	–	–	–	–	–	–	–	–	–	–	–	–

Table 5.3: Attribution robustness metrics (mean and standard deviation) of the vanilla (VAN.), adversarially trained (ADV.) and FAR-trained (FAR-IG) models, trained on our three datasets. We perform our AR estimation for the attributions S, DL, IG and A. The reported metrics are the cosine similarity between attributions of original and adversarial samples - $cos(A_{adv}, A)$ -, the semantic similarity of the two input text samples - *MedSemS* - as well as the estimated attribution robustness constant - $r(\boldsymbol{s})$ -. We conclude that the vanilla models perform poorly in terms of attribution robustness, while both adversarially and FAR-IG trained models are significantly more robust, yielding higher attribution similarities and lower $r(\boldsymbol{s})$ values. FAR-IG models outperform adversarially trained models, giving the most promising method to effectively train attributionally robust networks.

to train networks with significantly more robust attributions than vanilla training. For the HoC dataset and IG attributions, adversarial training increases the cosine similarity up to 0.46, while FAR-IG training up to 0.65. A similar trend is observable for the other models, datasets and attribution methods. FAR-IG training reduces the estimated robustness constants consistently by 40-60%, which is a significant increase in robustness. This convinces us that FAR is a feasible method to achieve robust attributions in DNNs. Figure 5.1 contains example attributions for our vanilla and robustly trained models, on the Drug Reviews dataset.

We further observe that even if our FAR-IG model is not evaluated on IG, but S, DL or A, it still consistently outperforms vanilla and adversarially trained models both in terms of $cos(A_{adv}, A)$ and $r(\boldsymbol{s})$. Therefore, we conclude that the robustness attained by FAR training with IG transfers to other attributions, further strengthening our confidence in FAR being an attractive option to train robust networks. In light of this observation, an interesting future research question is how utilizing self-attention (A) as attribution method in FAR affects robustness, as most state-of-the-art language models use transformer architectures based on self-attention.

	VANILLA	ADVERSARIAL	FAR-IG
Original	'took zoloft for 5 months. no side effects except sexual dysfunction. i didn't feel much better or happier and it made me feel really drowsy.' $F(\boldsymbol{s}, \hat{l} = "4.0") = 1.0$	'took zoloft for 5 months. no side effects except sexual dysfunction. i didn't feel much better or happier and it made me feel really drowsy.' $F(\boldsymbol{s}, \hat{l} = "4.0") = 1.0$	'took zoloft for 5 months. no side effects except sexual dysfunction. i didn't feel much better or happier and it made me feel really drowsy.' $F(\boldsymbol{s}, \hat{l} = "4.0") = 1.0$
Adversarial	'took zoloft for 5 months. no side effects except sexual dysfunction. i didn't feel much better or anything and it made me feel really drowsy.' $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = "4.0") = 1.0$ **Cosine Similarity** = -0.32 $SemS$ = 0.99	'took zoloft for 5 months. no side effects except sexual dysfunction. i didn't feel much better or stronger and it made me feel really drowsy.' $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = "4.0") = 1.0$ **Cosine Similarity** = -0.05 $SemS$ = 0.96	'took zoloft for 5 months. no side effects except nerve dysfunction. i didn't feel much better or happier and it made me feel really drowsy.' $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = "4.0") = 1.0$ **Cosine Similarity** = 0.79 $SemS$ = 0.93
Original	'i've been on this at 10mg for a few months to fight my hay fever. it has had some effect, my nose isn't running as frequently, but my eyes are watering and are itchy. if anything it has made my eyes worse. at 10mg it doesn't really do much.' $F(\boldsymbol{s}, \hat{l} = "6.0") = 1.0$	'i've been on this at 10mg for a few months to fight my hay fever. it has had some effect, my nose isn't running as frequently, but my eyes are watering and are itchy. if anything it has made my eyes worse. at 10mg it doesn't really do much.' $F(\boldsymbol{s}, \hat{l} = "6.0") = 1.0$	'i've been on this at 10mg for a few months to fight my hay fever. it has had some effect, my nose isn't running as frequently, but my eyes are watering and are itchy. if anything it has made my eyes worse. at 10mg it doesn't really do much.' $F(\boldsymbol{s}, \hat{l} = "6.0") = 1.0$
Adversarial	'i've been on this at 10mg for a few months to fight my hay fever. it has had some effect, my nose isn't running as frequently, but my eyes are watering and are itchy. if anything it has made my eyes worse. For 10mg it doesn't actually do much.' $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = "6.0") = 1.0$ **Cosine Similarity** = -0.14 $SemS$ = 0.99	'i've been on this at 10mg for a few months to fight my hay fever. it has had some effect, my nose isn't running as frequently, but my eyes are watering and are itchy. if anything it has made my eyes worse. For myself it doesn't really do much.' $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = "6.0") = 1.0$ **Cosine Similarity** = 0.24 $SemS$ = 0.99	'i've been on this at 10mg for a few months to fight my high fever. it has had some effect, my nose isn't running as frequently, but my eyes are watering and are itchy. if untreated it has made my eyes worse. at 10mg it doesn't really do much.' $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = "6.0") = 1.0$ **Cosine Similarity** = 0.46 $SemS$ = 0.88

Figure 5.1: Attribution methods in medical text classifiers trained without any robust objectives (VANILLA) are susceptible to imperceptible word substitutions. By changing a few words in the original sample (underlined), the words with originally positive attributions (red) are assigned negative values (blue), and vice versa, while keeping the prediction confidence F in the correct class unchanged. This is indicated by the Cosine Similarity between the explanations of original and adversarial samples. Attacks on attributions in networks trained with robust training objectives (ADVERSARIAL and FAR-IG) are less successful (higher Cosine Similarity values) while also being more perceptible (lower medical semantic similarity - SemS - values between original and adversarial samples).

5.4 FAR for Images

All of current methods that train attributionally robust DNN classifiers focus on the image input space. Therefore, in this section, we report the performance of FAR on image classifiers in order to be able to compare it to current methods. We provide two new instantiations of FAR — AAT and AdvAAT — that train robust attributions as well as predictions in DNN image classifiers. We introduce the adversarial solver of the optimization problems of FAR for images. Then, we describe the experiments and results on five image classification datasets, which show that our FAR instantiations outperform or perform similarly to current methods in terms of AR in image classification. Additionally, we are the first to experimentally show

the dependency of attribution robustness on the weight initialization of the networks and argue that training with our objectives lessens these dependencies. Moreover, we show that the tightness parameter β in the approximation of second order ReLU gradient significantly influences the robustness estimation.

5.4.1 Preliminaries

In case of images, we denote the dataset $\mathbb{S} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ with corresponding labels $\hat{l}_i$. Each $\boldsymbol{x}$ is expected to be in the continuous input space $\mathbb{R}^{d \times p \times t}$, where d, p and t denote the image height, width and depth respectively. Moreover, the elements of $\boldsymbol{x}$ can take real values in the interval $[0, 1]$. In this case, the embedding function is omitted, we denote $\boldsymbol{X} \coloneqq \boldsymbol{x}$ and $F \coloneqq f : \mathbb{R}^{d \times p \times t} \to \mathbb{R}^{|\mathbb{L}|}$, $f(\boldsymbol{X}) \coloneqq F(\boldsymbol{x}) = \boldsymbol{o}$, in other words, the classifier contains only a differentiable classification function f mapping the inputs directly to the output logits.

The concept of attribution robustness (AR) for image classifiers has been introduced in several current works [1, 30]. While there is no agreement of the definition of AR, most mathematical formulations build towards the conjecture that attribution maps should be similar for similar inputs, i.e., the prediction constraint of faithful explanations. Thus, we define AR for image domains as written in Equation (5.9).

$$r(\boldsymbol{x}) = \max_{\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_p < \varepsilon} d\big[A(\tilde{\boldsymbol{x}}, F, \hat{l}),\ A(\boldsymbol{x}, F, \hat{l})\big] \tag{5.9}$$

given the prediction constraint in the following Equation (5.10).

$$\arg\max_{i \in \{1...|\mathbb{L}|\}} F_i(\tilde{\boldsymbol{x}}) = \arg\max_{i \in \{1...|\mathbb{L}|\}} F_i(\boldsymbol{x}) \tag{5.10}$$

where $r(\boldsymbol{x})$ denotes the estimated attributional robustness constant of attribution map A on input $\boldsymbol{x}$ and $N(\boldsymbol{x})$ a small ε-bound on the ℓ_p-norm of the input change. The rest of the notation is kept from Section 2.1.

In image domains, we use a modified version of the iterative feature importance attack (IFIA) described in [1] to solve the inner optimization problem of FAR in Equation (5.1). It is an adapted projected gradient descent attack (PGD [127]) to alter attributions with gradient update steps and projections. We incorporate the prediction constraint, resulting in the adversarial IFIA attack written in Algorithm 4.

Algorithm 4 `Adversarial IFIA`

Input: Classifier F, input image $\boldsymbol{x}$, target class $\hat{l}$, classification loss l_c, attribution A, attribution distance d_s, norm p and bound ε, step size η, number of maximum iterations M, data input bounds b, prediction constraint P, flag to apply prediction constraint `keep_pred`, regularization constant γ

Output: Adversarial image $\boldsymbol{x}_{\text{adv}}$

$\boldsymbol{x}^0_{\text{adv}} \leftarrow \boldsymbol{x}$
while $i \leq N$ **do**
  $\mathbf{g}_t \leftarrow \nabla_{\boldsymbol{x}^{i-1}_{\text{adv}}} \left\{(1-\gamma)\cdot l_c(\boldsymbol{x}^{i-1}_{\text{adv}}, F, \hat{l}) + \gamma \cdot d\big[A(\boldsymbol{x}^{i-1}_{\text{adv}}, F, \hat{l}),\ A(\boldsymbol{x}, F, \hat{l})\big]\right\}$
  $\boldsymbol{x}^i_{\text{adv}} \leftarrow \boldsymbol{x}_{\text{adv}} + \eta \cdot \texttt{Normalize}_{\text{p}}\big(\mathbf{g}_t\big)$
  $\boldsymbol{x}^i_{\text{adv}} \leftarrow \texttt{Project}_{\text{p}}\big(\boldsymbol{x}^i_{\text{adv}}, \boldsymbol{x}, \varepsilon, b\big)$
  if not $\text{P}\big[F(\boldsymbol{x}_{\text{adv}}), \hat{l}\big]$ **and** `keep_pred` **then**
    return $\boldsymbol{x}^{i-1}_{\text{adv}}$
  $i \leftarrow i+1$
return $\boldsymbol{x}^{i-1}_{\text{adv}}$

5.4.2 Attributional Adversarial Training

Using the framework introduced in Section 5.2, we formalize our attributional adversarial training objectives for image classification, consisting of a regularization term — AAT — that is used to only optimize for robust attributions, and a robust training loss — AdvAAT — used to achieve both robust predictions and attributions. For AAT, we instantiate the adversarial sample extraction of Equation (5.1) with $\gamma = 1$, $0 < \delta < 1$ and the prediction constraint in tact. This allows for extracting adversarial samples that only optimize for robust attributions within a small neighbourhood of the input, as the classification loss is not maximized for the adversarial search. In contrary, AdvAAT is instantiated with $0 < \gamma < 1$, $0 < \delta < 1$ and no constraint on the predicted class of the adversarial sample, maximizing both the prediction and the attribution loss during adversarial search. Note that with $\gamma \to 0$ and $\delta \to 0$, AdvAAT becomes traditional adversarial training, as introduced in [112].

As in most datasets, target attributions are not given, we choose the attributions of the unperturbed inputs as targets. Moreover, we use the loss derived from the Pearson Correlation Coefficient (PCC) [101], PCL = $1 - \frac{\text{PCC}+1}{2}$, as attribution distance. PCL is a good proxy for optimizing for discrete rank correlations like Kendall's Correlation and Top-K Intersections, as these cannot be used directly due to their non-differentiable nature. Lastly, we utilize the ℓ_{p}-ball of size ε around the original input as neighbourhood function This leads to the following optimization regularization term AAT (5.11) and loss (5.12) AdvAAT respectively.

$$\theta^*_{\text{AAT}} = \arg\min_{\theta} \sum_{\boldsymbol{x}\in\mathbb{S}} \left\{ l_c(\boldsymbol{x}, F, \hat{l}) + \delta \cdot \max_{\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|_{\text{p}}<\varepsilon} \text{PCL}\big[A(\tilde{\boldsymbol{x}}, F, \hat{l}),\ A(\boldsymbol{x}, F, \hat{l})\big] \right\} \tag{5.11}$$

$$\theta^*_{\text{AdvAAT}} = \arg\min_{\theta} \sum_{\boldsymbol{x}\in\mathbb{S}} \max_{\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|_p<\varepsilon} \left\{ l_c(\tilde{\boldsymbol{x}}, F, \hat{l}) + \delta \cdot \text{PCL}\big[A(\tilde{\boldsymbol{x}}, F, \hat{l}),\ A(\boldsymbol{x}, F, \hat{l})\big] \right\} \tag{5.12}$$

where l_c denotes the standard cross-entropy classification loss. Note that we omit the $(1-\delta)$ from the original formulation of FAR for simplicity. The inner maximizations are solved with the aforementioned adversarial IFIA (Algorithm 4), while the outer minimizations with stochastic gradient descent.

During training with the robust objectives, we approximate the second derivative of ReLU network activations with the second derivative of the SoftPlus activation $\nabla^2_{\mathbf{x}}\text{ReLU}(\mathbf{x}) = \beta \cdot \text{sigmoid}(\beta\cdot\mathbf{x}) \cdot \big[\,1 - \text{sigmoid}(\beta\cdot\mathbf{x})\,\big]$, where $\text{sigmoid}(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}}}$ and β controls the approximation tightness of the ReLU [76]. As such, we decouple the attribution maps from the actual estimation of their robustness.

5.4.3 Experimental Setup

We compare three state-of-the-art robust attribution methods – Madry's Robust Prediction adversarial training [112], IG-SUM-NORM from Axiomatic Attribution Regularization [80] and the input-gradient Spatial Alignment regularization[29] (Adv, IG-SN and Align) – and a vanilla, naturally trained (Van.) models' attributional robustness to networks trained with our robust training objectives AAT and AdvAAT, on the datasets MNIST [128], Fashion-MNIST [129], CIFAR-10 [130], GTSRB [131] and Restricted Imagenet [25]. For MNIST and Fashion-MNIST, we train a two-layer convolutional neural network, for the other datasets we use a ResNet taken from [132]. In order to evaluate the attributional robustness of each model, the adversarial IFIA attack (Algorithm 4) [1] is used, with the prediction constraint and $\gamma = 1$, utilizing the proposed Top-K Intersection attribution similarity from [1]. We use the IG attribution map and report the Top-K Intersection (*IN*) of original and adversarial attribution map as well as their Kendall Rank Order Correlation (*CO*) as robustness metrics. The natural and adversarial accuracies (NA and AA) of the models are also reported. AA is estimated with the PGD attack from the authors of [112]. A detailed description of the architectures, training and evaluation details can be found in Appendix C. Table 5.4 contains the results of the comparison experiments. The results are run three times with different data splits and random seeds, and the average results are given.

5.4.4 Results

Based on Table 5.4, we make the following conclusions. First, our methods outperform all other state-of-the-art methods on MNIST and Fashion-MNIST. On the datasets CIFAR-10, GTSRB and Restricted Imagenet, our methods perform comparably to state-of-the-art in terms of *IN*, while giving slightly worse results in terms of *CO*. Hence, we conclude that while AAT and AdvAAT do not outperform Align, they give promising results while being more general and wider applicable, as described in Section 5.2. This is backed by the phenomenon that our methods perform significantly better on MNIST and Fashion-MNIST than Align. We argue

DATASET		VAN.	ADV.	ALIGN	ALIGN (S.)	IG-SN	**AAT**	**ADVAAT**
MNIST	NA	0.99	0.99	0.99	0.95	*0.98	0.98	0.99
	AA	0.12	0.93	0.03	0.12	*0.88	0.0	0.77
	IN	0.43	0.52	0.52	0.58	*0.72	0.76	**0.77**
	CO	0.10	0.19	0.40	0.43	*0.31	0.72	**0.73**
FASHION-MNIST	NA	0.92	0.87	0.90	0.85	*0.85	0.90	0.87
	AA	0.11	0.70	0.31	0.20	0.70	0.0	0.41
	IN	0.43	0.71	0.48	0.50	*0.72	0.80	**0.81**
	CO	0.20	0.58	0.60	0.45	*0.67	**0.82**	**0.82**
CIFAR-10	NA	0.90	0.80	*0.90	-	-	0.74	0.72
	AA	0.0	0.44	*0.38	-	-	0.0	0.25
	IN	0.17	0.66	***0.93**	-	-	0.86	0.90
	CO	-0.02	0.66	***0.92**	-	-	0.70	0.71
GTSRB	NA	0.99	0.95	*0.99	-	*0.96	0.96	0.92
	AA	0.15	0.67	*0.85	-	*0.77	0.27	0.66
	IN	0.39	0.72	***0.92**	-	*0.74	0.75	0.84
	CO	0.19	0.64	***0.89**	-	*0.77	0.79	0.80
RESTRICTED IMAGENET	NA	0.89	0.80	0.82	-	-	0.88	0.80
	AA	0.0	0.68	0.68	-	-	0.0	0.62
	IN	0.08	0.81	**0.92**	-	-	0.91	0.90
	CO	0.20	0.78	**0.86**	-	-	0.78	0.79

Table 5.4: Estimated attributional robustness (Top-K Intersection – *IN*, and Kendall's rank order correlation – *CO*) of the models trained naturally (VAN.), adversarially (ADV.), alignment-based (ALIGN), IG-SUM-NORM-based (IG-SN) as well as with our **AAT** and **ADVAAT** objectives. Their natural and adversarial accuracy is given in the NA and AA rows. Numbers indicated with an asterix * are taken from the respective work and not reproduced by us. ALIGN (S.) denotes the *alignment*-based method with input images scaled between -1 and 1.

that this is due to Align depending on the nature of the data. A large proportion of the data are black pixels. Along these dimensions, the alignment from Equation (5.6) is inherently zero, independently of the gradients. Therefore, Align does not provide an optimization target along these dimensions. Moreover, white pixels are targeted to have large gradients (in alignment terms), but gradient saturation leads to small gradients for these pixels, further worsening optimization with Align on the two MNIST datasets. Our methods do not suffer from these shortcomings, as they provide optimization targets for each input dimension, independently of their values. We also evaluated Align on the MNIST datasets with an input scaling between [-1, 1] as well, indicated as Align (s.) in Table 5.4. However, we see almost no improvement in terms of *IN* and *CO* compared to scaling between [0, 1] (Align). We believe that this is due to the arbitrary choice of input bounds. A lower bound of -1 encourages negative gradients, another arbitrary valid lower bound of 0.2 would encourage positive ones in the same dimensions. This highlights the flaws of the alignment-based method even more, namely that targets are not input shift invariant.

Our second conclusion comes from comparing our AAT method to AdvAAT. AAT achieves slightly worse attribution robustness than AdvAAT, but significantly worse adversarial accuracy

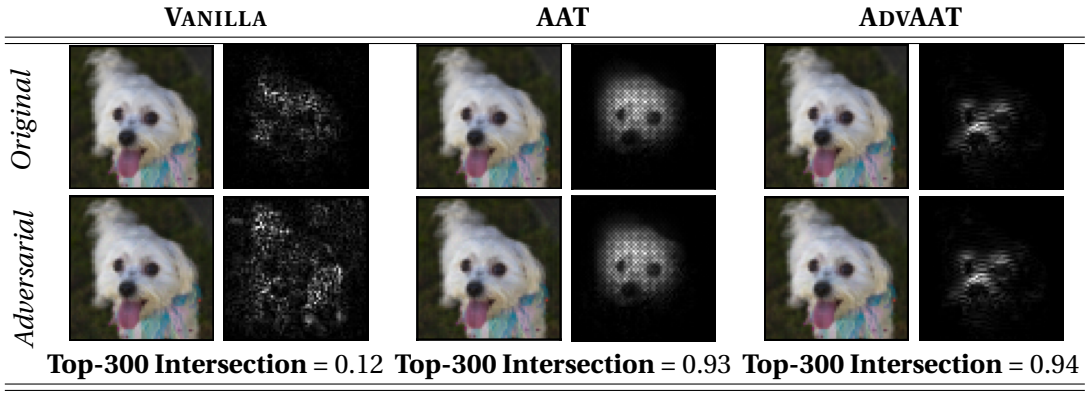


Figure 5.2: Original and adversarial Integrated Gradients (IG) of the natural model (Vanilla – left), our AAT (middle) and AdvAAT (right) method on Restricted Imagenet. For each model, the upper row contains the unperturbed image in the left column and its IG attribution map in the right column. The lower row contains the corresponding perturbed image on the left and the resulting adversarial IG attribution map on the right. Our methods yield less noisy and more robust attribution maps (measured by the *Top-300 Intersection* of the highest attributed pixels of the unperturbed and perturbed image), while correctly classifying all images.

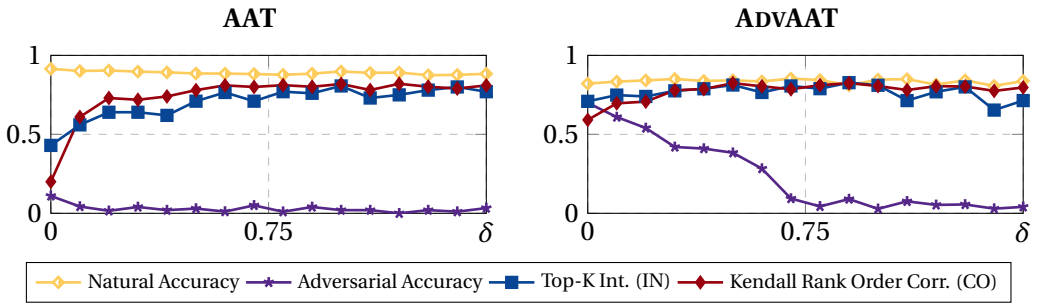


Figure 5.3: Estimated attributional robustness and adversarial accuracies for our AAT (left) and AdvAAT (right) methods, evaluated on Fashion-MNIST varying the regularization parameter δ.

for all datasets experimented on. This leads us to believe that while adversarial robustness does increase attributional robustness, the reverse is only limitedly true. We leave the theoretical analysis of this phenomenon to future work. Figure 5.2 contains an example of attribution maps of our naturally trained, AAT-trained and AdvAAT-trained networks.

5.4.5 Dependency on Regularization Parameter

We examine the influence of the regularization parameter δ on the robustness of attributions and predictions for our CNN trained on Fashion-MNIST. We chose this dataset as it is slightly more complex than MNIST, yet the computational cost of training is low. We train our AAT and AdvAAT models with δ-values varying from 0 to 1.5. Figure 5.3 contains the natural (NA) and adversarial (AA) accuracies as well as the attribution robustness metrics (*IN* and *CO*) for the AAT models to the left and AdvAAT models to the right. We observe that for both methods, higher δ values result in increased AR, with saturation occurring at values above 1. Moreover, for AdvAAT, the adversarial accuracy drops with increasing δ, while AR metrics

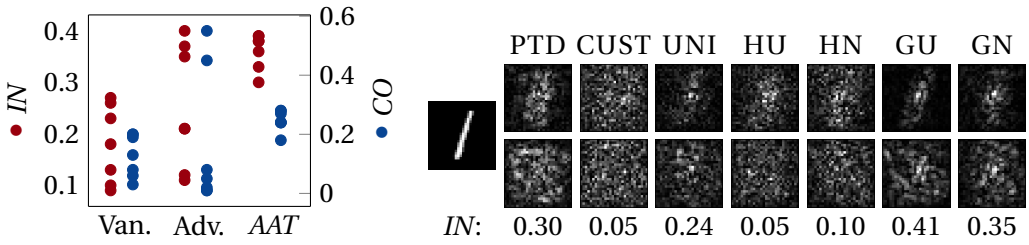


Figure 5.4: Estimated attributional robustness (*IN* and *CO*) of the S attribution method with seven different initializations for the natural (Van.), adversarially (Adv.) and AAT-trained models on MNIST (left). Gradient maps (S) and their attacked maps of the natural model trained on MNIST for different weight initializations (right).

increase, controlling the trade-off between adversarial and attribution robustness.

5.4.6 Dependency on Network Parameter Initialization

Our experiments have shown that gradient-based attribution maps and their robustness estimates can depend on the initialization of the weights in the network. While resulting in nearly identical natural and adversarial accuracies, differently initialized networks yield considerably different robustness of gradient maps. We exemplify this with our natural, adversarially and AAT trained models on MNIST, by reporting their corresponding performance and attribution robustness estimates for seven different network weight initializations. These are the default PyTorch [98] initialization (PTD), a custom initialization taken from [80] (CUST), a random uniform initialization of weights (UNI) as well as the default PyTorch He [133] and Glorot [134] uniform and normal (HU, HN, GU and GN) initializations (as listed in Figure 5.4 from left to right). Figure 5.4 also reports the resulting attributional robustness estimates for the initialization methods. Both for natural and adversarially robust models, the variance of *IN* and *CO* is significant across the different initializations. We expect this behaviour, as heuristic search algorithms like SGD depends strongly on initial conditions. The gradient maps look notably different as well, as reported in Figure 5.4 on the right. This dependency is partly mitigated by our AAT method, but still present.

5.4.7 Dependency on the Tightness Parameter of the ReLU Approximation

Figure 5.5 shows the estimated Top-K Intersection (*IN*) of the vanilla MNIST model while using different β values for the second gradient approximation. We observe that by varying this parameter, the Top-K Intersection changes considerably. We further observe that by setting β too extreme, second gradients vanish, resulting in the IFIA attack not being able to find good adversarial inputs. Previous work [76] has already shown the dependency of AR on β, however, we are the first to only use this approximation for the second order gradients. Therefore, we keep attribution maps unchanged, giving a better estimate for the true attribution robustness

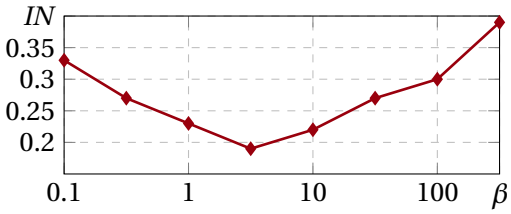


Figure 5.5: Estimated attributional robustness (*IN*) of the S attribution method for the natural model trained on MNIST, varying the β parameter of the ReLU approximation.

of ReLU networks.

5.5 Conclusion

This chapter introduced a generalized notion of attributional robustness with FAR providing objectives for increasing the robustness of explanations in DNNs. This allows for direct optimization for robust attributions, with optionally coupling it to robust predictions. We showed how current existing objectives to enhance AR in vision can be instantiated from this framework. Our results on three biomedical text classification datasets show that classifiers trained without robust objectives lack robustness to small input perturbations in this domain as well. Moreover, our findings show that FAR outperforms the current baseline method, adversarial training, in terms of robustness of attributions and gives a new state-of-the-art benchmark of AR in text classification.

Most current methods that train robust attributions were introduced in the continuous input image modality. Therefore, we showed that FAR can be used to train attributionally robust networks in this modality as well. This allows us to compare the efficacy of FAR to other state-of-the-art methods. To this end, we provided novel instantiations of FAR, AAT and AdvAAT, which directly optimize for high correlation of attributions as well as robust predictions for similar inputs. They perform comparably to or better than current state-of-the-art methods in terms of AR, utilizing fewer assumptions and generalizing better. We believe that borrowing current methods from the image domain, like curvature regularization [27] or other white-box methods could potentially lead to improved robustness in text classifiers while being significantly faster to train than FAR. Finally, we identified parameter dependencies of robust attributions that necessitate careful assessment of methods on their dependencies on these parameters.

We believe that FAR, along with the novel AR estimators in text paves the way to effectively utilize attribution maps in classifiers with sound and faithful explanations. While explainability and faithfulness are still open research questions, our work is a fundamental step towards achieving well-grounded explanations in deep neural networks.

6 Conclusion

6.1 Summary

In this thesis, we thoroughly investigated the robustness of attribution methods in text classification problems. We analyzed the robustness of attribution maps under the presence of imperceptible input perturbations. Our mathematical formulation of attribution robustness allowed us to quantify the robustness of attribution methods in text classifiers in light of plausible perturbations. We then used this mathematical framework to develop robust training objectives that enhance the robustness of explanations in text classifiers.

First, in Chapter 3, we developed a novel baseline method to probe the robustness of attribution methods in general text classifiers. With our word substitution-based, two step black-box attack TEF, we could significantly alter the explanations of our text classifiers without changing the prediction outcomes. By extensive experiments on multiple classification architectures, attribution methods and datasets, we showed that all attributions and classifiers we experimented on lack robustness to adversarial input perturbations. We then showed that such perturbations transfer across model architectures and even attribution methods, a similar trend to what has been shown in traditional adversarial settings [58]. Lastly, we took a step towards extracting universal perturbations that are able to attack heat maps with no knowledge of the model or attribution methods at attack time. These perturbations were able to successfully alter attributions, but we found TEF to yield tighter AR bounds.

In Chapter 4, we established a novel definition of attribution robustness in text that crucially reflects both attribution change and input perturbation size. This allows for comparison of the robustness of attributions in DNN classifiers, while maintaining plausible inputs and keeping the perturbation perceptibility low. Equipped with this notion, we introduced a novel AR estimator that outperforms the previous baseline. It makes use of pretrained language models to extract candidate substitutions. These are an attractive alternative to synonym embeddings which previous methods rely on, as they are increasingly available for many use cases and can be trained in an unsupervised fashion. Utilizing these MLMs, the novel attack computes input perturbations that alter explanations more while being less perceptible.

While having accurate estimators for the robustness of attributions in text classifiers is critical, it is equally important to mitigate the fragility of explanations and develop methods that train networks with robust attributions. Current methods to train robust attributions have thus far only been introduced in the image modality and assume end-to-end differentiable architectures from input to outputs. Thus, in Chapter 5, we introduced a general framework to train robust attributions that has minimal assumptions in terms of the input data and attribution methods at hand. We then showed how to use this framework to successfully train robust networks in text classification, achieving state-of-the-art performance in terms of attribution robustness. We also show that both our previously introduced AR estimators as well as the robust training objective are adaptable to different domains, such as biomedical or healthcare text. Lastly, we used FAR to train attributionally robust image classifiers and compared its performance to current state-of-the-art methods in this domain. We found that FAR achieves comparable AR in DNN image classifiers while having fewer assumptions. Overall, FAR is an attractive framework to train robust networks both in text and other input spaces, taking a first step towards enabling the deployment of state-of-the-art DNN architectures in a wide set of applications.

6.2 Future Directions

The field of faithful explanations and explanation robustness especially is fundamental to deploy DNNs in real-life scenarios. In this work, we marginalized the robustness of attributions and studied the impact of local perturbations on these explanation methods. Therefore, our main assumption throughout this thesis was that these attributions fulfill other desiderata, such as completeness, soundness or parsimony. It would be a very promising research direction to understand how robust explanations influence these other aspects. Arguably, more robust explanations are invariant towards factors in the input that are not relevant in the decision process, thus focus on features that truly contribute to and matter in the inference process, which could indicate sounder and more complete explanations.

It has been shown that one major cause of fragile gradient-based attributions is the large curvature of the decision boundary in DNNs when trained naively. Many methods that train robust attributions rely on second-order gradient information, which consequently is also irregular in this case. Investigating methods that regularize the curvature of gradients in the optimization space could potentially lead to further robustness improvement of interpretation methods. It would be interesting to understand how this influences other, non-gradient-based interpretation methods' robustness. Based on this, future research could lead to guarantees on AR, i.e., giving certified upper bounds on the maximal change of explanations within a small input region of interest. Guaranteed attribution robustness could be especially useful to enable the deployment of critical DNNs in real life.

Lastly, explainable AI methods might help detect certain undesired behaviours in the networks [5], such as racial or gender biases. In this work, we focused on attribution methods, but

together with other methods like Occlusion [11] or feature map visualizations, a promising future direction could be to investigate whether robust attributions better reflect these inherent biases in the networks than their non-robust counterparts. This could potentially enhance the fairness of DNNs, which is especially critical, as Large Language Models (LLMs) such as ChatGPT, Bard or LLaMA are increasingly being utilized in several fields of AI and revolutionizing many. Having explainable and robust LLMs therefore is fundamental to understand the uses and impact of these models.

A Appendix for Chapter 3

A.1 TEF Operation Example

Given classifier F, input sample $\boldsymbol{s}$ = *"a poignant comedy that offers food for thought ."*, original attribution scores $A(s, F, \hat{l})$, find the adversarial sequence of tokens $\boldsymbol{s}_{\text{adv}}$ that minimizes $\text{PCC}\big[A(\boldsymbol{s}, F, \hat{l}), A(\boldsymbol{s}_{\text{adv}}, F, \hat{l})\big]$ such that at most $\rho_{max} = 0.25 = 25\%$ of words are changed, with $\arg\max_{i \in \mathbb{L}} F_i(\boldsymbol{s}, \hat{l}) = \arg\max_{i \in \mathbb{L}} F_i(\boldsymbol{s}_{\text{adv}}, \hat{l})$ = **Positive** and $\boldsymbol{s}_{\text{adv}}$ fulfilling the locality constraints described in Section 3.3, namely each replacement is a synonym of the original word [92], the replacement word needs to have the same Part Of Speech computed by SpaCy [89] and stop words can not be replaced.

i	0	1	2	3	4	5	6	7	8
$\boldsymbol{s}$	a	poignant	**comedy**	that	offers	food	for	thought	.
I_{w_i}	0.0	0.63	0.4	0.05	0.16	0.23	0.03	0.22	0.0

The word importance ranking from Section 3.3 yields *poignant* and *comedy* (in this order) to be the $\lfloor 9 \cdot 0.25 \rfloor = 2$ most important tokens, therefore the candidate replacements for only those are considered. This results in the following two steps of TEF.

1. Step: Replace the most important word *poignant* (at index 1, indicated with yellow background) with its best candidate, measured by lowest PCC. This candidate is the word *distressing*. The table below indicates all possible sentences where *poignant* is replaced with its candidates. The latter two fail the POS and prediction constraints, thus are not taken into account.

i	0	1	2	3	4	5	6	7	8	PCC
$\tilde{\boldsymbol{s}}$	a	heartbreaking	comedy	that	offers	food	for	thought	.	
$A(\tilde{\boldsymbol{s}})$	0.01	-0.21	0.45	0.05	0.18	0.25	0.04	0.22	0.01	0.98
$\tilde{\boldsymbol{s}}$	a	distressing	comedy	that	offers	food	for	thought	.	
$A(\tilde{\boldsymbol{s}})$	-0.09	0.11	0.84	0.09	-1	0.54	-0.1	0.26	0.06	**0.44**
$\tilde{\boldsymbol{s}}$	a	alarm	comedy	that	offers	food	for	thought	.	
$A(\tilde{\boldsymbol{s}})$	Failed POS-Filter									-
$\tilde{\boldsymbol{s}}$	a	agonizing	comedy	that	offers	food	for	thought	.	
$A(\tilde{\boldsymbol{s}})$	Failed Prediction-Filter									-

2. Step: Replace the second-most important word *comedy* with the best valid candidate, in this case the token *comic*. Here, the original token *poignant* is already replaced with the best candidate, *distressing*.

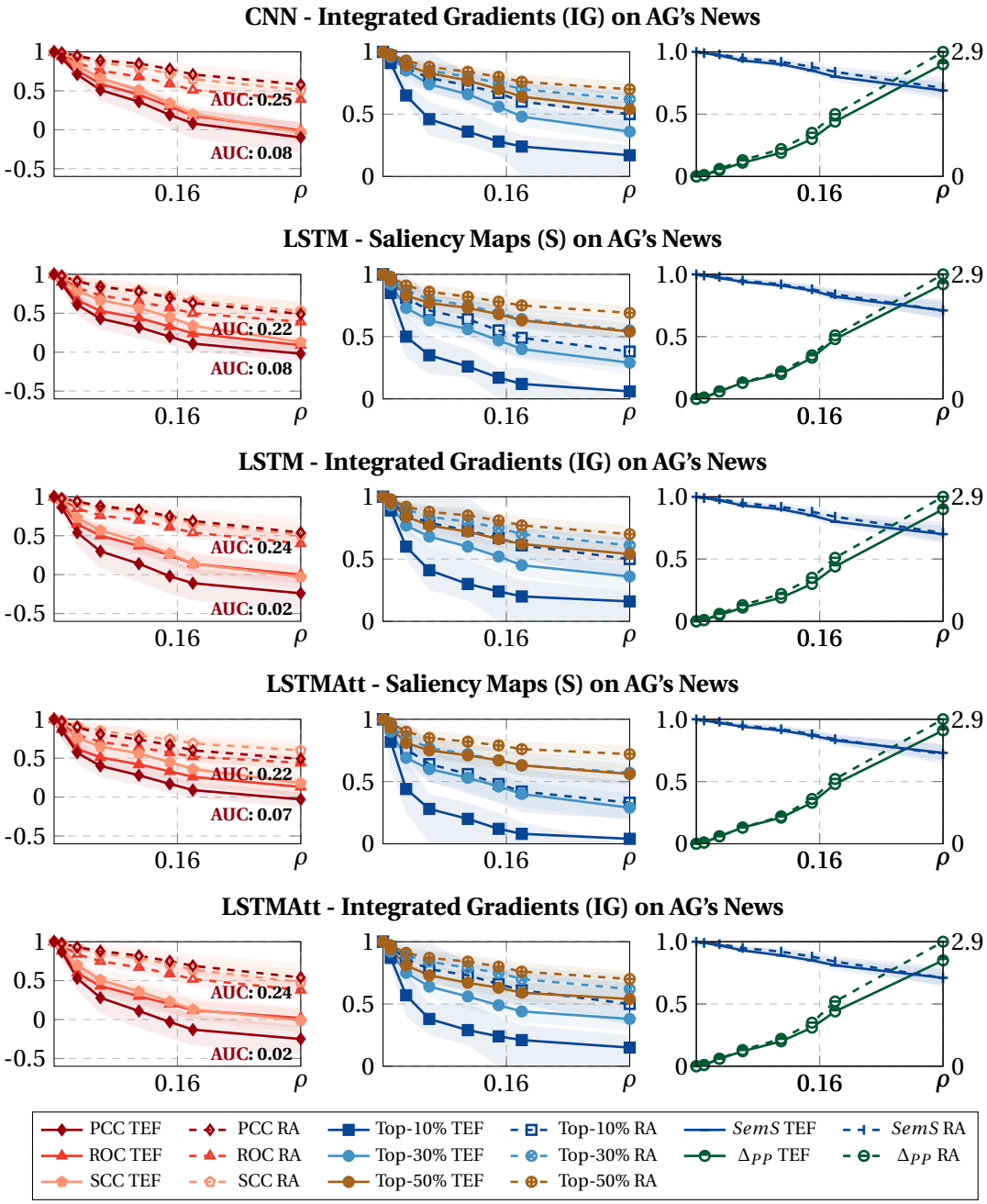
i	0	1	2	3	4	5	6	7	8	PCC
$\tilde{\boldsymbol{s}}$	a	distressing	**humor**	**that**	**offers**	food	for	**thought**	.	
$A(\tilde{\boldsymbol{s}})$	-0.09	0.01	0.34	0.27	0.2	-0.02	0.02	0.27	0.05	0.63
$\tilde{\boldsymbol{s}}$	a	distressing	comic	that	**offers**	food	for	thought	.	
$A(\tilde{\boldsymbol{s}})$	-0.02	0.04	0.05	0.02	-0.57	-0.13	0.01	0.46	0.04	**0.22**
$\tilde{\boldsymbol{s}}$	a	distressing	travesty	that	offers	food	for	thought	.	
$A(\tilde{\boldsymbol{s}})$	Failed Prediction-Filter									-
$\tilde{\boldsymbol{s}}$	a	distressing	humorous	that	offers	food	for	thought	.	
$A(\tilde{\boldsymbol{s}})$	Failed POS-Filter									-

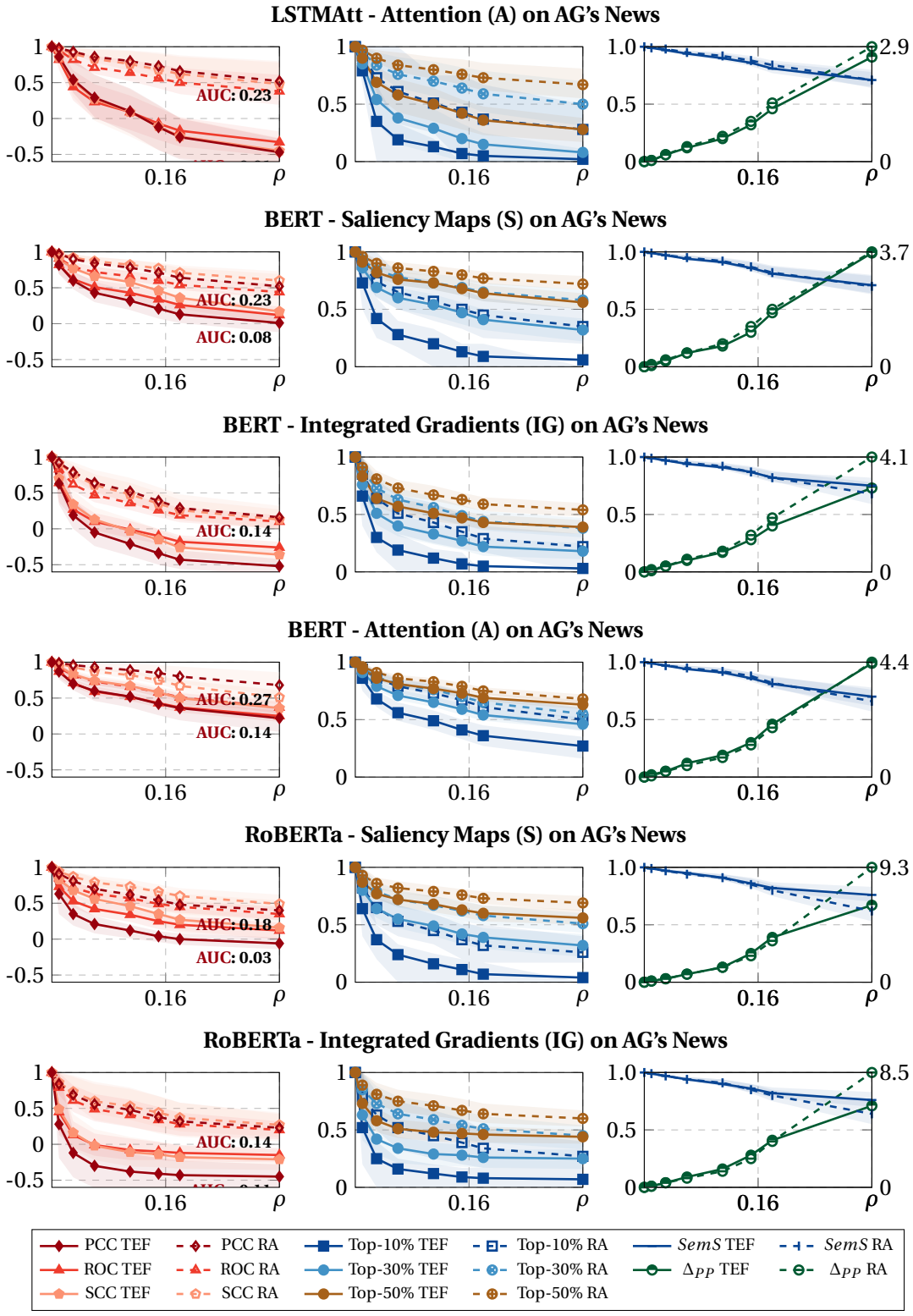
The final adversarial sequence $\boldsymbol{s}_{\text{adv}}$ becomes the valid $\tilde{\boldsymbol{s}}$ with the lowest PCC value, which is given in the following table.

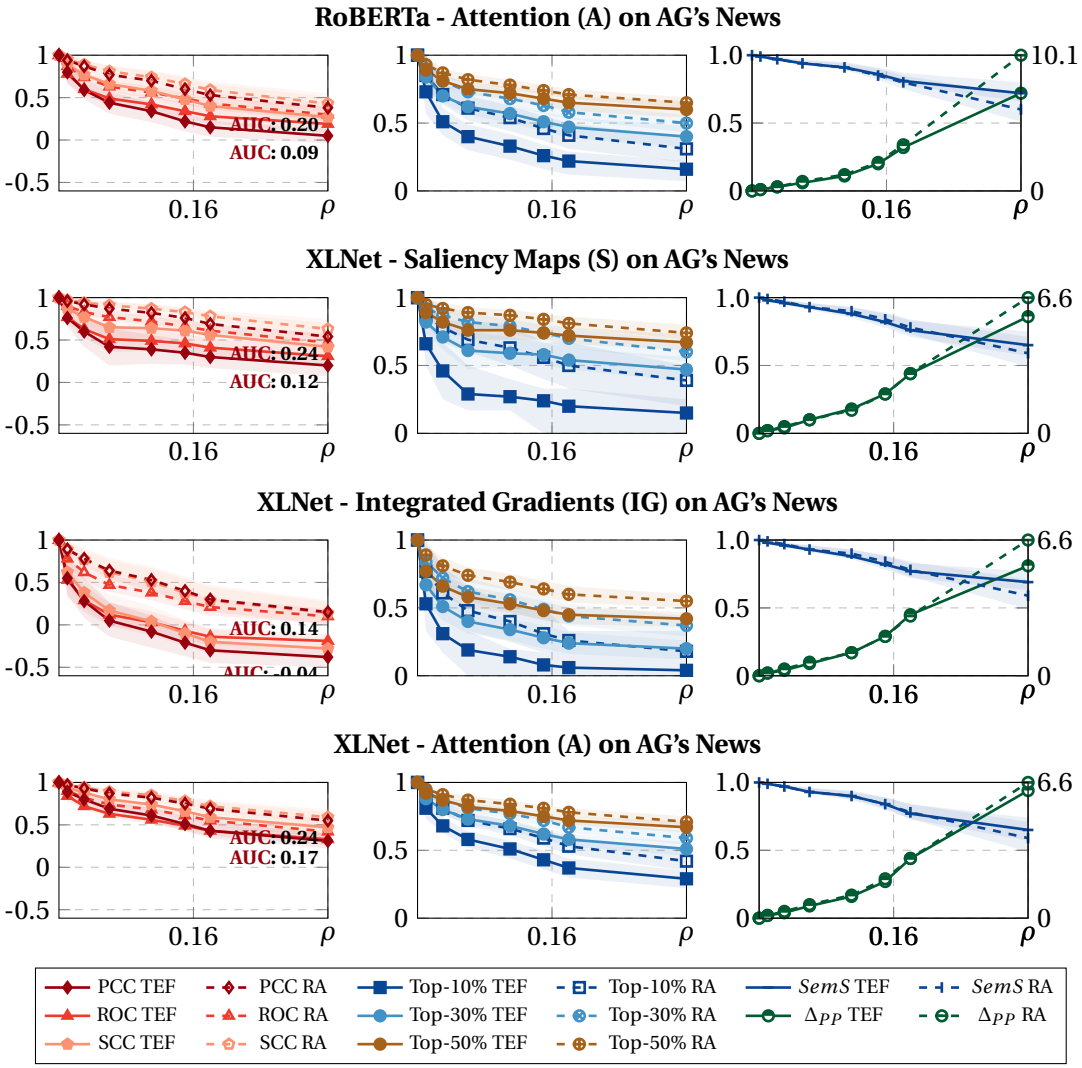
i	0	1	2	3	4	5	6	7	8	PCC
	a	poignant	comedy	that	offers	food	for	thought	.	
$A(\boldsymbol{s})$	0.0	-0.08	0.4	0.05	0.16	0.23	0.03	0.22	0.0	-
$\boldsymbol{s}_{\text{adv}}$	a	distressing	comic	that	**offers**	food	for	thought	.	
$A(\boldsymbol{s}_{\text{adv}})$	-0.02	0.04	0.05	0.02	-0.57	-0.13	0.01	0.46	0.04	**0.22**

A.2 Robustness of Attributions

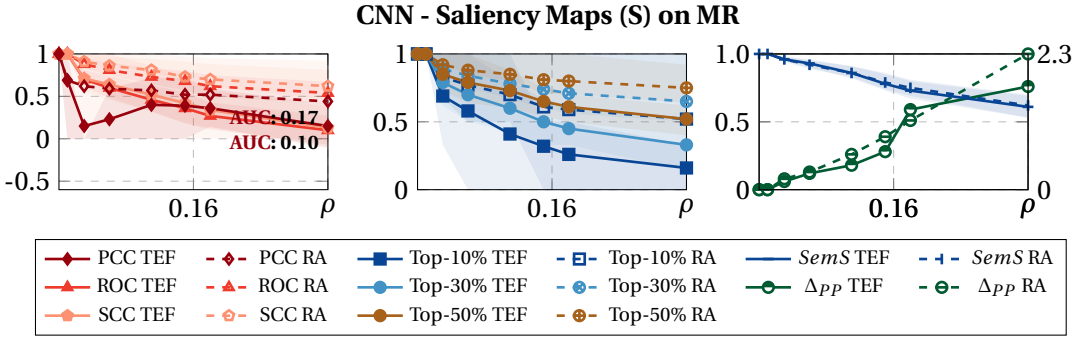
A.2.1 AG's News

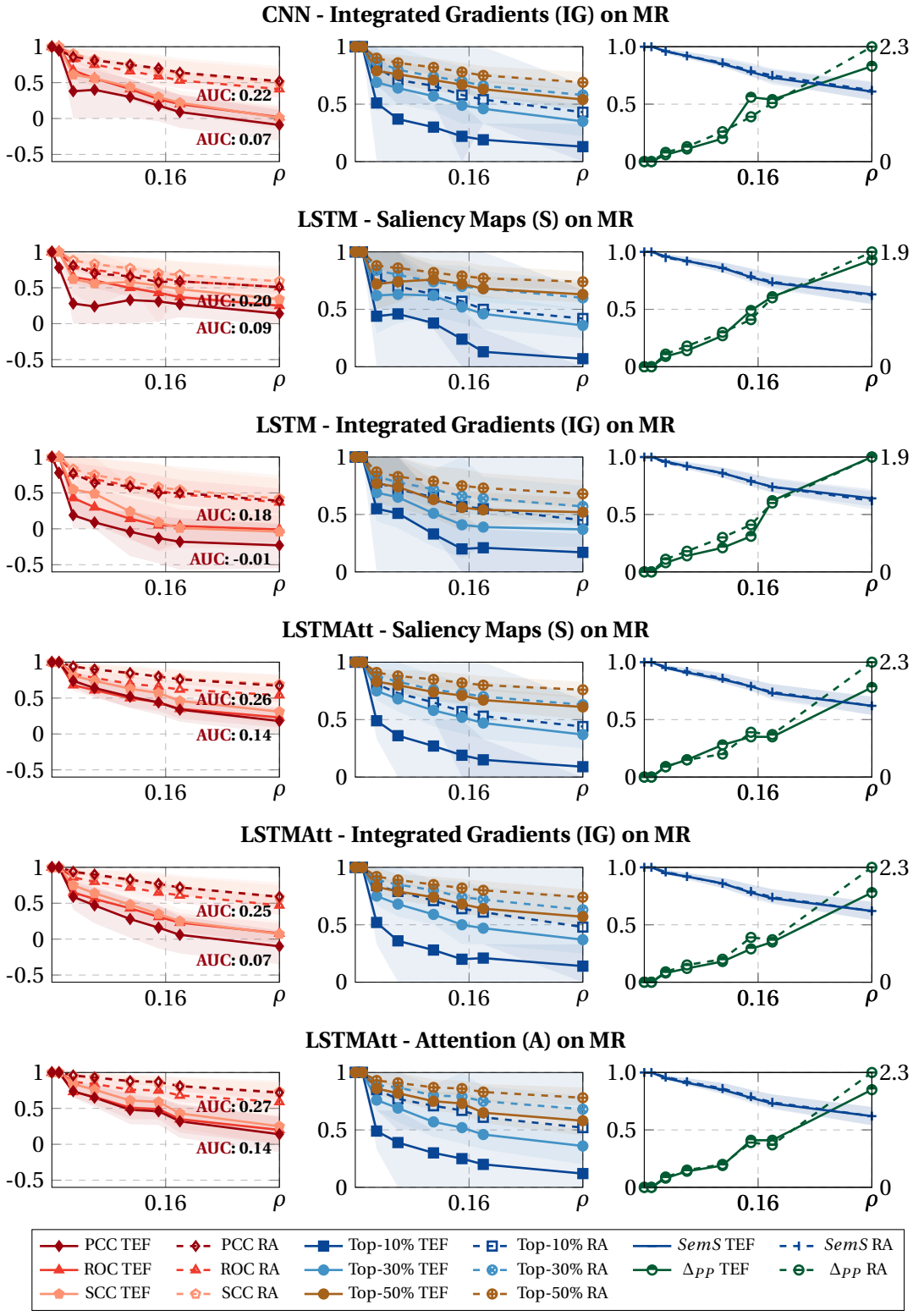


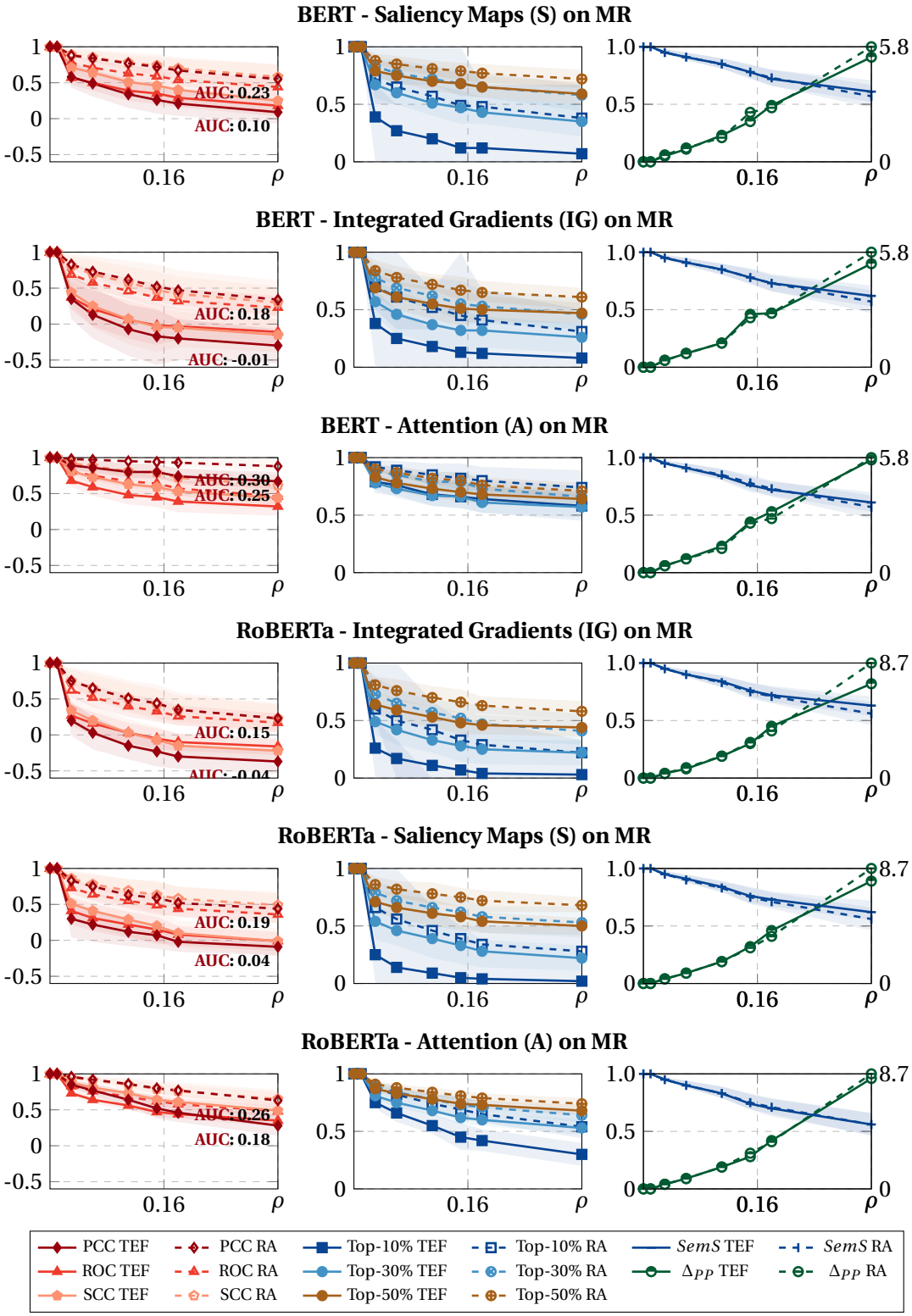

LSTMAtt - Attention (A) on AG's News
AUC: 0.23
2.9
BERT - Saliency Maps (S) on AG's News
AUC: 0.23
AUC: 0.08
3.7
BERT - Integrated Gradients (IG) on AG's News
AUC: 0.14
4.1
BERT - Attention (A) on AG's News
AUC: 0.27
AUC: 0.14
4.4
RoBERTa - Saliency Maps (S) on AG's News
AUC: 0.18
AUC: 0.03
9.3
RoBERTa - Integrated Gradients (IG) on AG's News
AUC: 0.14
8.5
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

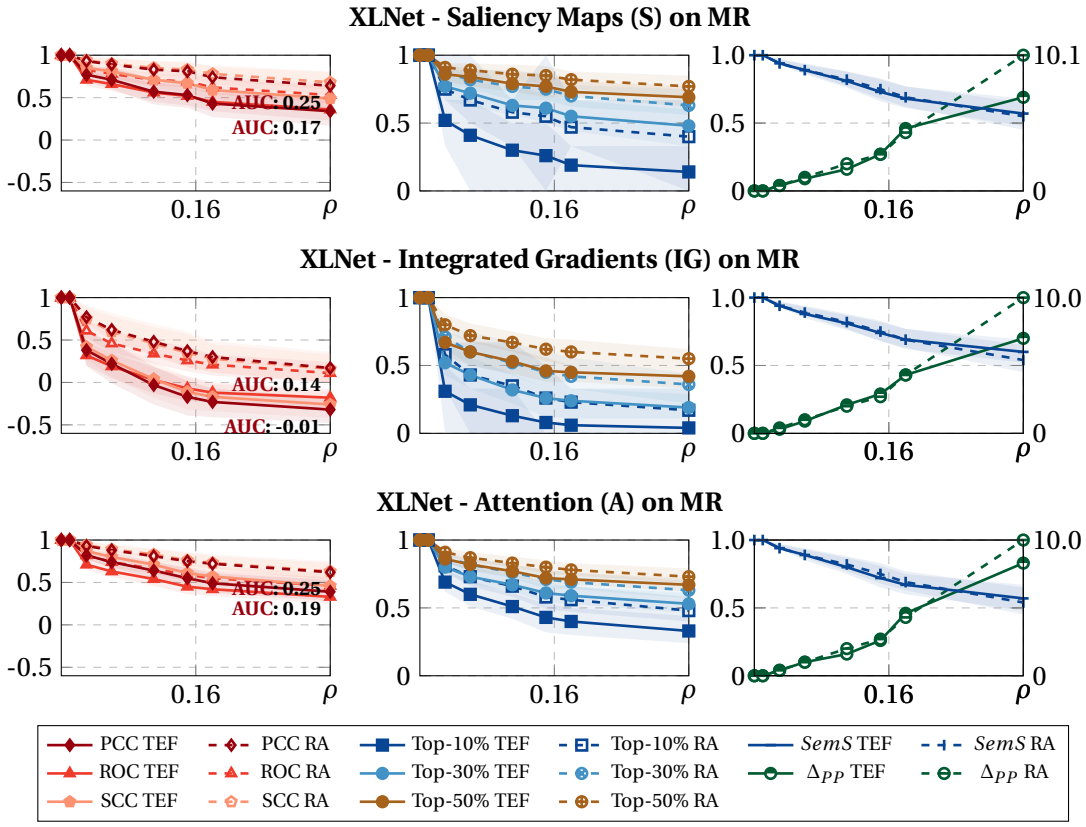

RoBERTa - Attention (A) on AG's News
AUC: 0.20
AUC: 0.09
XLNet - Saliency Maps (S) on AG's News
AUC: 0.24
AUC: 0.12
XLNet - Integrated Gradients (IG) on AG's News
AUC: 0.14
XLNet - Attention (A) on AG's News
AUC: 0.24
AUC: 0.17
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

A.2.2 MR

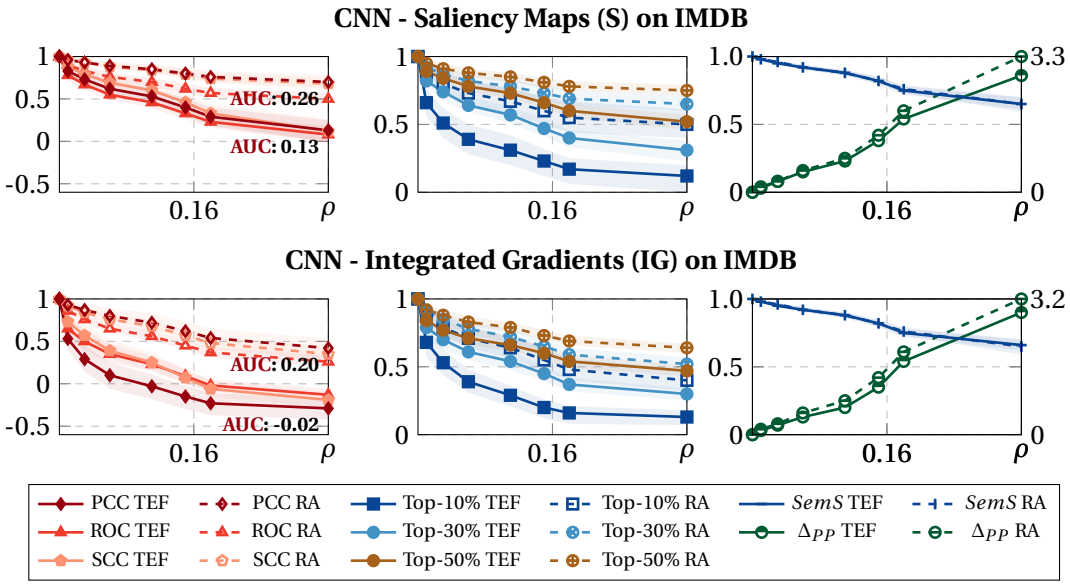

CNN - Saliency Maps (S) on MR
AUC: 0.17
AUC: 0.10
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

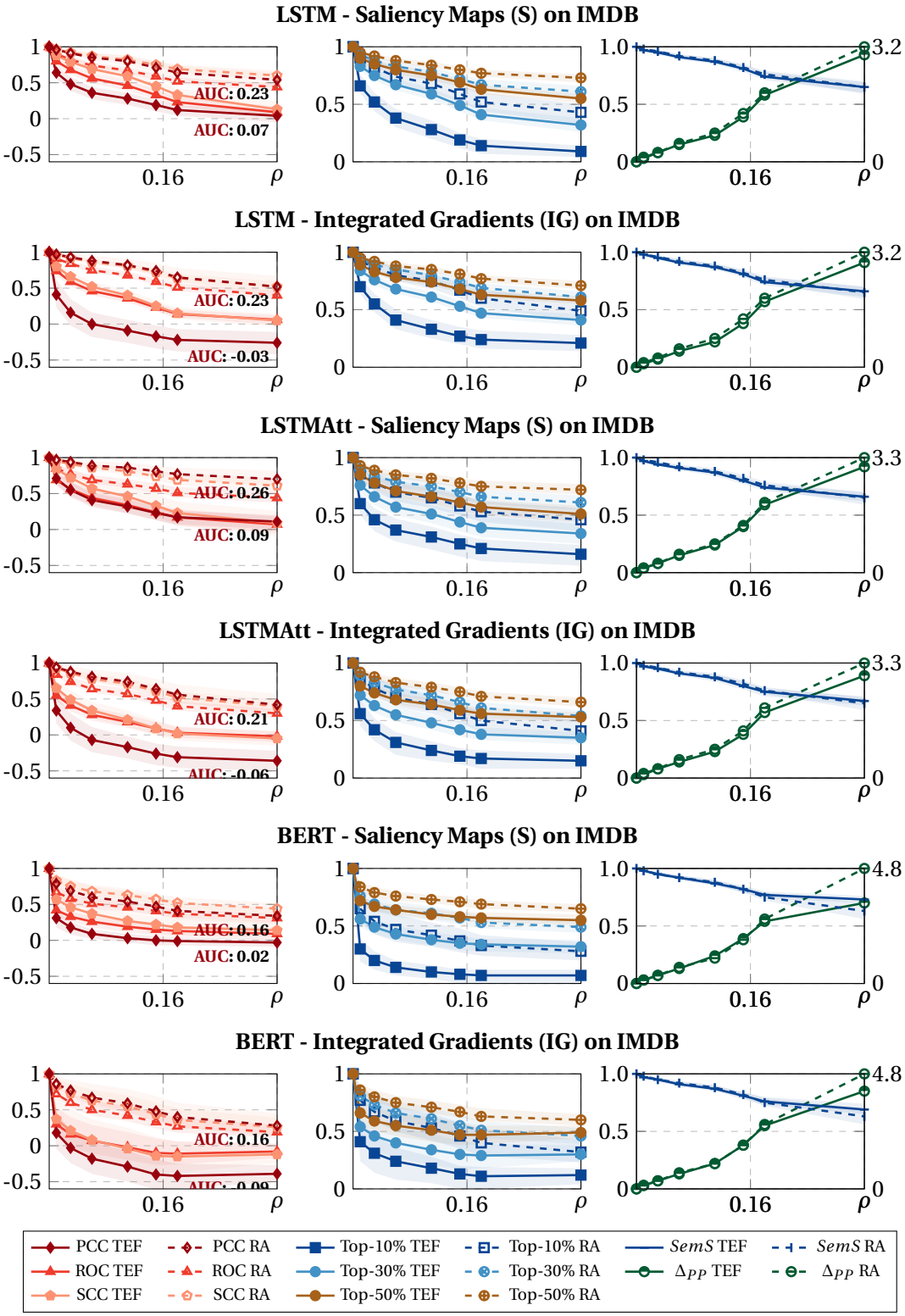

CNN - Integrated Gradients (IG) on MR
AUC: 0.22
AUC: 0.07
LSTM - Saliency Maps (S) on MR
AUC: 0.20
AUC: 0.09
LSTM - Integrated Gradients (IG) on MR
AUC: 0.18
AUC: -0.01
LSTMAtt - Saliency Maps (S) on MR
AUC: 0.26
AUC: 0.14
LSTMAtt - Integrated Gradients (IG) on MR
AUC: 0.25
AUC: 0.07
LSTMAtt - Attention (A) on MR
AUC: 0.27
AUC: 0.14
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

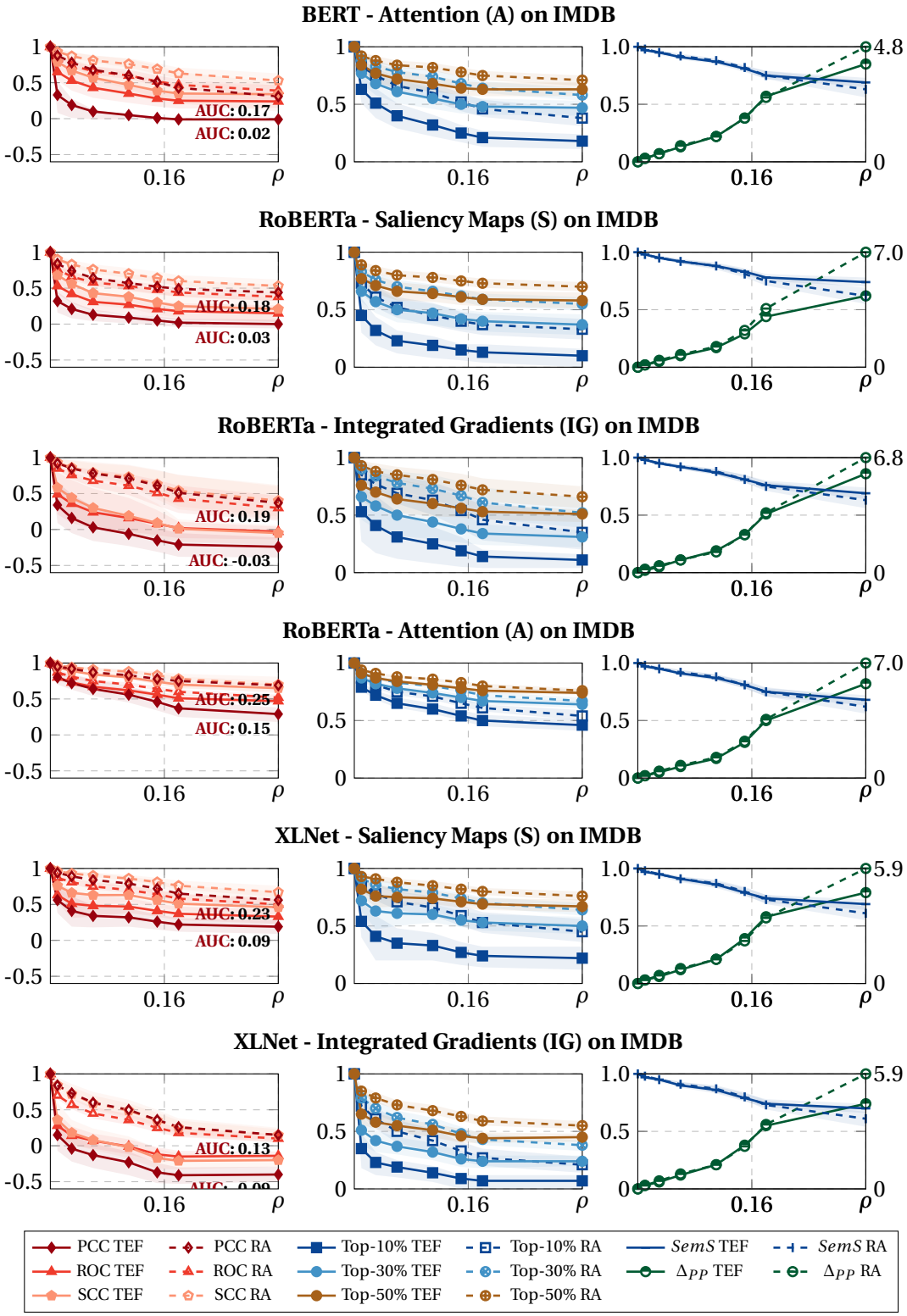

BERT - Saliency Maps (S) on MR
AUC: 0.23
AUC: 0.10
BERT - Integrated Gradients (IG) on MR
AUC: 0.18
AUC: -0.01
BERT - Attention (A) on MR
AUC: 0.30
AUC: 0.25
RoBERTa - Integrated Gradients (IG) on MR
AUC: 0.15
RoBERTa - Saliency Maps (S) on MR
AUC: 0.19
AUC: 0.04
RoBERTa - Attention (A) on MR
AUC: 0.26
AUC: 0.18
0.16
ρ
5.8
8.7
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

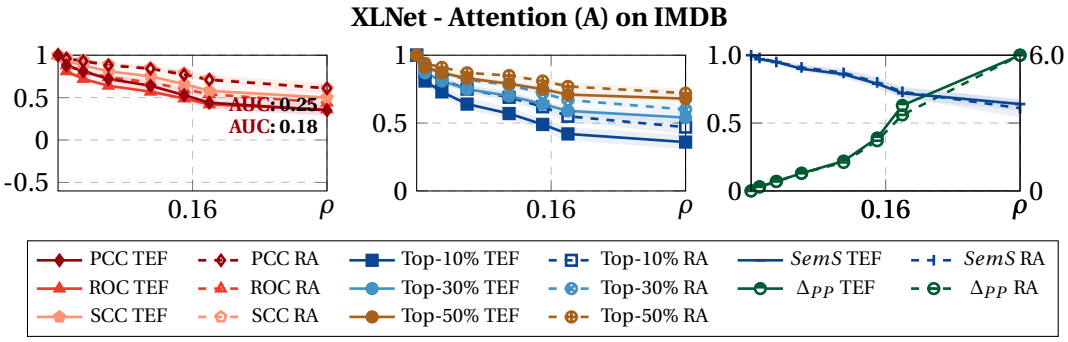

XLNet - Saliency Maps (S) on MR
AUC: 0.25
AUC: 0.17
10.1
XLNet - Integrated Gradients (IG) on MR
AUC: 0.14
AUC: -0.01
10.0
XLNet - Attention (A) on MR
AUC: 0.25
AUC: 0.19
10.0
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

A.2.3 IMDB

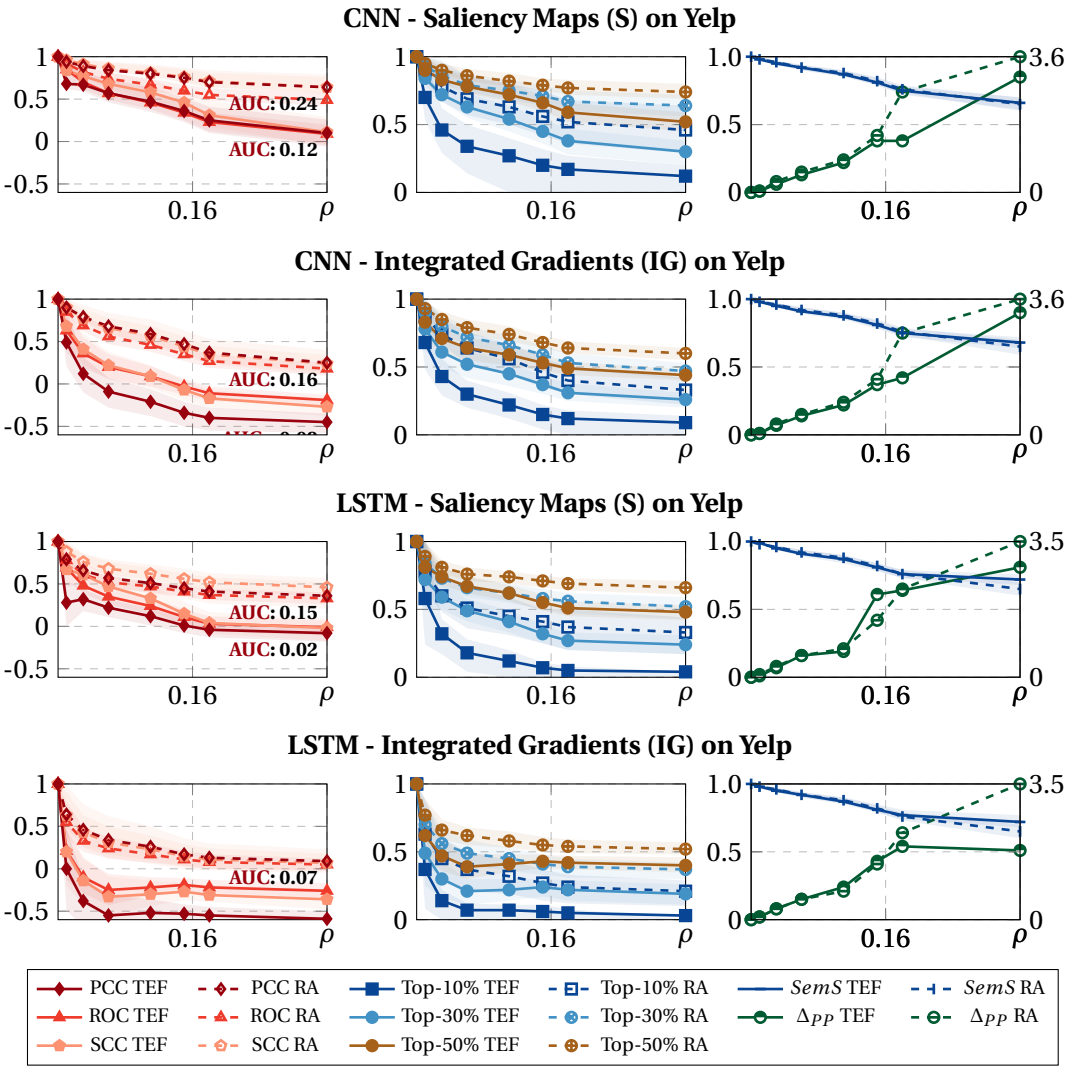

CNN - Saliency Maps (S) on IMDB
AUC: 0.26
AUC: 0.13
3.3
CNN - Integrated Gradients (IG) on IMDB
AUC: 0.20
AUC: -0.02
3.2
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

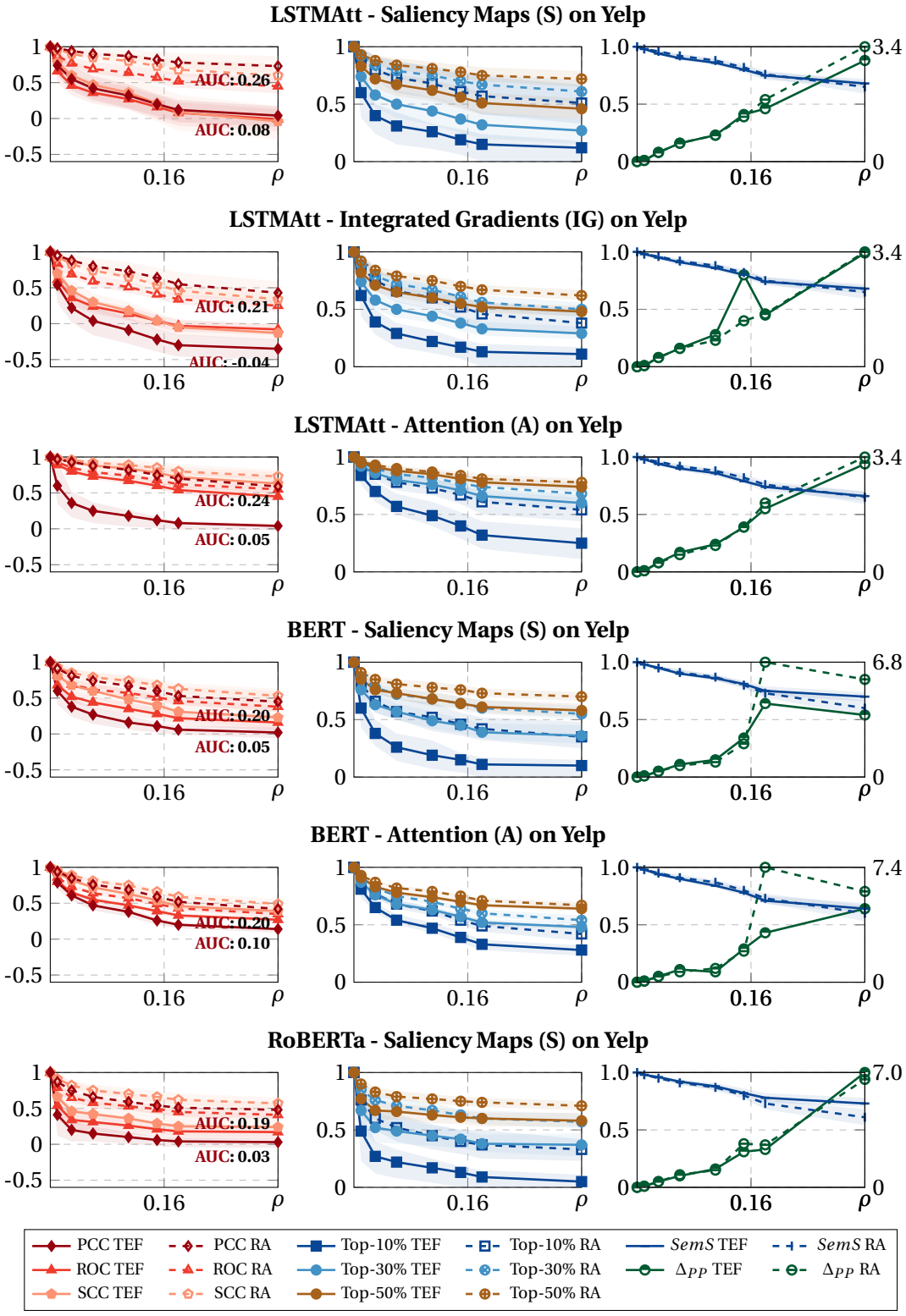

LSTM - Saliency Maps (S) on IMDB
AUC: 0.23
AUC: 0.07
LSTM - Integrated Gradients (IG) on IMDB
AUC: 0.23
AUC: -0.03
LSTMAtt - Saliency Maps (S) on IMDB
AUC: 0.26
AUC: 0.09
LSTMAtt - Integrated Gradients (IG) on IMDB
AUC: 0.21
AUC: -0.06
BERT - Saliency Maps (S) on IMDB
AUC: 0.16
AUC: 0.02
BERT - Integrated Gradients (IG) on IMDB
AUC: 0.16
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

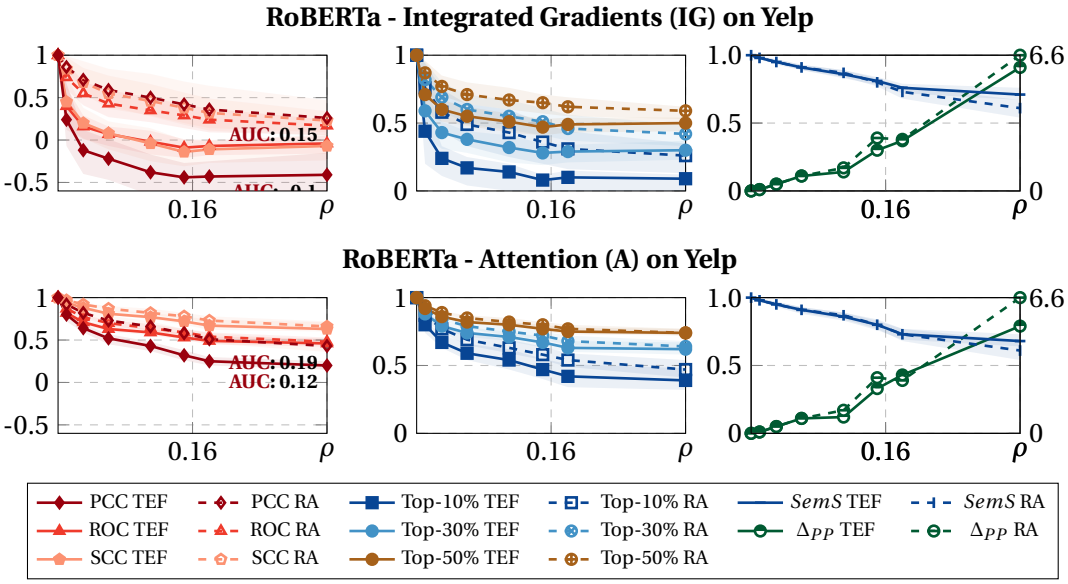

BERT - Attention (A) on IMDB
AUC: 0.17
AUC: 0.02
4.8
RoBERTa - Saliency Maps (S) on IMDB
AUC: 0.18
AUC: 0.03
7.0
RoBERTa - Integrated Gradients (IG) on IMDB
AUC: 0.19
AUC: -0.03
6.8
RoBERTa - Attention (A) on IMDB
AUC: 0.25
AUC: 0.15
7.0
XLNet - Saliency Maps (S) on IMDB
AUC: 0.23
AUC: 0.09
5.9
XLNet - Integrated Gradients (IG) on IMDB
AUC: 0.13
5.9
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA


XLNet - Attention (A) on IMDB
AUC: 0.25
AUC: 0.18
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

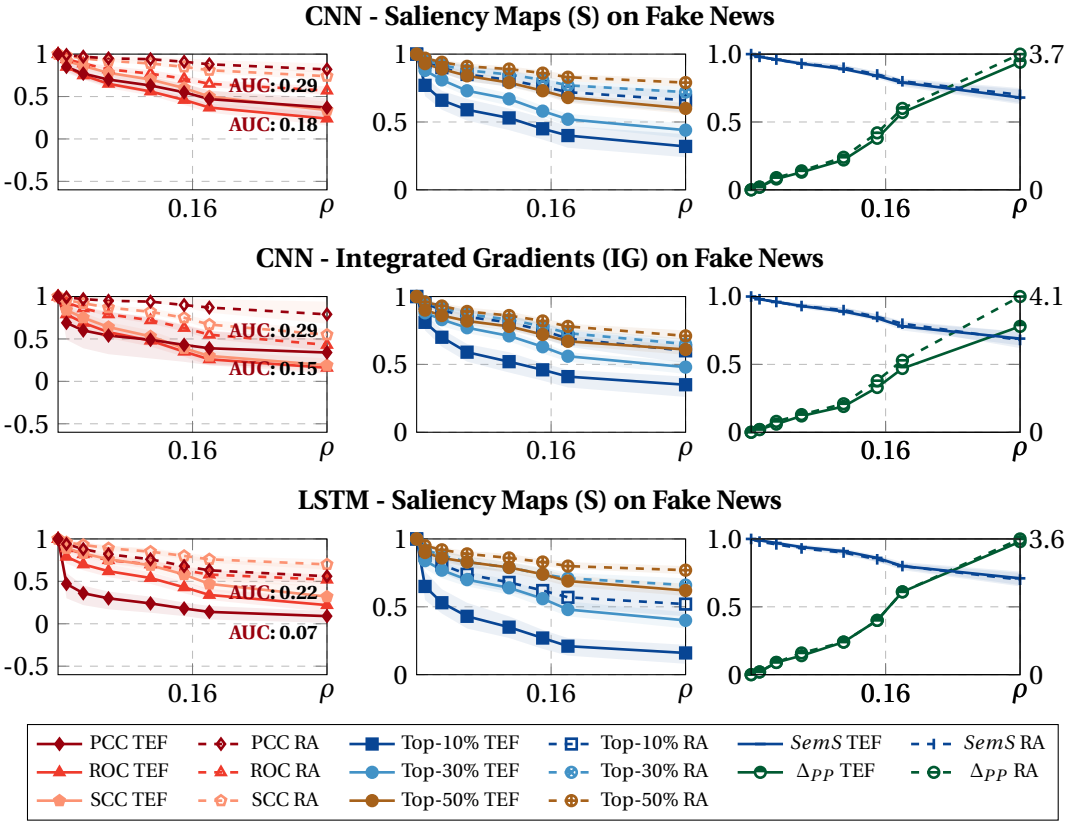
A.2.4 Yelp

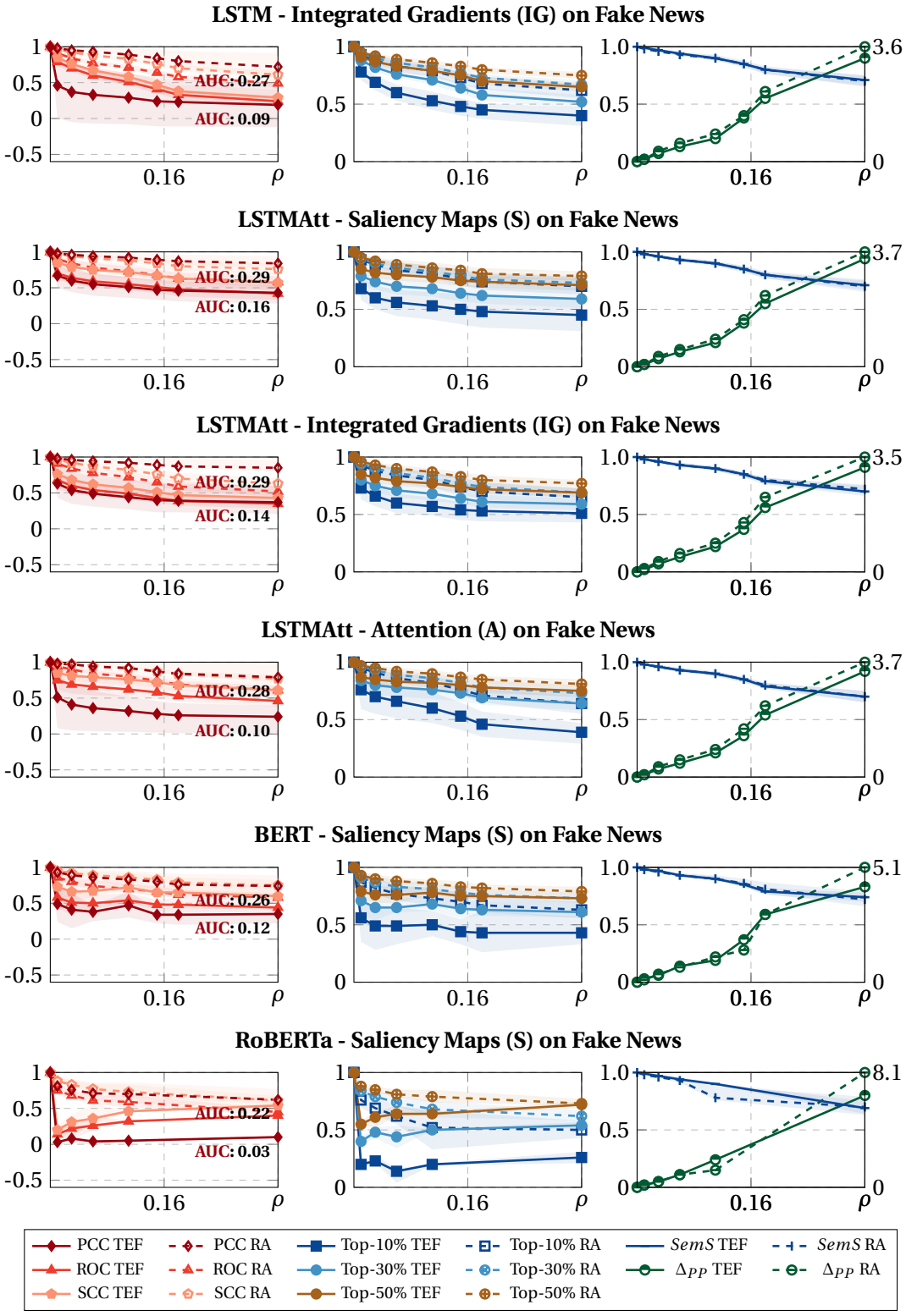

CNN - Saliency Maps (S) on Yelp
AUC: 0.24
AUC: 0.12
CNN - Integrated Gradients (IG) on Yelp
AUC: 0.16
LSTM - Saliency Maps (S) on Yelp
AUC: 0.15
AUC: 0.02
LSTM - Integrated Gradients (IG) on Yelp
AUC: 0.07
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

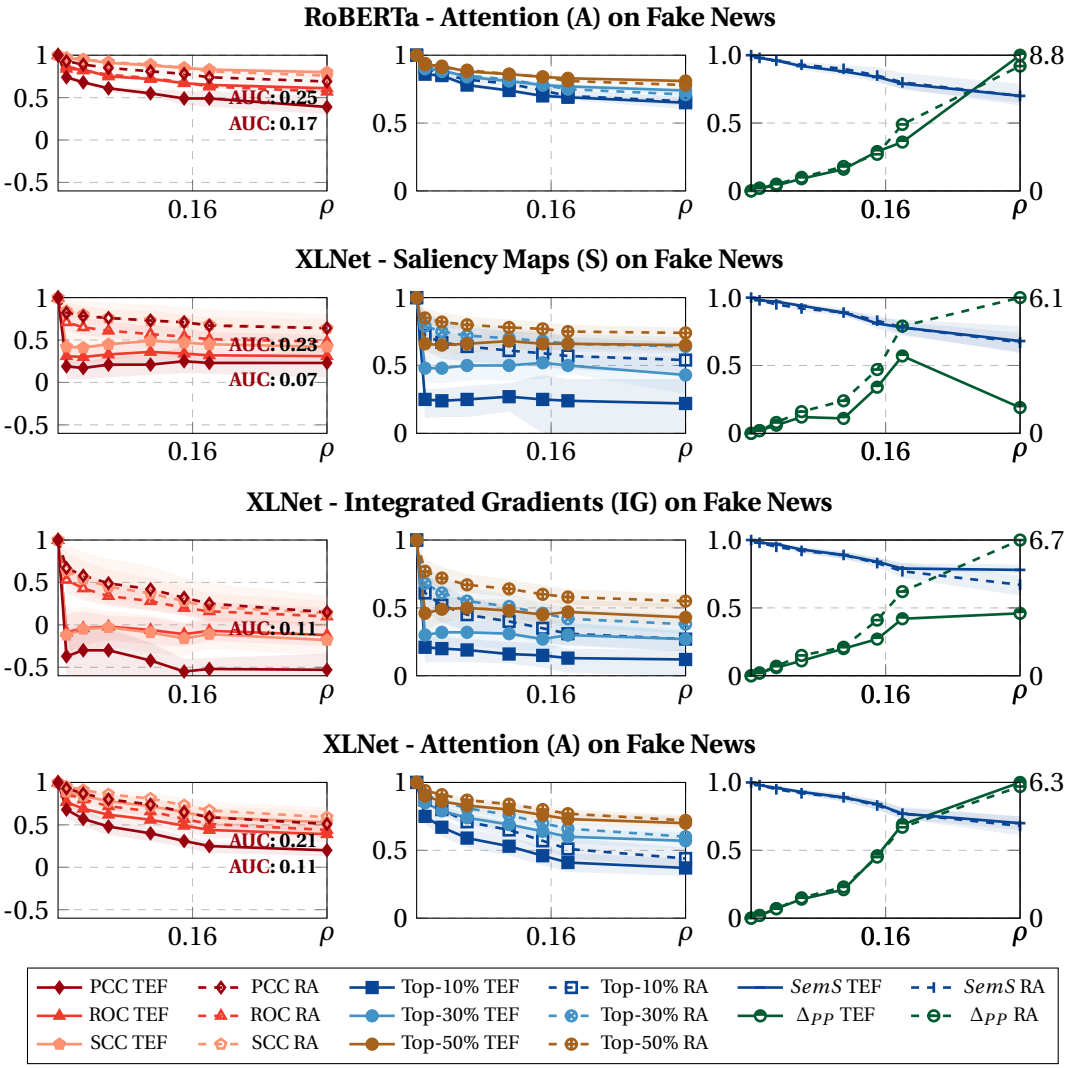

LSTMAtt - Saliency Maps (S) on Yelp
AUC: 0.26
AUC: 0.08
LSTMAtt - Integrated Gradients (IG) on Yelp
AUC: 0.21
AUC: -0.04
LSTMAtt - Attention (A) on Yelp
AUC: 0.24
AUC: 0.05
BERT - Saliency Maps (S) on Yelp
AUC: 0.20
AUC: 0.05
BERT - Attention (A) on Yelp
AUC: 0.20
AUC: 0.10
RoBERTa - Saliency Maps (S) on Yelp
AUC: 0.19
AUC: 0.03
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA



A.2.5 Fake News




LSTM - Integrated Gradients (IG) on Fake News
AUC: 0.27
AUC: 0.09
3.6
LSTMAtt - Saliency Maps (S) on Fake News
AUC: 0.29
AUC: 0.16
3.7
LSTMAtt - Integrated Gradients (IG) on Fake News
AUC: 0.29
AUC: 0.14
3.5
LSTMAtt - Attention (A) on Fake News
AUC: 0.28
AUC: 0.10
3.7
BERT - Saliency Maps (S) on Fake News
AUC: 0.26
AUC: 0.12
5.1
RoBERTa - Saliency Maps (S) on Fake News
AUC: 0.22
AUC: 0.03
8.1
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA


RoBERTa - Attention (A) on Fake News
AUC: 0.25
AUC: 0.17
8.8
XLNet - Saliency Maps (S) on Fake News
AUC: 0.23
AUC: 0.07
6.1
XLNet - Integrated Gradients (IG) on Fake News
AUC: 0.11
6.7
XLNet - Attention (A) on Fake News
AUC: 0.21
AUC: 0.11
6.3
0.16
ρ
PCC TEF
PCC RA
ROC TEF
ROC RA
SCC TEF
SCC RA
Top-10% TEF
Top-10% RA
Top-30% TEF
Top-30% RA
Top-50% TEF
Top-50% RA
SemS TEF
SemS RA
Δpp TEF
Δpp RA

B Appendix for Chapter 4

B.1 Datasets

We estimate the robustness of our attribution methods and models on five publicly available datasets. These are AG's News, MR movie review, IMDB movie review, Yelp and Fake News, all of which are in English. AG's News consists of 127552 news article samples, categorized into the classes World, News, Business and Sci/Tech. We use the concatenation of title and text of the samples to feed into our text classifiers, stripping any sample that is longer than 64 tokens. The MR movie review dataset contains 10592 short samples of positive or negative movie reviews. We only use the first 32 tokens in each sample as input to the classifiers. IMDB movie review is a dataset consisting of 49952 positive and negative movie reviews, with a maximum token length of 256. Yelp categorizes 700000 reviews of several topics into 5 classes, each representing a rating from 1 to 5. We strip the samples to a maximum length of 256. Fake News is a collection of 20080 news samples, each categorized into reliable or unreliable. These are rather long articles, thus we use a maximum sequence length of 512 for this dataset.

We apply basic preprocessing to all samples in each dataset, which includes converting them to lowercase, removing any special characters not in the English alphabet and emojis. We use 60% of the samples for training the classifier models, 20% for validation and 20% for testing and estimating the robustness of attribution methods.

B.2 Models

As described in Section 4.5.1, we train six classification architectures for each dataset, three DNN-based architectures, which are a CNN, an LSTM, an LSTM with an attention layer (LSTMAtt), as well as three transformer-based architectures, which are a finetuned BERT, RoBERTa and XLNet. The CNN, LSTM and LSTMAtt architectures use the 6B-300-dimensional Glove word embeddings, while the transformer-based architectures use the pretrained Hugging Face embeddings of the respective base-uncased versions. The DNN-based classifiers each contain a linear layer on top of their feature extractors and use the built-in SpaCy English tokenizer,

		AG's News	MR	IMDB	Yelp	Fake News
CNN	Input shape	(64, 300)	(32,300)	(256, 300)	(256, 300)	(512, 300)
	Num. classes	4	2	2	5	2
	Filter sizes	[3, 5, 7]	[3, 5]	[3, 5, 7]	[3, 5, 7]	[3, 5, 7]
	Feature sizes	[8, 8, 8]	[8, 8]	[16, 16, 16]	[128, 128, 128]	[32, 32, 32]
	Pooling sizes	[2, 2, 2]	[2, 2]	[2, 2, 2]	[2, 2, 2]	[2, 2, 2]
	Lin. layer dim.	8	8	16	64	32
	Num. params	67748	27946	567458	16428293	4091714
LSTM	Input shape	(64, 300)	(32,300)	(256, 300)	(256, 300)	(512, 300)
	Num. classes	4	2	2	5	2
	Hidden dim.	8	8	16	256	16
	Num. layers	1	1	2	2	1
	Pooling sizes	2	2	1	2	2
	Lin. layer dim.	8	8	16	32	16
	Num. params	10988	10458	18162	2146693	85986
LSTMAtt	Input shape	(64, 300)	(32,300)	(256, 300)	(256, 300)	(512, 300)
	Num. classes	4	2	2	5	2
	Hidden dim.	8	8	16	256	16
	Num. layers	4	1	2	2	1
	Lin. layer dim.	8	8	16	32	16
	Num. params	25004	19994	47666	2752901	41826
BERT	Input shape	(64,)	(32,)	(256,)	(256,)	(512,)
	Num. classes	4	2	2	5	2
	Model ID	bert-base-uncased	bert-base-uncased	bert-base-uncased	bert-base-uncased	bert-base-uncased
	Num. params	109485316	109483778	109483778	109486085	109483778
RoBERTa	Input shape	(64,)	(32,)	(256,)	(256,)	(512,)
	Num. classes	4	2	2	5	2
	Model ID	roberta-base	roberta-base	roberta-base	roberta-base	roberta-base
	Num. params	124648708	124647170	124647170	124649477	124647170
XLNet	Input shape	(64,)	(32,)	(256,)	(256,)	(512,)
	Num. classes	4	2	2	5	2
	Model ID	xlnet-base-cased	xlnet-base-cased	xlnet-base-cased	xlnet-base-cased	xlnet-base-cased
	Num. params	117312004	117310466	117310466	117312773	117310466

Table B.1: Model specifications.

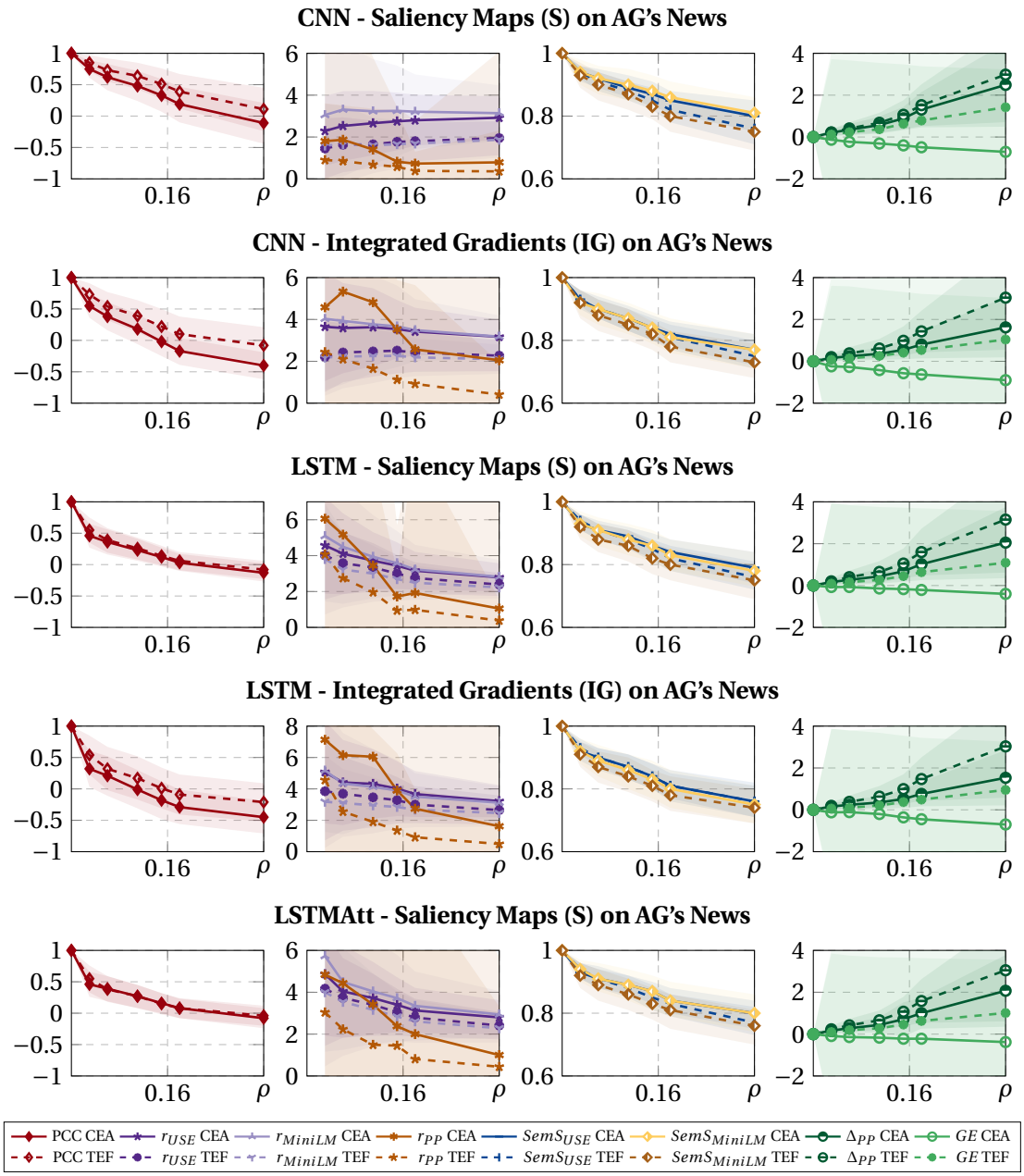
the transformers directly map the feature outputs to the output logits with a fully-connected layer and utilize the Hugging Face preptrained tokenizers for each architecture respectively. Table B.1 contains the model specifications. We train each model with a standard learning rate of 0.001, using the Adam optimizer with the cross-entropy loss and early stopping. We utilize NVIDIA A100 GPUs to speed up training and AR estimation.

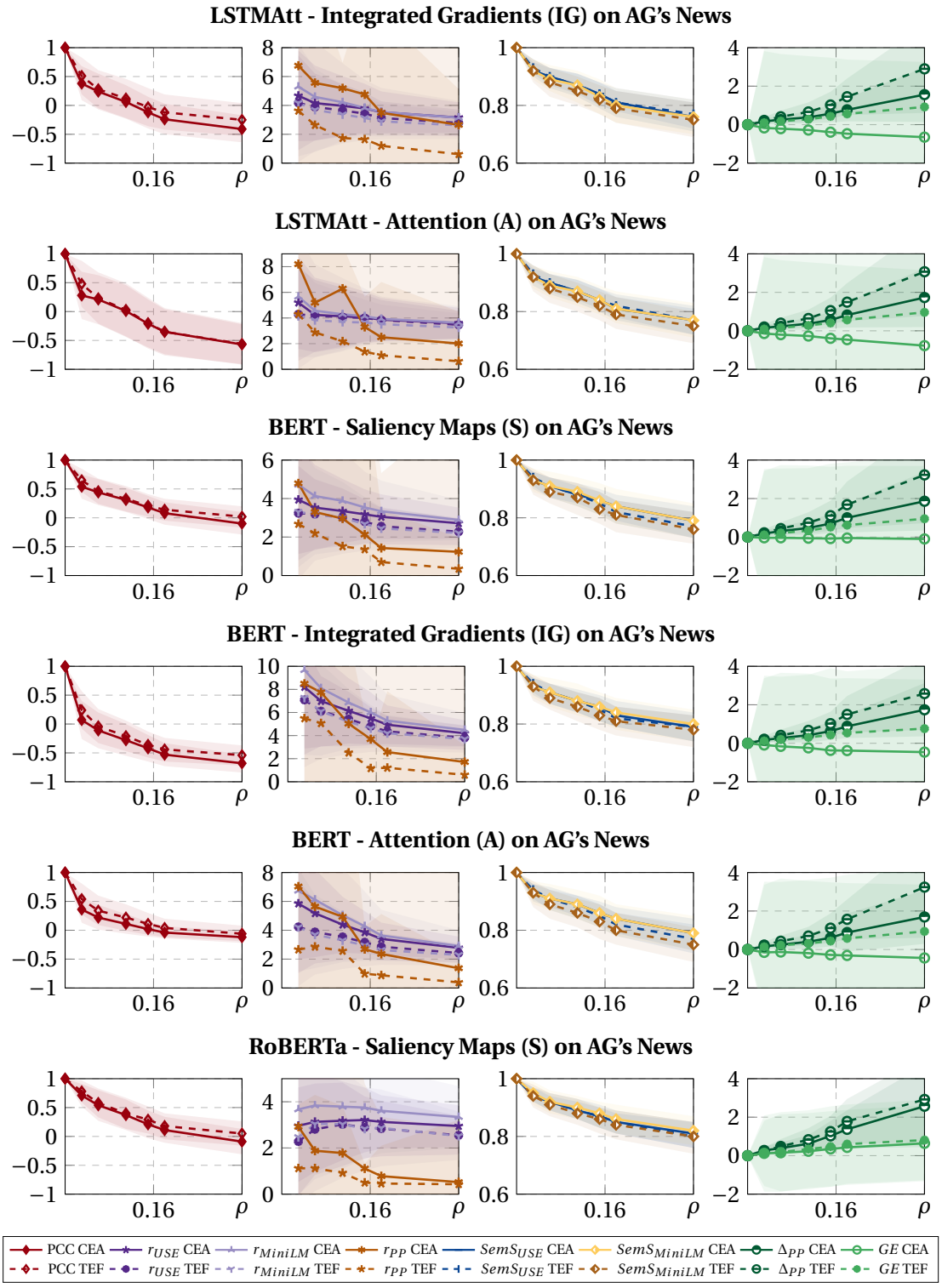
B.3 Additional AR Results

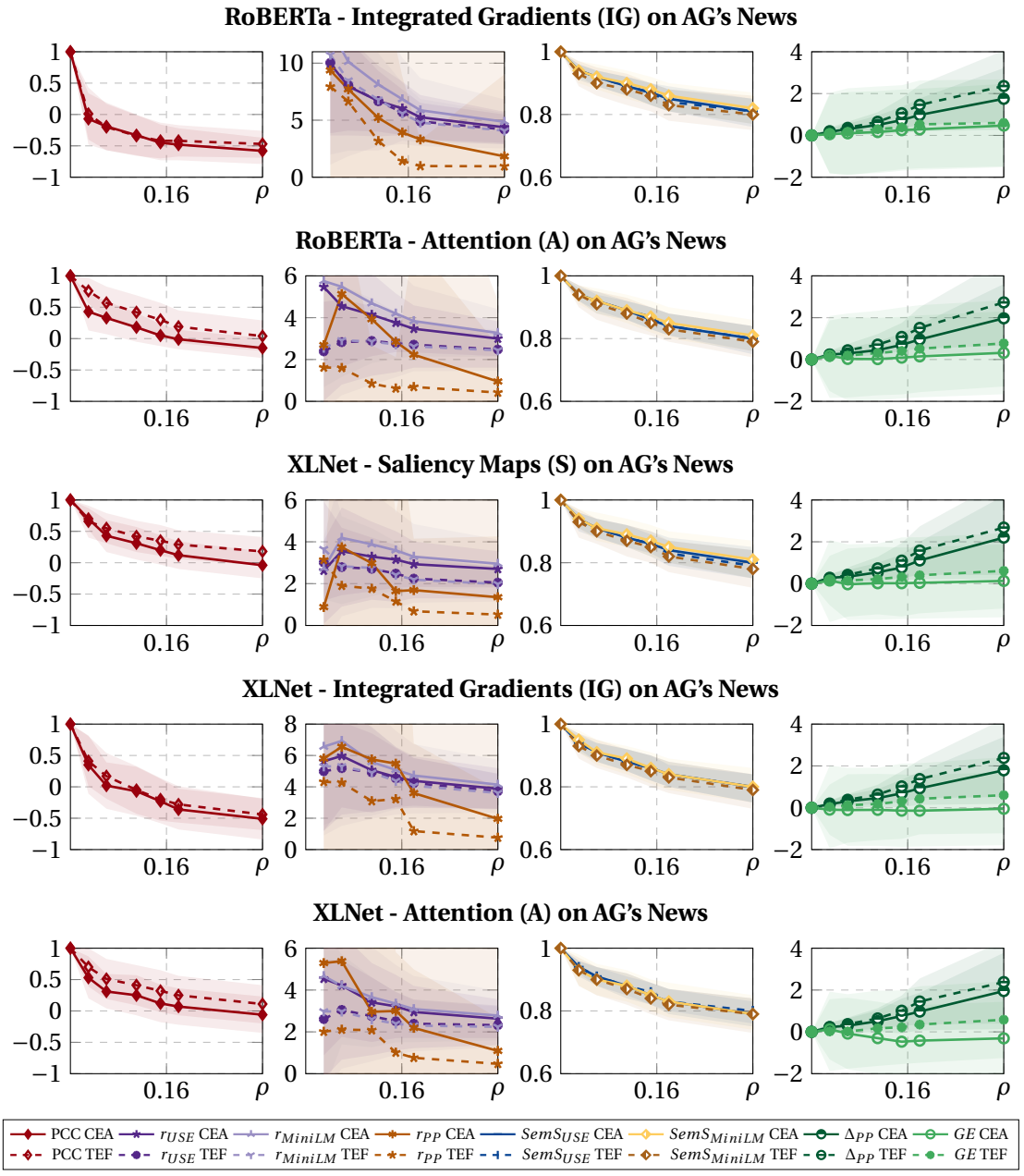
As described in Section 4.5.2, we plot the Pearson Correlation Coefficient between original and adversarial attribution values of the words (1^{st} column from left), the estimated robustness constants r (2^{nd} column from left) as well as the semantic similarities between unperturbed and perturbed input texts, the perplexity increase and the increase in number of grammatical errors (3^{rd} and 4^{th} column from left) after perturbation. We consider a high estimated robustness constant r as *successful* attack, thus low PCC values accompanied by high semantic similarities, low perplexity increase values and grammatical errors. Based on the graphs below, we conclude that CEA consistently yields higher estimated Lipschitz robustness constants r than the reference method TEF, due to lower Pearson correlation between adversarial and

original attribution maps, higher semantic similarities and smaller perplexity increases after applying the adversarial perturbations.

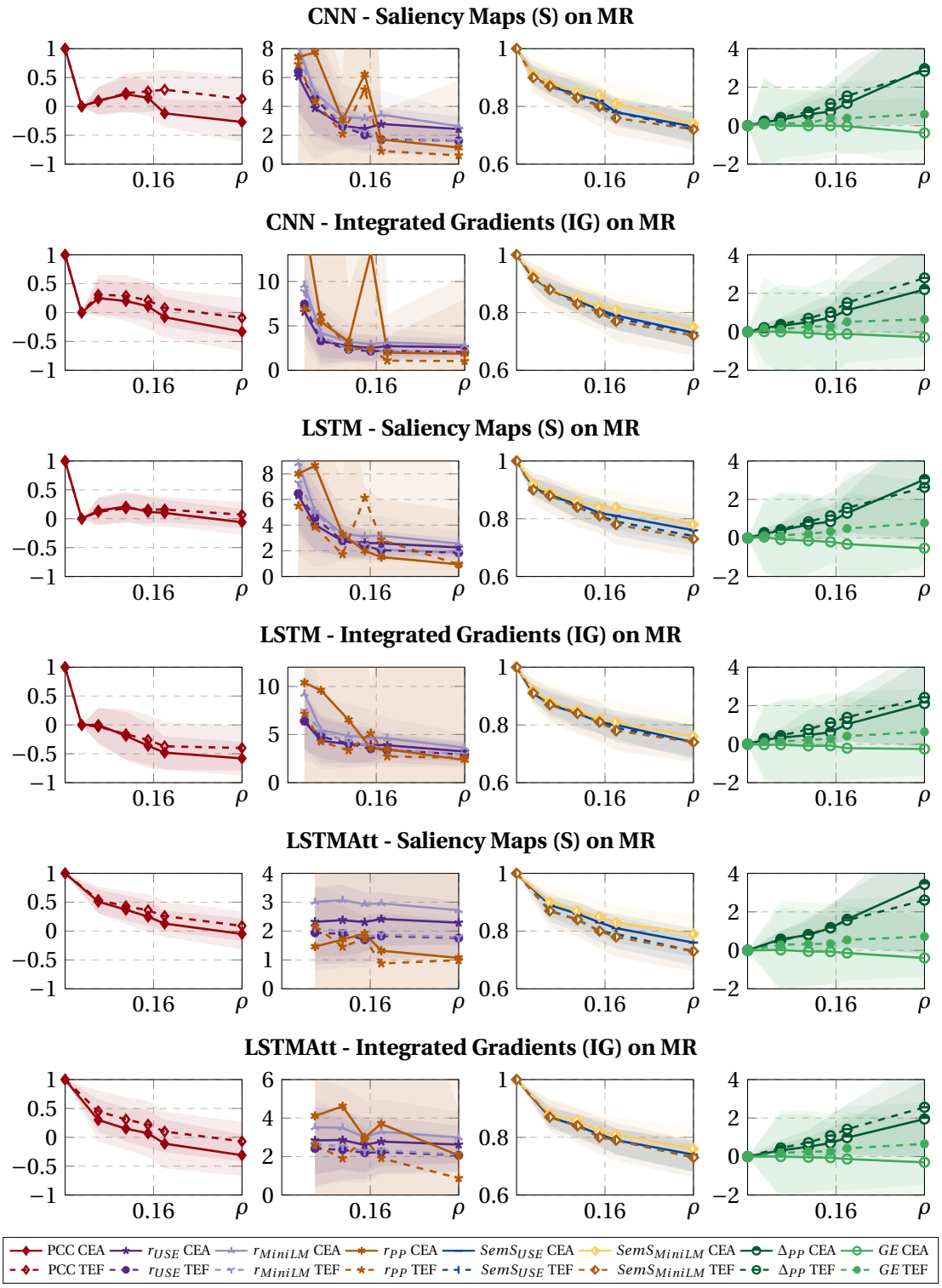
B.3.1 AG's News

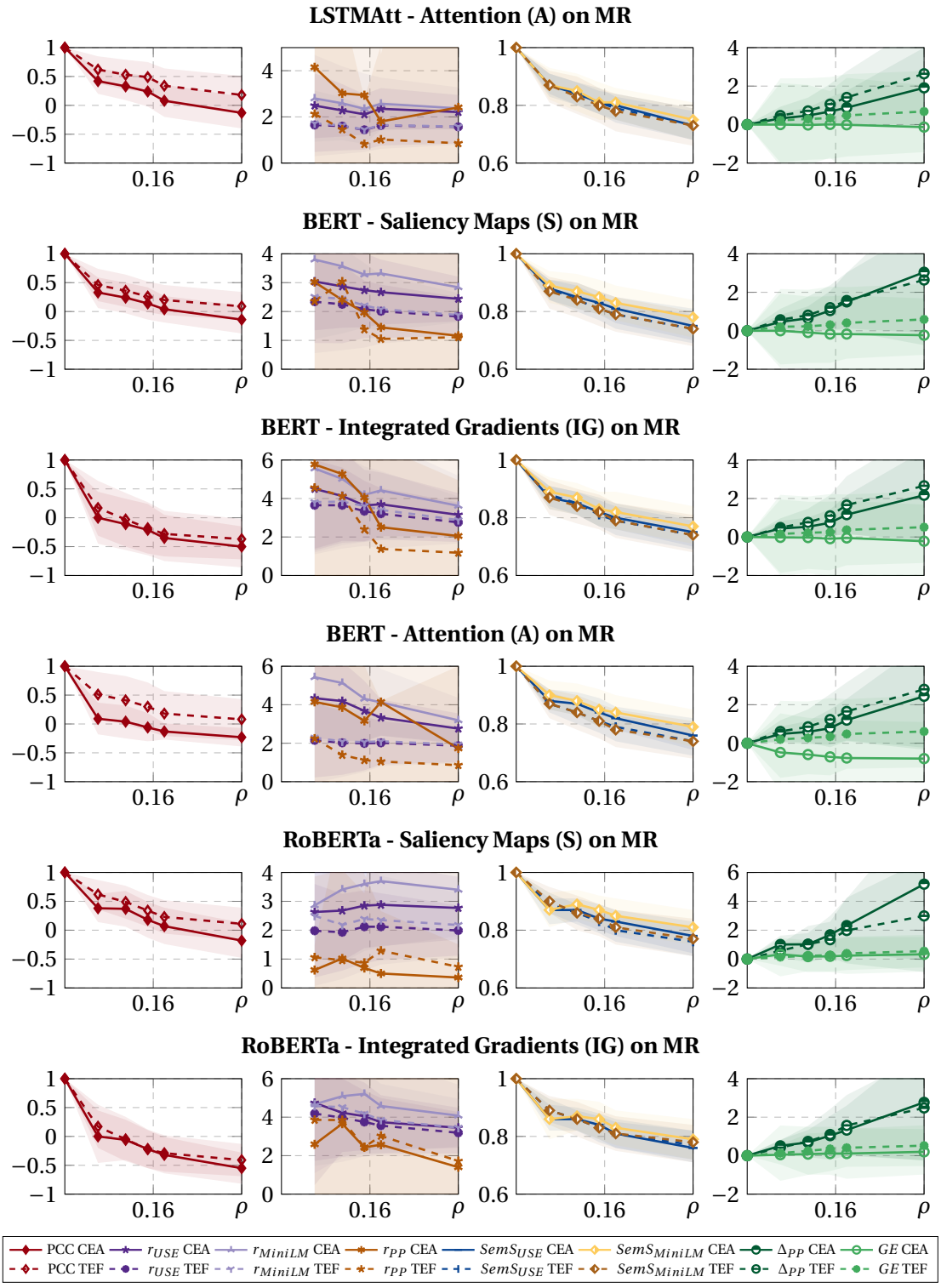


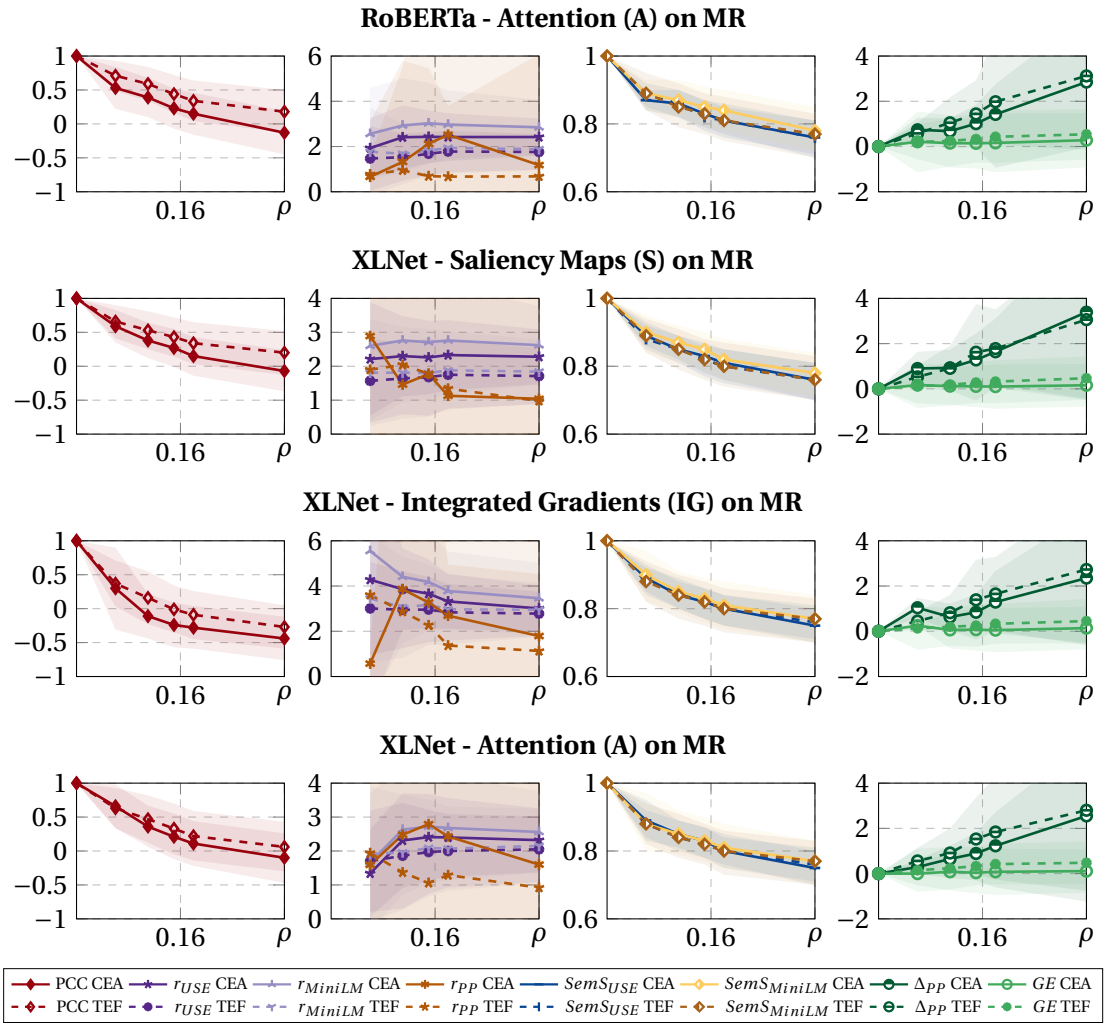

LSTMAtt - Integrated Gradients (IG) on AG's News
LSTMAtt - Attention (A) on AG's News
BERT - Saliency Maps (S) on AG's News
BERT - Integrated Gradients (IG) on AG's News
BERT - Attention (A) on AG's News
RoBERTa - Saliency Maps (S) on AG's News
0.16
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF


RoBERTa - Integrated Gradients (IG) on AG's News
RoBERTa - Attention (A) on AG's News
XLNet - Saliency Maps (S) on AG's News
XLNet - Integrated Gradients (IG) on AG's News
XLNet - Attention (A) on AG's News
0.16
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF

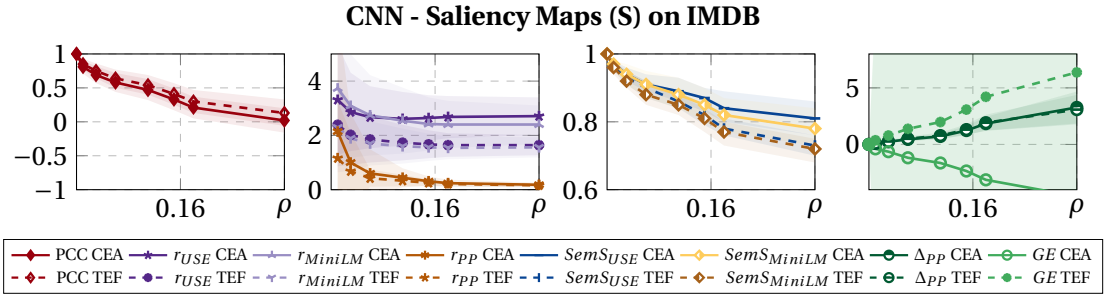
B.3.2 MR

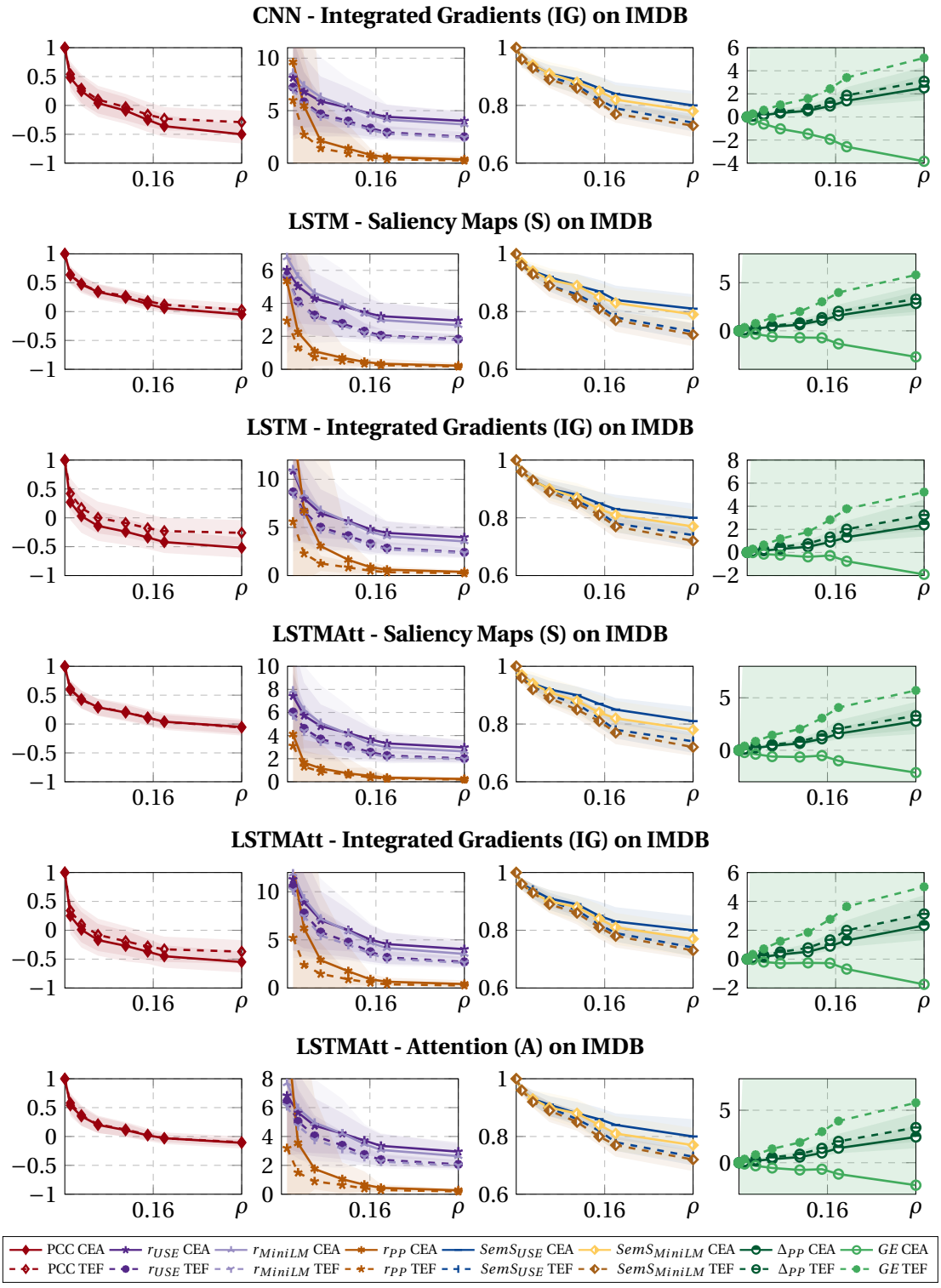

CNN - Saliency Maps (S) on MR
CNN - Integrated Gradients (IG) on MR
LSTM - Saliency Maps (S) on MR
LSTM - Integrated Gradients (IG) on MR
LSTMAtt - Saliency Maps (S) on MR
LSTMAtt - Integrated Gradients (IG) on MR
ρ
0.16
PCC CEA
r_{USE} CEA
r_{MiniLM} CEA
r_{PP} CEA
$SemS_{USE}$ CEA
$SemS_{MiniLM}$ CEA
Δ_{PP} CEA
GE CEA
PCC TEF
r_{USE} TEF
r_{MiniLM} TEF
r_{PP} TEF
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ TEF
Δ_{PP} TEF
GE TEF

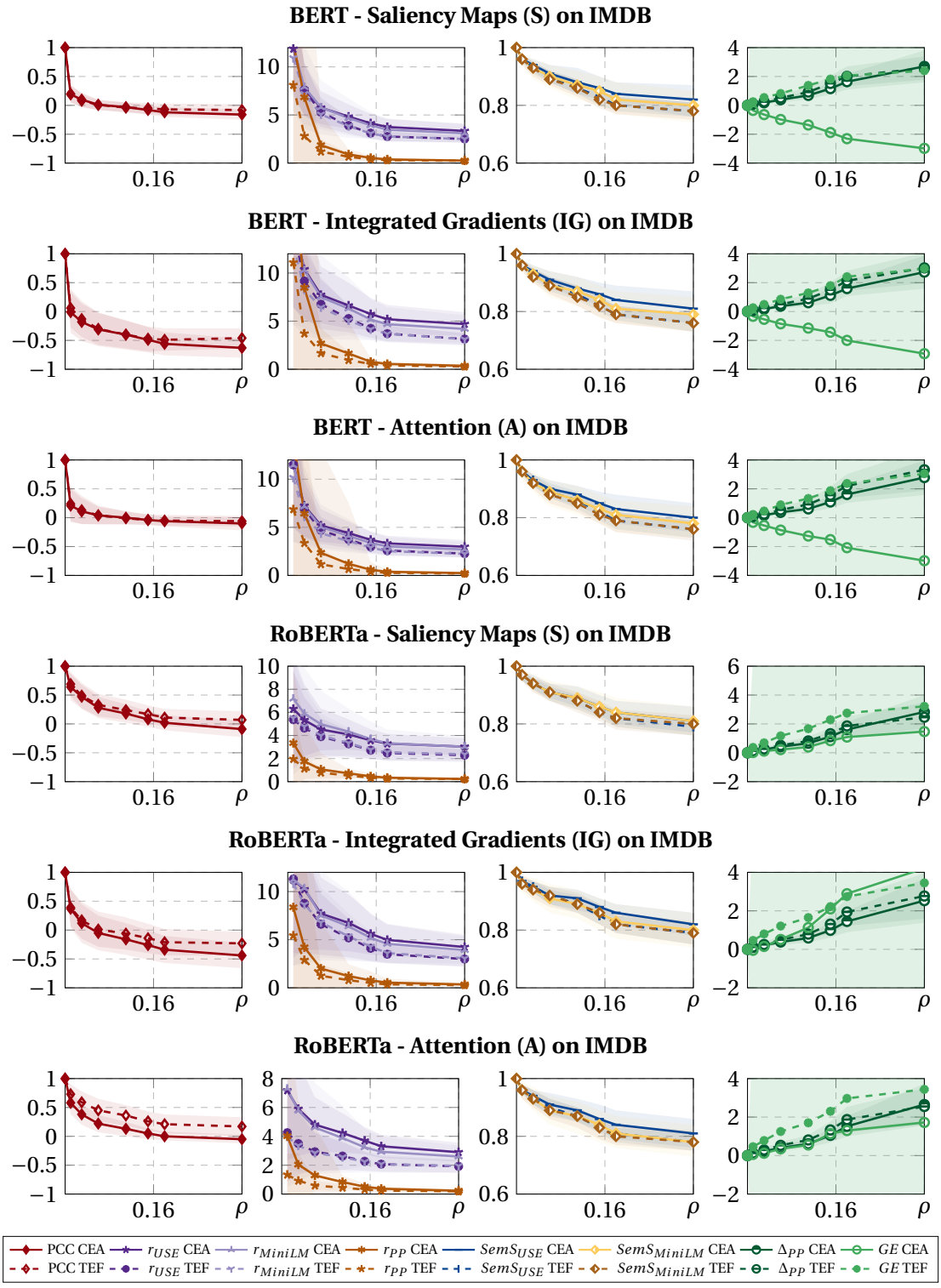

LSTMAtt - Attention (A) on MR
BERT - Saliency Maps (S) on MR
BERT - Integrated Gradients (IG) on MR
BERT - Attention (A) on MR
RoBERTa - Saliency Maps (S) on MR
RoBERTa - Integrated Gradients (IG) on MR
0.16
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF

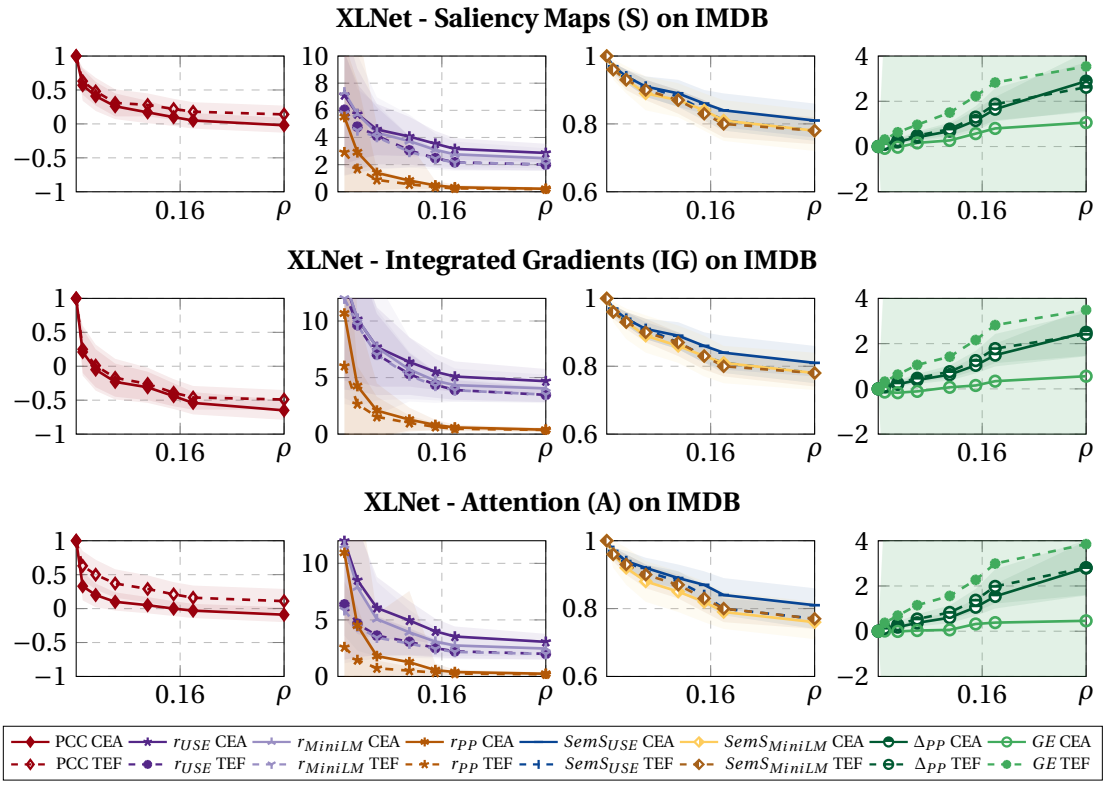

RoBERTa - Attention (A) on MR
XLNet - Saliency Maps (S) on MR
XLNet - Integrated Gradients (IG) on MR
XLNet - Attention (A) on MR
PCC CEA
r_{USE} CEA
r_{MiniLM} CEA
r_{PP} CEA
$SemS_{USE}$ CEA
$SemS_{MiniLM}$ CEA
Δ_{PP} CEA
GE CEA
PCC TEF
r_{USE} TEF
r_{MiniLM} TEF
r_{PP} TEF
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ TEF
Δ_{PP} TEF
GE TEF
0.16
ρ

B.3.3 IMDB

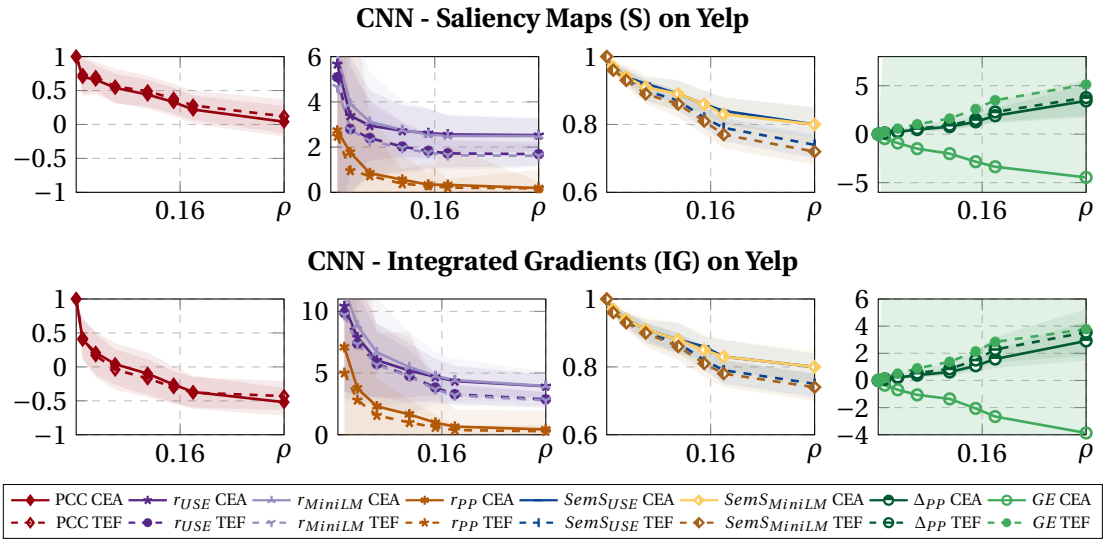

CNN - Saliency Maps (S) on IMDB
PCC CEA
r_{USE} CEA
r_{MiniLM} CEA
r_{PP} CEA
$SemS_{USE}$ CEA
$SemS_{MiniLM}$ CEA
Δ_{PP} CEA
GE CEA
PCC TEF
r_{USE} TEF
r_{MiniLM} TEF
r_{PP} TEF
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ TEF
Δ_{PP} TEF
GE TEF
0.16
ρ

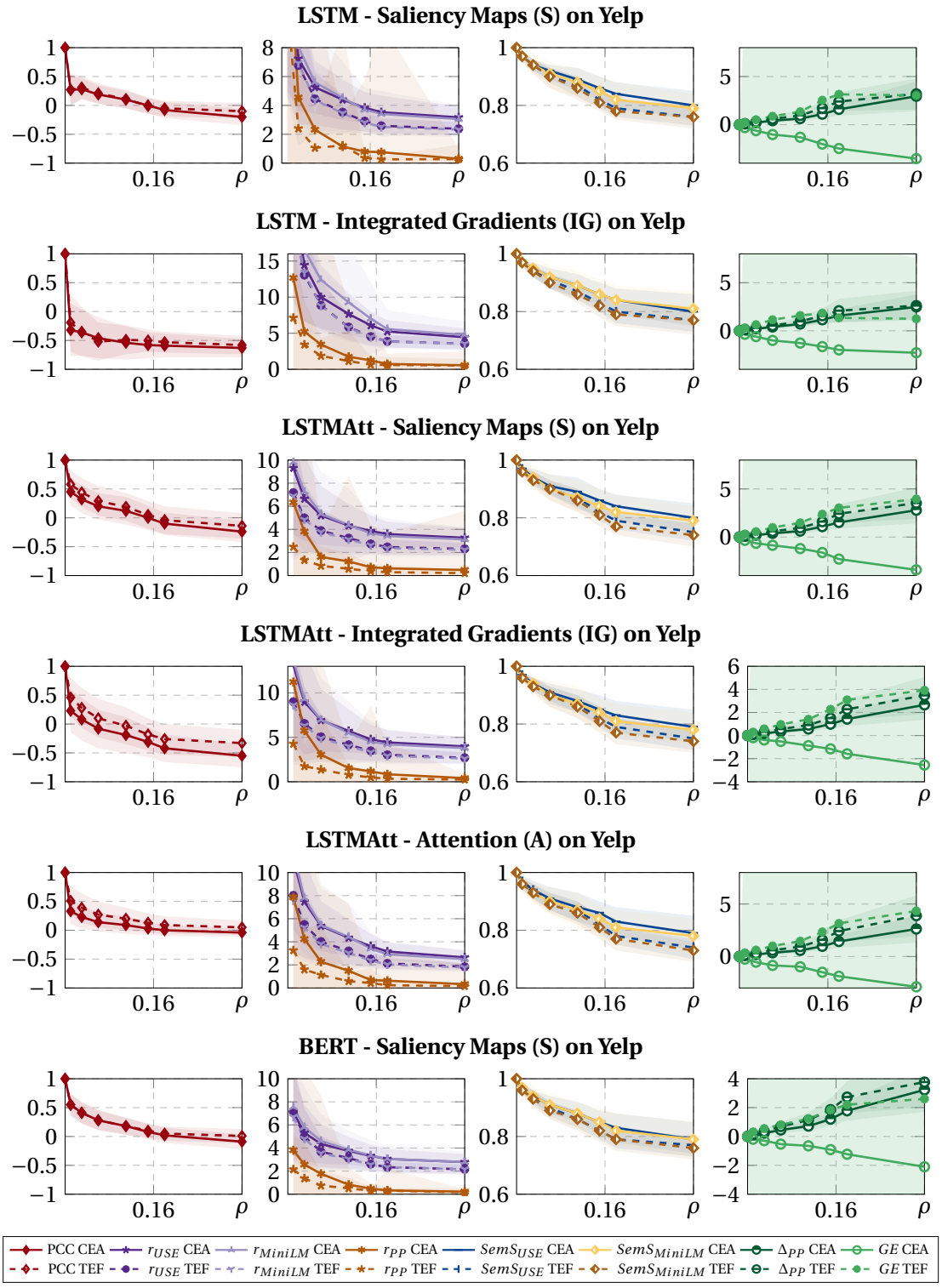

CNN - Integrated Gradients (IG) on IMDB
LSTM - Saliency Maps (S) on IMDB
LSTM - Integrated Gradients (IG) on IMDB
LSTMAtt - Saliency Maps (S) on IMDB
LSTMAtt - Integrated Gradients (IG) on IMDB
LSTMAtt - Attention (A) on IMDB
0.16
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF

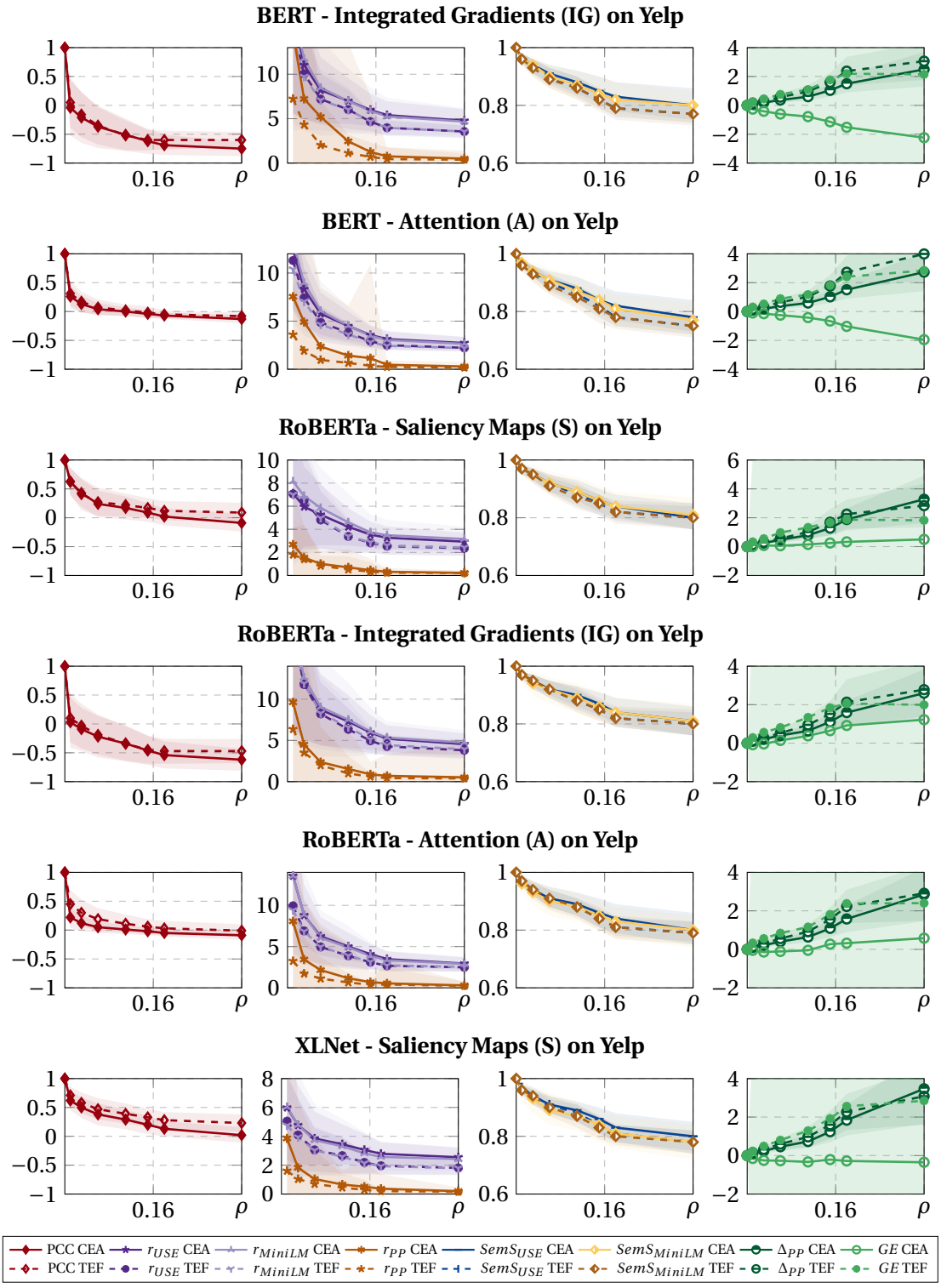

BERT - Saliency Maps (S) on IMDB
BERT - Integrated Gradients (IG) on IMDB
BERT - Attention (A) on IMDB
RoBERTa - Saliency Maps (S) on IMDB
RoBERTa - Integrated Gradients (IG) on IMDB
RoBERTa - Attention (A) on IMDB
ρ
0.16
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF

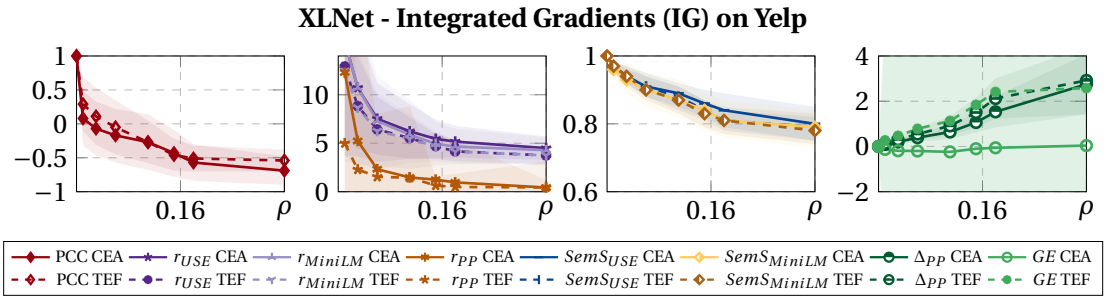

XLNet - Saliency Maps (S) on IMDB
XLNet - Integrated Gradients (IG) on IMDB
XLNet - Attention (A) on IMDB
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF

B.3.4 Yelp

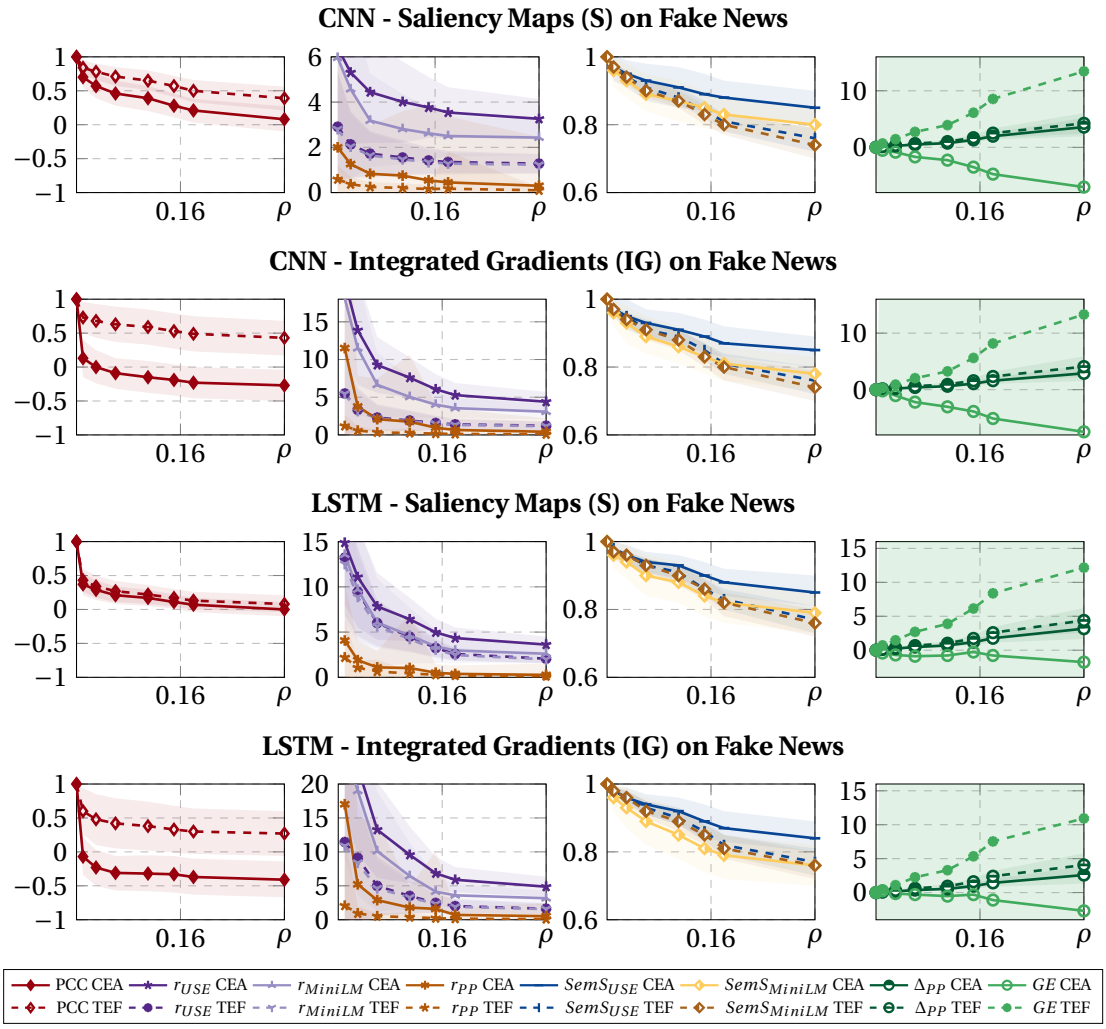

CNN - Saliency Maps (S) on Yelp
CNN - Integrated Gradients (IG) on Yelp
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF

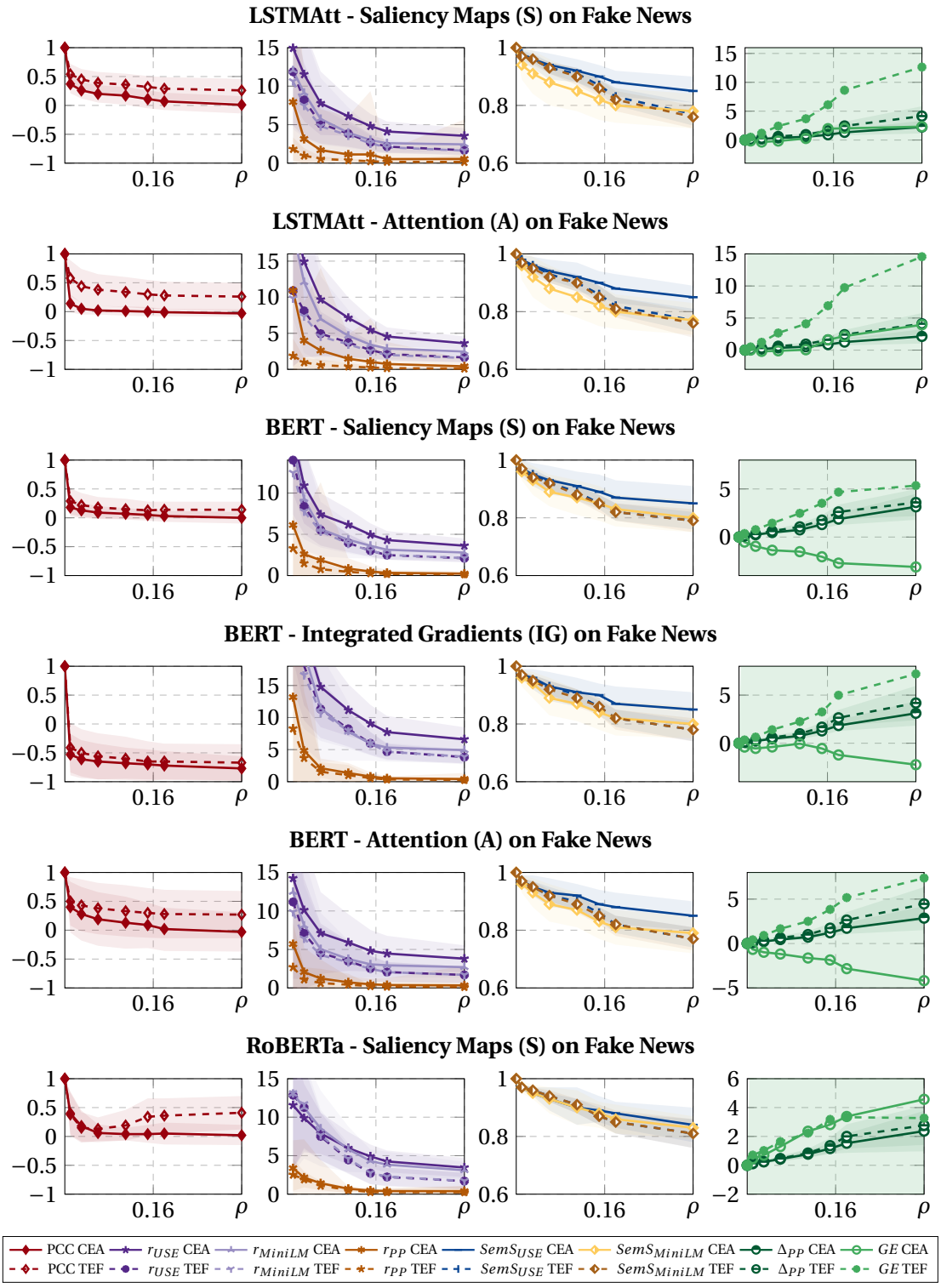

LSTM - Saliency Maps (S) on Yelp
LSTM - Integrated Gradients (IG) on Yelp
LSTMAtt - Saliency Maps (S) on Yelp
LSTMAtt - Integrated Gradients (IG) on Yelp
LSTMAtt - Attention (A) on Yelp
BERT - Saliency Maps (S) on Yelp
0.16
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF

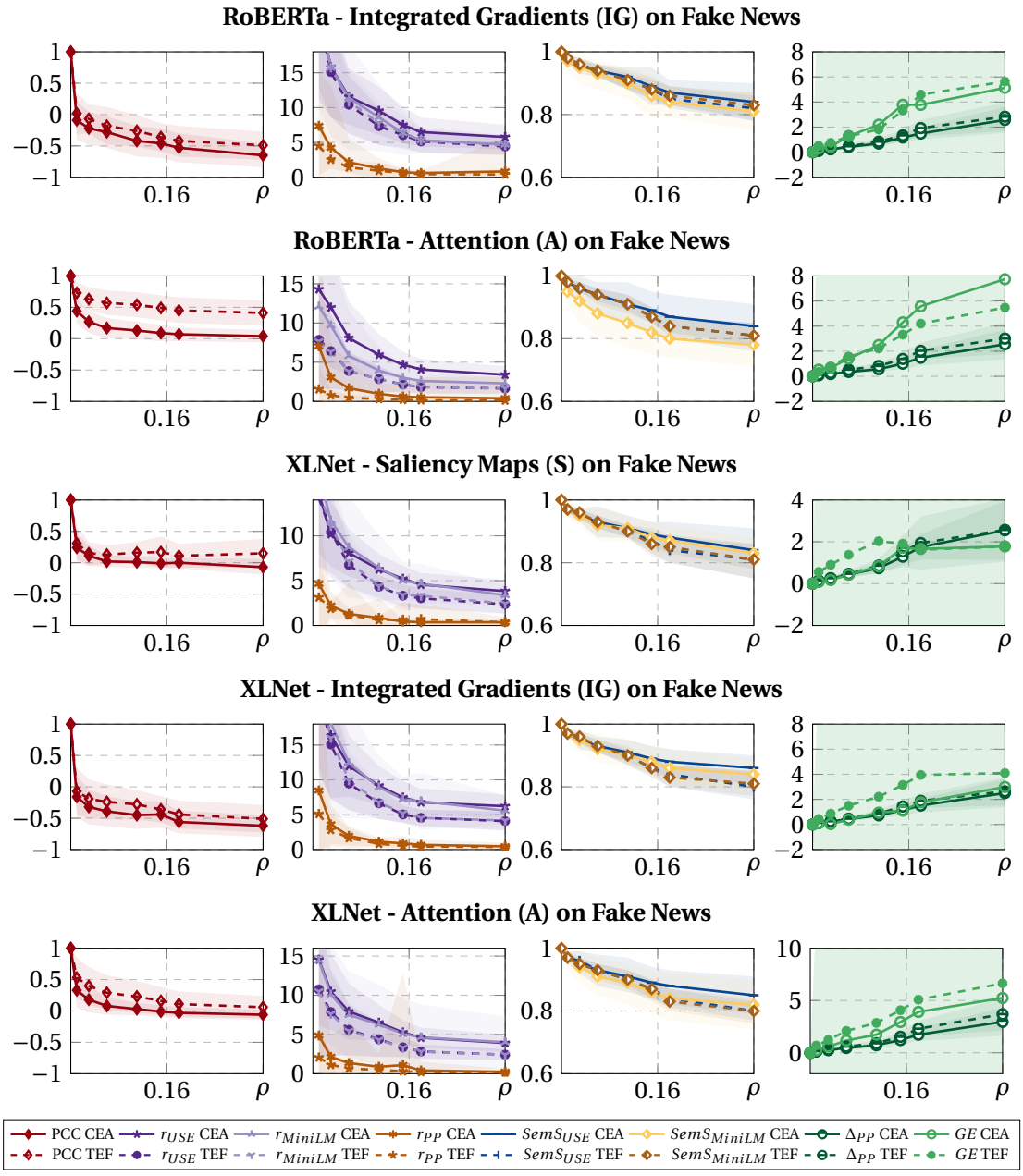

BERT - Integrated Gradients (IG) on Yelp
BERT - Attention (A) on Yelp
RoBERTa - Saliency Maps (S) on Yelp
RoBERTa - Integrated Gradients (IG) on Yelp
RoBERTa - Attention (A) on Yelp
XLNet - Saliency Maps (S) on Yelp
0.16
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF



B.3.5 Fake News




LSTMAtt - Saliency Maps (S) on Fake News
LSTMAtt - Attention (A) on Fake News
BERT - Saliency Maps (S) on Fake News
BERT - Integrated Gradients (IG) on Fake News
BERT - Attention (A) on Fake News
RoBERTa - Saliency Maps (S) on Fake News
0.16
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF


RoBERTa - Integrated Gradients (IG) on Fake News
RoBERTa - Attention (A) on Fake News
XLNet - Saliency Maps (S) on Fake News
XLNet - Integrated Gradients (IG) on Fake News
XLNet - Attention (A) on Fake News
0.16
ρ
PCC CEA
PCC TEF
r_{USE} CEA
r_{USE} TEF
r_{MiniLM} CEA
r_{MiniLM} TEF
r_{PP} CEA
r_{PP} TEF
$SemS_{USE}$ CEA
$SemS_{USE}$ TEF
$SemS_{MiniLM}$ CEA
$SemS_{MiniLM}$ TEF
Δ_{PP} CEA
Δ_{PP} TEF
GE CEA
GE TEF

C Appendix for Chapter 5

C.1 Supplements for Text

C.1.1 Models and Datasets

Parameter	Hallmarks of Cancer	Drug Reviews	MIMIC-III
Input shape	(256,)	(128,)	(4096,)
Num. classes	10	5	50
HF Model ID	michiyasunaga/BioLinkBERT-base	roberta-base	yikuan8/Clinical-Longformer
Num. params	108240394	124649477	148697906

Table C.1: Parameters of our classification models.

We use three public datasets to evaluate the attribution robustness of biomedical text classifiers. Our main goal is to show how robust attribution methods are on these datasets, thus we do not aim to advance the state-of-the-art for classification accuracy, but train models that achieve close to state-of-the-art performance while being relatively easy to train. For each dataset, we use a 60%-20%-20% split for training, test and validation splits, apply basic preprocessing by lower casing the text, removing characters that are not in the Latin alphabet and remove double spaces, new line symbols and double quotes.

The Drug Reviews (DR) dataset consists of patient reviews of different medical drugs, classified into a rating of 1 to 10 for patient satisfaction. In order to increase classification performance, we reduce the number of classes to 5 by merging classes 1 and 2, 3 and 4, 5 and 6 etc. The dataset contains 215063 samples, and we train a RoBERTa model for classification, with the standard cross entropy loss on the first 128 tokens.

The Hallmarks of Cancer (HoC) dataset comprises 1852 biomedical publication abstract associated with 0 or more hallmarks of cancer, thus is a 10-class multilabel classification dataset. We finetune a pretrained BioLinkBERT model for classification, use the first 256 tokens as inputs to the model after tokenization and utilize the binary cross entropy as classification weight.

PARAMETER	HALLMARKS OF CANCER	DRUG REVIEWS	MIMIC-III
CLASSIFICATION LOSS	Multilabel binary cross entropy	Cross entropy	Multilabel binary cross entropy
LR	0.00001	0.000001	0.00004
BATCH SIZE	128	64	4
EPOCHS	50	50	50
PRECISION	32	32	16
ACCUMULATE GRADIENT BATCHES	1	1	4

Table C.2: Parameters used to train our non-robust, vanilla models.

PARAMETER	HALLMARKS OF CANCER	DRUG REVIEWS	MIMIC-III
CANDIDATE EXTRACTOR	PubMedBERT	DistilRoBERTa	Clinical-Longformer
$\rho_{\max}$	0.05	0.05	0.005
$\lvert\mathbb{C}\rvert$	5	5	3
CLASSIFICATION LOSS	Multilabel binary cross entropy	Cross entropy	Multilabel binary cross entropy
RATIO OF ATTACKED SAMPLES IN BATCH	0.3	0.3	0.3
LR	0.00001	0.000001	0.000001
BATCH SIZE	32	64	16
EPOCHS	30	20	20

Table C.3: Parameters used to train our adversarially robust networks.

Our last dataset, the MIMIC-III Discharge Summary dataset consists of patients' ICU discharge summaries, associated with their ICD-9 codes. In order to reduce the overall number of classes from 1800, we only take the 50 most frequent ICD-9 codes. This results in a total of 59647 samples. As the summaries are very long, we finetune a pretrained Clinical-Longformer model for classification, with a maximum sequence length of 4096, default attention window size and global attention on the `[CLS]` (or equivalent) token.

Table C.1 summarizes our models, Table C.2 contains the used hyperparameters for our finetuning process and Table C.5 the resulting accuracies of all our trained models. We use the AdamW optimizer throughout all our experiments.

The Hallmarks of Cancer and Drug Reviews dataset are public datasets and the requirements for MIMIC-III [1] were completed and we comply with their DUA.

C.1.2 AR Estimation and Robust Training

In order to achieve robust attributions, in addition to the vanilla models we train models with robust training objectives. During adversarial training, we augment the training batches with adversarial samples that maximize classification loss. We use the A2T training method for

[1] https://physionet.org/content/mimiciii/1.4/

PARAMETER	HALLMARKS OF CANCER	DRUG REVIEWS
CANDIDATE EXTRACTOR	PubMedBERT	DistilRoBERTa
$\rho_{\max}$	0.05	0.05
A	IG	IG
$d(A_{\text{adv}}, A)$	cosine	cosine
$\lvert\mathbb{C}\rvert$	5	5
CLASSIFICATION LOSS	Multilabel binary cross entropy	Cross entropy
PREDICTION CONSTRAINT	No	Yes
γ	0.85	0.0
δ	0.85	0.7
LR	0.00001	0.000001
BATCH SIZE	4	8
EPOCHS	30	20
RATIO OF ATTACKED SAMPLES IN BATCH	0.6	0.6

Table C.4: Parameters used to train our FAR-IG networks.

extracting adversarial samples, with the parameters summarized in Table C.3. Our FAR models are trained with the robust objectives from Section 5.2, and the hyperparameters are written in Table C.4.

		HALLMARKS OF CANCER			DRUG REVIEWS			MIMIC-III		
	MODEL	*Van.*	*Adv.*	*FAR-IG*	*Van.*	*Adv.*	*FAR-IG*	*Van.*	*Adv.*	*FAR-IG*
NATURAL	ACCURACY	0.95	0.94	0.92	0.9	0.92	0.92	0.92	0.9	-
	PRECISION	0.78	0.74	0.62	0.89	0.92	0.92	0.59	0.57	-
	RECALL	0.89	0.82	0.90	0.9	0.92	0.92	0.71	0.61	-
	F1-SCORE	0.82	0.78	0.73	0.9	0.92	0.92	0.64	0.6	-
	LOSS	0.24	0.27	0.27	0.68	0.36	0.32	0.3	0.33	-
ADVERSARIAL	ACCURACY	0.88	0.89	0.87	0.61	0.67	0.65	0.89	0.9	-
	PRECISION	0.55	0.59	0.5	0.61	0.66	0.65	0.54	0.55	-
	RECALL	0.75	0.7	0.8	0.6	0.67	0.65	0.67	0.61	-
	F1-SCORE	0.61	0.63	0.62	0.6	0.67	0.65	0.59	0.62	-
	LOSS	0.64	0.53	0.44	2.5	1.1	1.2	0.41	0.39	-

Table C.5: Natural and adversarial classification metrics of the non-robust (Van.), adversarially robust (Adv.) and FAR-trained (FAR-IG) models. All metrics are macro-averaged over the samples, as our datasets are highly class-imbalanced.

C.2 Supplements for Images

C.2.1 Parameters and Architectures

We conduct experiments on five vision datasets (MNIST, Fashion-MNIST, CIFAR-10, GTSRB and Restricted Imagenet) to compare our attributional robustness method to state-of-the-art

		MNIST	FASHION-MNIST	CIFAR-10	GTSRB	RESTR. IMAGENET
VAN.	OPTIMIZER	Adam				
	EPOCHS	50				
	BATCH SIZE	50	50	128	128	32
	LR	0.001	0.001	0.01	0.01	0.01
ADV.	OPTIMIZER	Adam				
	EPOCHS	50				
	BATCH SIZE	50	50	128	128	32
	LR	0.0001	0.001	0.001	0.001	0.001
	ADV. RATIO	0.7				
ALIGN	OPTIMIZER	Adam				
	EPOCHS	50				
	BATCH SIZE	50	50	-	-	32
	LR	0.0001	0.0001	-	-	0.0001
	δ	0.5	0.5	-	-	0.5
AAT	OPTIMIZER	Adam				
	EPOCHS	50				
	BATCH SIZE	50	50	128	128	32
	LR	0.0001	0.0001	0.0001	0.0001	0.0001
	δ	0.5	1.0	2.0	0.5	1.5
ADVAAT	OPTIMIZER	Adam				
	EPOCHS	50				
	BATCH SIZE	50	50	128	128	32
	LR	0.0001	0.0001	0.0001	0.0001	0.0001
	δ	0.5	0.5	0.5	0.2	0.5

Table C.6: Parameters to train our vanilla and robust models for image classification.

algorithms. Each model is implemented in PyTorch v1.3.1 and is trained distributedly on six NVIDIA Tesla V100 GPUs with the PyTorch Distributed Data Parallel wrapper. We fix all seeds to 42. Table C.7 contains the evaluation parameters of our experiments, Table C.6 the training parameters. We finetune the natural model to train our robust methods. If we do not mention a specific parameter, it is set to the default value in PyTorch v1.3.1. Moreover, the parameters values of IFIA during training are kept as the values during evaluation.

C.2.2 Initialization Methods

We use seven different initialization methods for addressing the dependency of attributional robustness on the initialization. These are detailed in the next paragraphs. If a parameter is not mentioned, it is kept as the default value defined in PyTorch. The training setup is kept constant for each initialization, and corresponds to the setup mentioned in the previous section for the different models.

PTD. Default PyTorch initialization for linear and convolutional layers. This is the He uniform initialization with $a = \sqrt{5}$ for the weights and a uniform initialization with bounds $b = \pm 1/\sqrt{\texttt{fan\_in}}$ for the bias terms.

CUST. Custom initialization method. Weights are initialized utilizing a zero-centered normal

		MNIST	FASHION-MNIST	CIFAR-10	GTSRB	RESTR. IMAGENET
ARCHITECTURE		CNN [135]	CNN [135]	ResNet [132]	ResNet [132]	ResNet [132]
AA	ATTACK	PGD				
	STEPS	40				
	REL. STEPSIZE	0.03				
AR	ATTACK	Adversarial IFIA with pred. constraint and $\delta = 1$				
	EXPLAINER	Integrated Gradients with baseline 0				
	d_s	Sum-Top-K				
	STEPS	7				
	REL. STEPSIZE	1.2/7				
	β	1.0				
	K	50	50	100	100	300
ε		0.3	0.1	0.03	0.03	0.01
NUMBER OF RESTARTS		3				

Table C.7: Parameters to evaluate adversarial accuracy (**AA**) and attribution robustness (**AR**) of our image classifiers.

distribution with a standard deviation of 0.1, and biases are initialized to be 0.1, both for linear and convolutional layers.

UNI. Uniform initialization method. Weights and biases are initialized utilizing a uniform distribution with bounds $b = \pm 0.1$ for all layers.

HU. He uniform initialization method. Weights are initialized utilizing the default PyTorch He uniform initialization, biases are set to zero.

HN. He uniform initialization method. Weights are initialized utilizing the default PyTorch He normal initialization, biases are set to zero.

GU. Glorot uniform initialization method. Weights are initialized utilizing the default PyTorch Glorot uniform initialization, biases are set to zero.

GN. Glorot normal initialization method. Weights are initialized utilizing the default PyTorch Glorot normal initialization, biases are set to zero.

		PTD	CUST	UNI	HU	HN	GU	GN
NA	VAN.	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	ADV.	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	AAT	0.99	0.99	0.99	0.98	0.99	0.99	0.99
AA	VAN.	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ADV.	0.94	0.90	0.94	0.94	0.94	0.94	0.94
	AAT	0.09	0.09	0.06	0.07	0.04	0.06	0.09
IN	VAN.	0.23	0.09	0.18	0.13	0.10	0.27	0.26
	ADV.	0.35	0.21	0.40	0.12	0.11	0.21	0.37
	AAT	0.39	0.30	0.33	0.38	0.36	0.38	0.39
CO	VAN.	0.20	0.03	0.13	0.08	0.06	0.19	0.20
	ADV.	0.05	0.02	0.08	0.01	0.01	0.45	0.55
	AAT	0.28	0.18	0.24	0.24	0.24	0.27	0.28

Table C.8: Estimated attributional robustness (**IN** and **CO**) for several different initialization methods. The results are reported for models trained naturally (VAN.), adversarially (ADV.) as well as with our **AAT** objective on MNIST. The natural and adversarial accuracy is given in the NA and AA rows. While accuracies of the models are similar, their estimated attributional robustness varies significantly throughout the initializations.

C.3 More Text Examples

	VANILLA	ADVERSARIAL	FAR-IG
Original	'i have been on **invokana** **since** september 2013, so **a** little **over** **a** **year.** i **have** **experienced** **hair** **loss,** tiredness, **and** **yeast** **infections.** **i** **talked** to my **doctor** about the **hair** **loss,** which i experienced for over **a** **year.** he **has** upped **my** metformin to **the** maximum **dosage.** **my** hair **has** **stopped** falling **out.** i **am** **also** using **rosemary** **essential** **oil** to **help** with **hair** **loss,** and probiotics for **the** yeast **infection.** i have **had** amazing **results** with this **medication** in regards to **blood** sugar **control.** **my** a1c **went** from 12.3 to 7.1 i have never had $F(\boldsymbol{s}, \hat{l} = \text{"8.0"}) = 1.0$	'i have been on **invokana** **since** september 2013, so **a** little **over** **a** **year.** i **have** **experienced** **hair** **loss,** tiredness, **and** **yeast** **infections.** **i** **talked** to my **doctor** about the **hair** **loss,** which i experienced for over **a** **year.** he **has** upped **my** metformin to **the** maximum **dosage.** **my** hair **has** **stopped** falling **out.** i **am** **also** using **rosemary** **essential** **oil** to **help** with **hair** **loss,** and probiotics for **the** yeast **infection.** i have **had** amazing **results** with this **medication** in regards to **blood** sugar **control.** **my** a1c **went** from 12.3 to 7.1 i have never had $F(\boldsymbol{s}, \hat{l} = \text{"8.0"}) = 0.88$	'i have been on **invokana** **since** september 2013, so **a** little **over** **a** **year.** i **have** **experienced** **hair** **loss,** tiredness, **and** **yeast** **infections.** **i** **talked** to my **doctor** about the **hair** **loss,** which i experienced for over **a** **year.** he **has** upped **my** metformin to **the** maximum **dosage.** **my** hair **has** **stopped** falling **out.** i **am** **also** using **rosemary** **essential** **oil** to **help** with **hair** **loss,** and probiotics for **the** yeast **infection.** i have **had** amazing **results** with this **medication** in regards to **blood** sugar **control.** **my** a1c **went** from 12.3 to 7.1 i have never had $F(\boldsymbol{s}, \hat{l} = \text{"8.0"}) = 0.97$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	'i have been on invokana since september 2013, so a little over a year. i have noticed scalp loss, tiredness, and yeast infections. i talked to my doctor about the hair loss, which i experienced for over a year. he has upped my metformin to the maximum dosage. my hair has stopped falling out. i am also using rosemary essential oil to help with hair loss, and probiotics for the yeast infection. i have had numerous results with this medication in regards to blood glucose control. my a1c went from 12.3 to 7.1 i have never had $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"8.0"}) = 0.77$ **Cosine Similarity** = -0.07 $SemS = 1.0$	'i have been on invokana since september 2013, taking a tad over a year. i have experienced hair loss, tiredness, and yeast infections. i complained to my doctor about the hair loss, which i experienced for over a year. he has upped my metformin to the recommended dosage. my hair has stopped falling out. i am also using rosemary essential oil to help with hair loss, and probiotics for the yeast infection. i have had amazing results with this medication in regards to blood sugar control. my a1c went from 12.3 to 7.1 i have never had $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"8.0"}) = 0.73$ **Cosine Similarity** = 0.35 $SemS = 1.0$	'i have been on invokana since september 2013, so a little over a year. i have experienced hair loss, tiredness, and yeast infections. i talked to my doctor about the hair loss, which i experienced for over a year. he has upped my metformin to the recommended dosage. my hair has stopped falling out. i am also using rosemary olive oil to help with hair loss, and probiotics for the yeast infection. i have had amazing success with this medication in regards to blood pressure control. my a1c went from 12.3 to 7.1 i have never had $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"8.0"}) = 0.93$ **Cosine Similarity** = 0.58 $SemS = 1.0$
Original	'i have had intractable migraine for 28 years, and migraines from the age of 10 until 28 years ago when it never quit.. i went through many different trials of treatment & nothing worked, so finally the headache specialist gave me vicodin.it worked and i was able to begin living life again. then a new md took the vicodin away and gave me topamax. my life was hell. i live alone in a 2 story house and i had to scoot up/down on my butt. i am 66 & disabled (from strokes) and i was terrified. $F(\boldsymbol{s}, \hat{l}=\text{"2.0"}) = 1.0$	'i have had intractable migraine for 28 years, and migraines from the age of 10 until 28 years ago when it never quit.. i went through many different trials of treatment & nothing worked, so finally the headache specialist gave me vicodin.it worked and i was able to begin living life again. then a new md took the vicodin away and gave me topamax. my life was hell. i live alone in a 2 story house and i had to scoot up/down on my butt. i am 66 & disabled (from strokes) and i was terrified. $F(\boldsymbol{s}, \hat{l}=\text{"2.0"}) = 0.98$	'i have had intractable migraine for 28 years, and migraines from the age of 10 until 28 years ago when it never quit.. i went through many different trials of treatment & nothing worked, so finally the headache specialist gave me vicodin.it worked and i was able to begin living life again. then a new md took the vicodin away and gave me topamax. my life was hell. i live alone in a 2 story house and i had to scoot up/down on my butt. i am 66 & disabled (from strokes) and i was terrified. $F(\boldsymbol{s}, \hat{l}=\text{"2.0"}) = 0.99$
Adversarial	'i have had intractable migraine for 28 years, and migraines from the age of 10 until 28 years old when it never quit.. i went through many different trials of treatment & nothing worked, so finally the arthritis specialist gave me vicodin.it worked and i was able to begin living life again. then a new md took the vicodin away and gave me topamax. my life was saying i was alone in a 2 story house and i had to scoot up/down on my butt. i am 66 & disabled (from strokes) and i was terrified. $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"2.0"}) = 0.71$ **Cosine Similarity** = 0.01 $SemS = 0.89$	'i have had intractable epilepsy for 28 years, and migraines from the age of 10 until 28 years ago when it never quit.. i went through many different trials of treatment & nothing worked, so finally the epilepsy specialist gave me vicodin.it worked and i was able to begin living life again. then a new md took the vicodin away and gave me topamax. my life was ruined i live alone in a 2 nd house and i had to scoot up/down on my butt. i am 66 & disabled (from strokes) and i was terrified. $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"2.0"}) = 0.98$ **Cosine Similarity** = 0.26 $SemS = 0.81$	'i have had intractable epilepsy for 28 years, and migraines from the age of 10 until 28 years ago when it never quit.. i went through many different trials of treatment & nothing worked, so finally the epilepsy specialist gave me vicodin.it worked and i was able to begin living life again. then a new md took the vicodin away and gave me topamax. my life was hell. i live alone in a 2 nd flat and i had to scoot up/down on my butt. i am 66 & disabled (from strokes) and i was terrified $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"2.0"}) = 0.99$ **Cosine Similarity** = 0.58 $SemS = 0.81$

Appendix C. Appendix for Chapter 5

	VANILLA	ADVERSARIAL	FAR-IG
Original	'i have only been using nuva ring for 5 days... i have not been sick in any way.. or had mood swings.. ive noticed i have alittle nore energy to get things done around the house. my sex drive i believe has increased a tiny bit... already was high but i haven't had sex yet since i have had it in due to my partners work schedule.i do feel blowed everyday n i get pains in my stomach here and their like period cramps but nothing to intence. i so far do really like this birth control.. i hope it makes my period leas painful and... $F(\boldsymbol{s}, \hat{l}=\text{"8.0"}) = 1.0$	'i have only been using nuva ring for 5 days... i have not been sick in any way.. or had mood swings.. ive noticed i have alittle nore energy to get things done around the house. my sex drive i believe has increased a tiny bit... already was high but i haven't had sex yet since i have had it in due to my partners work schedule.i do feel blowed everyday n i get pains in my stomach here and their like period cramps but nothing to intence. i so far do really like this birth control.. i hope it makes my period leas painful and... $F(\boldsymbol{s}, \hat{l}=\text{"8.0"}) = 0.93$	'i have only been using nuva ring for 5 days... i have not been sick in any way.. or had mood swings.. ive noticed i have alittle nore energy to get things done around the house. my sex drive i believe has increased a tiny bit... already was high but i haven't had sex yet since i have had it in due to my partners work schedule.i do feel blowed everyday n i get pains in my stomach here and their like period cramps but nothing to intence. i so far do really like this birth control.. i hope it makes my period leas painful and... $F(\boldsymbol{s}, \hat{l}=\text{"8.0"}) = 0.91$
Adversarial	'i have only been using nuva ring for 5 days... i have not been manic in any way.. or had mood swings.. ive noticed i have alittle nore energy to get things done around the house. my sex drive i believe has increased a tiny bit... already was high but i haven't had sex yet since i have had it in due to my partners work schedule.i do feel blowed everyday n i get pains in my stomach here and their like period cramps but nothing to intence. lol so NOT do i like this under control.. i hope it makes my period leas painful and... $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"8.0"}) = 1.0$ **Cosine Similarity** = -0.35 $SemS = 0.91$	'i have only been using nuva ring for 5 days... i have not been manic in any way.. or had mood swings.. ive noticed i have alittle nore energy to get things done around the house. my sex drive i believe has increased a tiny bit... already was high but i haven't had sex yet since i have had it in due to my partners work schedule.i do feel blowed everyday n i get pains in my stomach here and their like period cramps but nothing to intence. i so i would not take this birth control.. i hope it makes my period leas painful and... $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"8.0"}) = 0.96$ **Cosine Similarity** = -0.12 $SemS = 0.91$	'i have only been using nuva ring for 5 days... i have not been depressed in any way.. or had mood swings.. ive glad i have alittle nore energy to get things done around the house. my gas density i believe has increased a tiny bit... already was high but i haven't had intercourse yet since i have had it in due to my partners work schedule.i do feel blowed everyday n i get pains in my stomach here and their like period cramps but nothing to intence. i so far do really like this birth control.. i hope it makes my period leas painful and... $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"8.0"}) = 0.48$ **Cosine Similarity** = 0.59 $SemS = 0.73$
Original	'i started using this product a little more than a week ago.i applied it three nights in a row as instructed, and went to a party the next day to test it out.i still sweated, but not nearly as much, and i had hope that with time i would be totally sweat free.i applied it once again the following night, only to continue to sweat the next day.since then (it's been about four days) i have applied hypercare every night without any improvements in the amount i sweat.today was the first day of school and i was sweat the entire day, unable to lift my $F(\boldsymbol{s}, \hat{l}=\text{"2.0"}) = 1.0$	'i started using this product a little more than a week ago.i applied it three nights in a row as instructed, and went to a party the next day to test it out.i still sweated, but not nearly as much, and i had hope that with time i would be totally sweat free.i applied it once again the following night, only to continue to sweat the next day.since then (it's been about four days) i have applied hypercare every night without any improvements in the amount i sweat.today was the first day of school and i was sweat the entire day, unable to lift my $F(\boldsymbol{s}, \hat{l}=\text{"2.0"}) = 0.97$	'i started using this product a little more than a week ago.i applied it three nights in a row as instructed, and went to a party the next day to test it out.i still sweated, but not nearly as much, and i had hope that with time i would be totally sweat free.i applied it once again the following night, only to continue to sweat the next day.since then (it's been about four days) i have applied hypercare every night without any improvements in the amount i sweat.today was the first day of school and i was sweat the entire day, unable to lift my $F(\boldsymbol{s}, \hat{l}=\text{"2.0"}) = 1.0$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	'i started using this product a little more than a week ago.i applied it three nights in a row as instructed, and went to a party the next day to test it out.i still sweated, but not nearly as much, and i was hope that by time i would be totally sweat free.i applied it once again the following night, only to continue to sweat the next day.since then (it's been about four days) i have applied this every night without any breaks in the night i sweat.today was the first day of school and i was sweat the entire day, unable to lift my $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"2.0"}) = 0.87$ **Cosine Similarity** = -0.27 $SemS = 1.0$	'i started using this product a little more than a week ago.i applied it three nights in a row as instructed, and went to a party the next day to test it out.i still sweated, but not nearly as much, and i had thought that with time i would be totally much less applied it once again the following night, only to continue to sweat the next day.since then (it's been about four days) i have applied myself every night without any decrease in the amount i sweat.today was the first day of school and i was sweat the entire day, unable to lift my $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"2.0"}) = 0.55$ **Cosine Similarity** = 0.01 $SemS = 1.0$	'i started using this mask a little more than a week ago.i applied it three nights in a row as instructed, and went to a clinic the next day to test it out.i still sweated, but not nearly as much, and i kept convinced that with time i would be totally sweat free.i applied it once again the following night, only to continue to sweat the next day.since then (it's been about four days) i have applied hypercare every night without any changes in the amount i sweat.today was the first day of school and i was sweat the entire day, unable to lift my $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"2.0"}) = 1.0$ **Cosine Similarity** = 0.42 $SemS = 0.91$
Original	'not every medicine is for everyone, but as one who has tried most of the major pharmaceuticals for major depression, panic attacks, severe anxiety and anxiety related bouts of obsessive compulsive disorder, i can tell you lexapro is the only medicine that i've been able to stay on and be effective for my mental well-being...it is the only one i've had no side effects with. other ssri's have either: made me more anxious and/or depressed, dry mouth, bad weight gain, or extreme fatigue making me into a walking zombie during the day. i've been on lexapro 6 years $F(\boldsymbol{s}, \hat{l}=\text{"10.0"}) = 1.0$	'not every medicine is for everyone, but as one who has tried most of the major pharmaceuticals for major depression, panic attacks, severe anxiety and anxiety related bouts of obsessive compulsive disorder, i can tell you lexapro is the only medicine that i've been able to stay on and be effective for my mental well-being...it is the only one i've had no side effects with. other ssri's have either: made me more anxious and/or depressed, dry mouth, bad weight gain, or extreme fatigue making me into a walking zombie during the day. i've been on lexapro 6 years $F(\boldsymbol{s}, \hat{l}=\text{"10.0"}) = 0.99$	'not every medicine is for everyone, but as one who has tried most of the major pharmaceuticals for major depression, panic attacks, severe anxiety and anxiety related bouts of obsessive compulsive disorder, i can tell you lexapro is the only medicine that i've been able to stay on and be effective for my mental well-being...it is the only one i've had no side effects with. other ssri's have either: made me more anxious and/or depressed, dry mouth, bad weight gain, or extreme fatigue making me into a walking zombie during the day. i've been on lexapro 6 years $F(\boldsymbol{s}, \hat{l}=\text{"10.0"}) = 1.0$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	'not every medicine is for everyone, but as one who has tried most of the major pharmaceuticals for major depression, panic attacks, severe anxiety and anxiety related bouts of obsessive compulsive disorder, i can reassure you lexapro is the only medicine that i've been able to stay on and be effective for my personal well-being...it is the only one i've had no side effects with. My ssri's have never made me more anxious and/or depressed, dry mouth, bad weight gain, or extreme fatigue making me into a walking zombie during the day. i've been on lexapro 6 years $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"10.0"}) = 1.0$ **Cosine Similarity** = -0.2 $SemS = 0.97$	'not every medicine is for everyone, but as one who has prescribed most of the major pharmaceuticals for major depression, panic attacks, severe anxiety and anxiety related bouts of obsessive compulsive disorder, i can tell you lexapro is the only medicine that i've been able to stay on and be effective for my mental well-being...it is the only one i've had no side effects . pill antidepressants have either: made me more anxious and/or depressed, dry mouth, bad weight gain, or extreme fatigue making me into a walking zombie during the day. i've been on lexapro 6 years $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"10.0"}) = 1.0$ **Cosine Similarity** = 0.04 $SemS = 0.95$	'not every medicine is for everyone, but as one who has done most of the major medication for major depression, panic attacks, severe anxiety and anxiety related bouts of obsessive compulsive disorder, i can tell you lexapro is the only medicine that i've been able to focus on and be positive for my mental well-being...it is the only one i've had no side effects with. other ssri's have either: made me more anxious and/or depressed, dry mouth, bad weight gain, or extreme fatigue making me into a walking zombie during the day. i've been on lexapro 6 years $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"10.0"}) = 1.0$ **Cosine Similarity** = 0.52 $SemS = 0.92$
Original	'just took my first dose 5 mg of brintellix - have been on every possible medication including wellbutrin for 15 years, seroquel for 9 years, lexapro for 2 years, just weaned off lexapro.i feel quite odd, butterflies in stomach and brain fog - my daughter has been on brintellix for 2 months and is still vomiting - if this continues, another failed med.' $F(\boldsymbol{s}, \hat{l}=\text{"4.0"}) = 1.0$	'just took my first dose 5 mg of brintellix - have been on every possible medication including wellbutrin for 15 years, seroquel for 9 years, lexapro for 2 years, just weaned off lexapro.i feel quite odd, butterflies in stomach and brain fog - my daughter has been on brintellix for 2 months and is still vomiting - if this continues, another failed med.' $F(\boldsymbol{s}, \hat{l}=\text{"4.0"}) = 1.0$	'just took my first dose 5 mg of brintellix - have been on every possible medication including wellbutrin for 15 years, seroquel for 9 years, lexapro for 2 years, just weaned off lexapro.i feel quite odd, butterflies in stomach and brain fog - my daughter has been on brintellix for 2 months and is still vomiting - if this continues, another failed med.' $F(\boldsymbol{s}, \hat{l}=\text{"4.0"}) = 1.0$
Adversarial	'just took my first full 5 mg of brintellix - have been on every possible medication including wellbutrin for 15 years, seroquel for 9 years, lexapro for 2 years, just weaned off lexapro.i feel quite and butterflies in stomach and brain fog - my daughter has been on brintellix for 2 months and is still vomiting - if this continues, another new med.' $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"4.0"}) = 0.6$ **Cosine Similarity** = -0.18 $SemS = 1.0$	'just took my first batch 5 mg of brintellix - have been on every possible medication including wellbutrin for 15 years, seroquel for 9 years, lexapro for 2 years, just weaned off lexapro.i feel quite sick butterflies in stomach and brain fog - my daughter has been on brintellix for 2 months and is still vomiting - if this continues, another miracle med.' $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"4.0"}) = 0.76$ **Cosine Similarity** = 0.04 $SemS = 1.0$	'just took my first full 5 mg of brintellix - have been on every possible medication including wellbutrin for 15 years, seroquel for 9 years, lexapro for 2 years, just weaned off lexapro.i feel quite odd, butterflies in stomach and brain fog - my daughter has been on brintellix for 2 months and is still awake - if this continues, another antidepressant med.' $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"4.0"}) = 1.0$ **Cosine Similarity** = 0.5 $SemS = 1.0$

	VANILLA	ADVERSARIAL	FAR-IG
Original	'after trying zoloft and lexapro, without any success and made my symptoms worse. luvox helped me getting my life back, the best medicine. i feel much more in control of my ocd. excellent.. even when i feel sleepy sometimes as a side effect.. its worth it!' $F(\boldsymbol{s}, \hat{l}=\text{"10.0"}) = 1.0$	'after trying zoloft and lexapro, without any success and made my symptoms worse. luvox helped me getting my life back, the best medicine. i feel much more in control of my ocd. excellent.. even when i feel sleepy sometimes as a side effect.. its worth it!' $F(\boldsymbol{s}, \hat{l}=\text{"10.0"}) = 1.0$	'after trying zoloft and lexapro, without any success and made my symptoms worse. luvox helped me getting my life back, the best medicine. i feel much more in control of my ocd. excellent.. even when i feel sleepy sometimes as a side effect.. its worth it!' $F(\boldsymbol{s}, \hat{l}=\text{"10.0"}) = 1.0$
Adversarial	'after trying zoloft and lexapro, without any success and made my symptoms worse. luvox helped me getting my life back, the best medicine. i feel much more in spite of my euph excellent.. even when i feel sleepy sometimes as a side effect.. its worth it!' $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"10.0"}) = 1.0$ **Cosine Similarity** = -0.34 $SemS = 0.79$	'after trying zoloft and lexapro, without any success and made my symptoms worse. luvox helped me getting my life back, the best medicine. i feel much more in spite of my sleeping excellent.. even when i feel sleepy sometimes as a side effect.. its worth it!' $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"10.0"}) = 1.0$ **Cosine Similarity** = -0.08 $SemS = 0.81$	'after trying zoloft and lexapro, without any success and made my symptoms worse. luvox helped me getting my life back, the best thing i feel much more in control of my ocd. excellent.. even when i feel pain sometimes as a side effect.. its worth it!' $F(\boldsymbol{s}_{\text{adv}}, \hat{l}=\text{"10.0"}) = 1.0$ **Cosine Similarity** = 0.41 $SemS = 0.8$
Original	'i was put on tri sprintec when i started getting my periods every two weeks, and was on it for three months. it fixed my period problem, but i had never had a problem with acne until starting this. my acne got so much worse and would clear up instantly once i started the sugar pills. i went from a d to a dd which is kind of annoying but i didn't gain much weight at least. the worst part, however, was the tenderness in my breasts, it was horrible. painful to the touch, running or going to the gym was horribly uncomfortable. just like the acne, during the $F(\boldsymbol{s}, \hat{l}=\text{"6.0"}) = 1.0$	'i was put on tri sprintec when i started getting my periods every two weeks, and was on it for three months. it fixed my period problem, but i had never had a problem with acne until starting this. my acne got so much worse and would clear up instantly once i started the sugar pills. i went from a d to a dd which is kind of annoying but i didn't gain much weight at least. the worst part, however, was the tenderness in my breasts, it was horrible. painful to the touch, running or going to the gym was horribly uncomfortable. just like the acne, during the $F(\boldsymbol{s}, \hat{l}=\text{"6.0"}) = 0.92$	'i was put on tri sprintec when i started getting my periods every two weeks, and was on it for three months. it fixed my period problem, but i had never had a problem with acne until starting this. my acne got so much worse and would clear up instantly once i started the sugar pills. i went from a d to a dd which is kind of annoying but i didn't gain much weight at least. the worst part, however, was the tenderness in my breasts, it was horrible. painful to the touch, running or going to the gym was horribly uncomfortable. just like the acne, during the $F(\boldsymbol{s}, \hat{l}=\text{"6.0"}) = 1.0$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	'i was put on tri sprintec when i started getting my periods every two weeks, and was on it for three months. it cured my period problem, but i had never had a problem with sugar until starting this. my stomach got so much better and would clear up instantly once i started the sugar pills. i went from a d to a dd which is kind of annoying but i didn't gain much weight at first the worst part, however, was the tenderness in my breasts, it was horrible. painful to the touch, running or going to the gym was horribly uncomfortable. just like the acne, during the $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"6.0"}) = 0.98$ **Cosine Similarity** = -0.37 $SemS = 0.76$	'i was put on tri sprintec when i started getting my periods every two weeks, and was on it for three months. it caused my period problem, but i had never had a problem with this until starting this. my stomach got so much easier and would clear up immediately once i started the sugar pills. i went from a d to a dd which is kind of annoying but i didn't gain much weight at least. the worst part, however, was the tenderness in my breasts, it was horrible. painful to the touch, running or going to the gym was horribly uncomfortable. just like the acne, during the $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"6.0"}) = 0.72$ **Cosine Similarity** = -0.11 $SemS = 0.75$	'i was put on tri sprintec when i started getting my periods every two weeks, and was on it for three months. it fixed my period problem, but i had never had a problem with this until starting this. my headaches got so much worse and would clear up instantly once i started the sugar syrup which ranging from a d to a dd which is kind of annoying but i didn't gain much weight at least. the worst part, however, was the tenderness in my breasts, it was horrible. painful to the touch, running or going to the gym was horribly uncomfortable. just like the acne, during the $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"6.0"}) = 0.97$ **Cosine Similarity** = 0.68 $SemS = 0.82$
Original	'mestinon helps everyone differently. i started out with ocular mg. apparently, this drug helps most people with their ocular mg, but it doesn't do anything at all to improve my eye. my case of mg rapidly generalized, and difficulty breathing was my 2nd symptom to manifest. mestinon improves my breathing issues somewhat, but doesn't take the shortness of breath away completely. same with my arms and thighs; it helps, but doesn't make the weakness disappear altogether. mestinon can cause diarrhea, but most likely won't if taken alongside a meal or with a small snack. $F(\boldsymbol{s}, \hat{l} = \text{"6.0"}) = 1.0$	'mestinon helps everyone differently. i started out with ocular mg. apparently, this drug helps most people with their ocular mg, but it doesn't do anything at all to improve my eye. my case of mg rapidly generalized, and difficulty breathing was my 2nd symptom to manifest. mestinon improves my breathing issues somewhat, but doesn't take the shortness of breath away completely. same with my arms and thighs; it helps, but doesn't make the weakness disappear altogether. mestinon can cause diarrhea, but most likely won't if taken alongside a meal or with a small snack. $F(\boldsymbol{s}, \hat{l} = \text{"6.0"}) = 1.0$	'mestinon helps everyone differently. i started out with ocular mg. apparently, this drug helps most people with their ocular mg, but it doesn't do anything at all to improve my eye. my case of mg rapidly generalized, and difficulty breathing was my 2nd symptom to manifest. mestinon improves my breathing issues somewhat, but doesn't take the shortness of breath away completely. same with my arms and thighs; it helps, but doesn't make the weakness disappear altogether. mestinon can cause diarrhea, but most likely won't if taken alongside a meal or with a small snack. $F(\boldsymbol{s}, \hat{l} = \text{"6.0"}) = 1.0$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	'mestinon helps everyone .' i started out with ocular mg. apparently, this drug helps most kids with their ocular mg, but it doesn't do anything at all to improve my eye. my case of mg is generalized, and difficulty breathing was my 2nd symptom to manifest. mestinon improves my breathing issues somewhat, but doesn't take the shortness of breath away completely. same with my arms and thighs; it helps, but doesn't make the nausea disappear altogether. mestinon can cause diarrhea, but most likely won't if taken alongside a meal or with a small snack. $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"6.0"}) = 1.0$ **Cosine Similarity** = -0.08 $SemS = 0.92$	'mestinon helps everyone .' i started out with ocular mg. apparently, this supplement helps most people with their ocular mg, but it doesn't do anything at all to relieve my eye. my intake of mg rapidly generalized, and difficulty breathing was my 2nd symptom to manifest. mestinon improves my breathing issues somewhat, but doesn't take the shortness of breath away completely. same with my arms and thighs; it helps, but doesn't make the weakness disappear altogether. mestinon can cause diarrhea, but most likely won't if taken alongside a meal or with a small snack. $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"6.0"}) = 0.99$ **Cosine Similarity** = 0.38 $SemS = 0.94$	'mestinon helps everyone .' i started out with ocular mg. apparently, this pill helps most people with their ocular mg, but it doesn't do anything at all to improve my eye. my case of mg rapidly generalized, and difficulty breathing was my 2nd cause to manifest. mestinon improves my breathing issues somewhat, but doesn't take the shortness of coughing away completely. same with my arms and thighs; it helps, but doesn't make the weakness disappear altogether. mestinon can cause diarrhea, but most likely won't if taken alongside a meal or with a small snack. $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"6.0"}) = 1.0$ **Cosine Similarity** = 0.7 $SemS = 0.95$
Original	'picked up a nasty h pylori strain from an casual blind date, i know i should have gotten to know the person better. took a while before symptoms showed up. had severe upset stomach, occasional diarrhea, nausea and slow but steady weight loss. took a long time and several doctors to diagnose my steadily worsening condition. tried prevpak first, seemed to work at first butmy infection came back. the new gi then prescribed pylera after my 3rd endoscopy. pylera has worked, it's been a year and i am still h pylera negative. but it's been brutal and $F(\boldsymbol{s}, \hat{l} = \text{"6.0"}) = 1.0$	'picked up a nasty h pylori strain from an casual blind date, i know i should have gotten to know the person better. took a while before symptoms showed up. had severe upset stomach, occasional diarrhea, nausea and slow but steady weight loss. took a long time and several doctors to diagnose my steadily worsening condition. tried prevpak first, seemed to work at first butmy infection came back. the new gi then prescribed pylera after my 3rd endoscopy. pylera has worked, it's been a year and i am still h pylera negative. but it's been brutal and $F(\boldsymbol{s}, \hat{l} = \text{"6.0"}) = 0.98$	'picked up a nasty h pylori strain from an casual blind date, i know i should have gotten to know the person better. took a while before symptoms showed up. had severe upset stomach, occasional diarrhea, nausea and slow but steady weight loss. took a long time and several doctors to diagnose my steadily worsening condition. tried prevpak first, seemed to work at first butmy infection came back. the new gi then prescribed pylera after my 3rd endoscopy. pylera has worked, it's been a year and i am still h pylera negative. but it's been brutal and $F(\boldsymbol{s}, \hat{l} = \text{"6.0"}) = 0.9$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	'picked up a nasty h pylori ' from an casual doctor date, i know i should have gotten to know the person better. took a while until they showed up. had severe upset stomach, occasional diarrhea, nausea and slow but steady weight loss. took a long time and several doctors to diagnose my steadily worsening condition. tried prevpak first, seemed to work at first butmy infection came back. the new gi then prescribed pylera after my 3rd endoscopy. pylera has worked, it's been a year and i am still h pylera negative. but it's been brutal and $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"6.0"}) = 0.97$ **Cosine Similarity** = -0.19 $SemS = 0.98$	'picked up a nasty h pylori pill from an anonymous internet date, i know i should have gotten to know the person better. took a while until symptoms showed up. had severe upset stomach, occasional diarrhea, nausea and slow but steady weight loss. took a long time and several doctors to diagnose my steadily worsening condition. tried prevpak first, seemed to work at first butmy infection came back. the new gi then prescribed pylera after my 3rd endoscopy. pylera has worked, it's been a year and i am still h pylera negative. but it's been brutal and $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"6.0"}) = 0.82$ **Cosine Similarity** = 0.19 $SemS = 0.9$	'picked up a nasty h pylori rash from an infected infection so i know i should have gotten to know the person better. took a while before symptoms showed up. had severe upset stomach, occasional diarrhea, nausea and slow but steady weight loss. took a long time and several doctors to diagnose my steadily worsening condition. tried prevpak first, seemed to work at first butmy infection came back. the new gi then prescribed pylera after my 3rd endoscopy. pylera has worked, it's been a year and i am still h pylera negative. but it's been brutal and $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"6.0"}) = 0.78$ **Cosine Similarity** = 0.52 $SemS = 0.81$

	VANILLA	ADVERSARIAL	FAR-IG
Original	missense substitutions of uncertain clinical significance in the brca1 gene are a vexing problem in genetic counseling for women who have a family history of breast cancer . in this study , we evaluated the functions of 29 missense substitutions of brca1 in two dna repair pathways . repair of double - strand breaks by homology - directed recombination (hdr) had been previously analyzed for 16 of these brca1 variants , and 13 more variants were analyzed in this study . all 29 variants were also analyzed for function in double - strand break repair by the single - strand annealing (ssa) pathway . we found that among the pathogenic mutations in brca1 , all were defective for dna repair by either pathway . the hdr assay was accurate because all pathogenic mutants were defective for hdr , and all nonpathogenic variants were fully functional for hdr . repair by ssa accurately identified pathogenic mutants , but several nonpathogenic variants were scored as defective or partially defective . these results indicated that specific amino acid residues of the brca1 protein have different effects in the two related dna repair pathways , and these results validate the hdr assay as highly correlative with brca1 - associated breast cancer .	missense substitutions of uncertain clinical significance in the brca1 gene are a vexing problem in genetic counseling for women who have a family history of breast cancer . in this study , we evaluated the functions of 29 missense substitutions of brca1 in two dna repair pathways . repair of double - strand breaks by homology - directed recombination (hdr) had been previously analyzed for 16 of these brca1 variants , and 13 more variants were analyzed in this study . all 29 variants were also analyzed for function in double - strand break repair by the single - strand annealing (ssa) pathway . we found that among the pathogenic mutations in brca1 , all were defective for dna repair by either pathway . the hdr assay was accurate because all pathogenic mutants were defective for hdr , and all nonpathogenic variants were fully functional for hdr . repair by ssa accurately identified pathogenic mutants , but several nonpathogenic variants were scored as defective or partially defective . these results indicated that specific amino acid residues of the brca1 protein have different effects in the two related dna repair pathways , and these results validate the hdr assay as highly correlative with brca1 - associated breast cancer .	missense substitutions of uncertain clinical significance in the brca1 gene are a vexing problem in genetic counseling for women who have a family history of breast cancer . in this study , we evaluated the functions of 29 missense substitutions of brca1 in two dna repair pathways . repair of double - strand breaks by homology - directed recombination (hdr) had been previously analyzed for 16 of these brca1 variants , and 13 more variants were analyzed in this study . all 29 variants were also analyzed for function in double - strand break repair by the single - strand annealing (ssa) pathway . we found that among the pathogenic mutations in brca1 , all were defective for dna repair by either pathway . the hdr assay was accurate because all pathogenic mutants were defective for hdr , and all nonpathogenic variants were fully functional for hdr . repair by ssa accurately identified pathogenic mutants , but several nonpathogenic variants were scored as defective or partially defective . these results indicated that specific amino acid residues of the brca1 protein have different effects in the two related dna repair pathways , and these results validate the hdr assay as highly correlative with brca1 - associated breast cancer .
	$F(\boldsymbol{s}, \hat{l}=$"<multilabel>"$) = 1.0$	$F(\boldsymbol{s}, \hat{l}=$"<multilabel>"$) = 1.0$	$F(\boldsymbol{s}, \hat{l}=$"<multilabel>"$) = 1.0$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	missense substitutions of unknown clinical significance in the brca1 gene are a vexing problem in risk counseling for women who have a family history of breast cancer . in this study , we evaluated the functions of 29 missense variants of brca1 in two dna repair pathways . repair of two - strand breaks by homology - directed recombination (hdr) had been previously analyzed for 16 of these brca1 variants , and 13 more variants were analyzed in this study . all 29 variants were also analyzed for function in single - strand break repair by the single - strand annealing (ssa) pathway . we found that among the 29 variants in brca1 , all were defective for dna repair by either pathway . the hdr assay was accurate because all pathogenic mutants were defective for hdr , and all nonpathogenic variants were fully functional for hdr . repair by ssa accurately identified most variants , but several nonpathogenic variants were scored as defective or partially defective . these results indicated that specific amino acid residues of the brca1 protein have different effects in the two related dna repair pathways , and these results validate the hdr assay as highly correlative with brca1 - associated breast cancer . $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 1.0 **Cosine Similarity** = 0.13 $SemS$ = 0.82	missense variants of uncertain clinical significance in the brca1 gene are a vexing problem in genetic counseling for women who have a family history of hereditary cancer . in this study , we evaluated the functions of 29 nonsynonymous polymorphisms of brca1 in two dna repair pathways . repair of double - strand breaks by homology - directed recombination (hdr) had been previously analyzed for 16 of these brca1 variants , and 13 more variants were analyzed in this study . all 29 variants were also analyzed for function in double - strand break repair by the single - strand annealing (ssa) pathway . we found that among the pathogenic mutations in brca1 , all were defective for dna repair by either pathway . the hdr assay was accurate because all missense mutants were defective for hdr , and all nonpathogenic variants were fully functional for hdr . repair by ssa accurately identified pathogenic variants , but several nonpathogenic variants were scored as defective or partially defective . these results indicated that specific amino acid residues of the brca1 protein have different effects in the two related dna repair pathways , and these results validate the hdr assay as highly correlative with single - and mutation studies . $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 1.0 **Cosine Similarity** = 0.56 $SemS$ = 0.8	genetic variants of uncertain clinical significance in the brca1 gene are a vexing problem in genetic counseling for women who have a family history of breast cancer . in this study , we evaluated the functions of 29 missense variants of brca1 in two dna repair pathways . repair of double - strand breaks by homology - directed recombination (hdr) had been previously analyzed for 16 of these brca1 variants , and 13 more variants were analyzed in this study . all 29 variants were also analyzed for function in double - strand break repair by the single - strand annealing (ssa) pathway . we found that among the 28 variants in brca1 , all were defective for dna repair by either pathway . the hdr assay was accurate because all 15 mutants were defective for hdr , and all nonpathogenic variants were fully functional for hdr . repair by ssa accurately identified functional mutants , but several nonpathogenic mutations were scored as defective or partially defective . these results indicated that specific amino acid residues of the brca1 protein have different effects in the two related dna repair pathways , and these results validate the hdr assay as highly useful with single - associated breast cancer . $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 1.0 **Cosine Similarity** = 0.73 $SemS$ = 0.81

	Vanilla	Adversarial	FAR-IG
Original	the punica granatum l . var . granatum (pomegranate) has been demonstrated to exert antitumor effects on various types of cancer cells . the present study aimed to evaluate the medicinal herbs punica granatum l . var . spinosa (apple punice) that are native to iran . this study was determined to test the possible cytotoxic activity and induction of apoptosis on human prostate cell lines . the effect of ethanol extracts of the herbs on the inhibition of cell proliferation was assessed by mtt colorimetric assay . pc3 cell lines treated with the extracts were analyzed for the induction of apoptosis by cell death detection (elisa) and tunel assay . dye exclusion analysis was performed for viability rate . our results demonstrated that the punica granatum l . var . spinosa extract dose dependently suppressed the proliferation of pc3 cells (ic (50) = 250 . 21 μg / ml) when compared with a chemotherapeutic anticancer drug (toxol) (vesper pharmaceuticals) with increased nucleosome production from apoptotic cells . the punica granatum l . var . spinosa extract attenuated the human prostate cell proliferation in vitro possibly by inducing apoptosis . the punica granatum l . var . spinosa is likely to be valuable for the treatment of some forms of human prostate cell line .	the punica granatum l . var . granatum (pomegranate) has been demonstrated to exert antitumor effects on various types of cancer cells . the present study aimed to evaluate the medicinal herbs punica granatum l . var . spinosa (apple punice) that are native to iran . this study was determined to test the possible cytotoxic activity and induction of apoptosis on human prostate cell lines . the effect of ethanol extracts of the herbs on the inhibition of cell proliferation was assessed by mtt colorimetric assay . pc3 cell lines treated with the extracts were analyzed for the induction of apoptosis by cell death detection (elisa) and tunel assay . dye exclusion analysis was performed for viability rate . our results demonstrated that the punica granatum l . var . spinosa extract dose dependently suppressed the proliferation of pc3 cells (ic (50) = 250 . 21 μg / ml) when compared with a chemotherapeutic anticancer drug (toxol) (vesper pharmaceuticals) with increased nucleosome production from apoptotic cells . the punica granatum l . var . spinosa extract attenuated the human prostate cell proliferation in vitro possibly by inducing apoptosis . the punica granatum l . var . spinosa is likely to be valuable for the treatment of some forms of human prostate cell line .	the punica granatum l . var . granatum (pomegranate) has been demonstrated to exert antitumor effects on various types of cancer cells . the present study aimed to evaluate the medicinal herbs punica granatum l . var . spinosa (apple punice) that are native to iran . this study was determined to test the possible cytotoxic activity and induction of apoptosis on human prostate cell lines . the effect of ethanol extracts of the herbs on the inhibition of cell proliferation was assessed by mtt colorimetric assay . pc3 cell lines treated with the extracts were analyzed for the induction of apoptosis by cell death detection (elisa) and tunel assay . dye exclusion analysis was performed for viability rate . our results demonstrated that the punica granatum l . var . spinosa extract dose dependently suppressed the proliferation of pc3 cells (ic (50) = 250 . 21 μg / ml) when compared with a chemotherapeutic anticancer drug (toxol) (vesper pharmaceuticals) with increased nucleosome production from apoptotic cells . the punica granatum l . var . spinosa extract attenuated the human prostate cell proliferation in vitro possibly by inducing apoptosis . the punica granatum l . var . spinosa is likely to be valuable for the treatment of some forms of human prostate cell line .
	$F(\boldsymbol{s}, \hat{l}=$"<multilabel>"$) = 1.0$	$F(\boldsymbol{s}, \hat{l}=$"<multilabel>"$) = 1.0$	$F(\boldsymbol{s}, \hat{l}=$"<multilabel>"$) = 1.0$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	the punica granatum l . var . granatum (pomegranate) has been demonstrated to exert antitumor effects on various types of cancer cells . the present study aimed to evaluate the medicinal herbs punica granatum l . var . spinosa (apple punice) that are native to iran . this study was determined to test the possible antioxidant activity and induction of cytotoxicity on human pca cell lines . the effect of ethanol extracts of the herbs on the inhibition of cell proliferation was assessed by mtt colorimetric assay . pc3 cell lines treated with the extracts were analyzed for the induction of p53 by cell apoptosis detection (elisa) and tunel assay . dye exclusion analysis was performed for viability rate . our results demonstrated that the punica granatum l . var . spinosa extract dose dependently suppressed the proliferation of pc3 cells (ic (50) = 250 . 21 micrograms / ml) when compared with a chemotherapeutic model drug (toxol) (vesper pharmaceuticals) with increased nucleosome number from apoptotic cells . the punica granatum l . var . spinosa extract attenuated the human pc cell proliferation in vitro possibly by inducing pge2 . the punica granatum l . var . spinosa is likely to be valuable for the treatment of some forms of human pc cell line . $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 1.0 **Cosine Similarity** = 0.32 $SemS$ = 0.93	the punica granatum l . var . granatum (pomegranate) has been demonstrated to exert antitumor effects on various types of cancer cells . the present study aimed to evaluate the medicinal herbs punica granatum l . var . spinosa (apple punice) that are native to iran . this study was determined to test the possible antioxidant activity and induction of cytotoxicity on human pc3 cell lines . the effect of ethanol extracts of the leaves on the inhibition of cell proliferation was assessed by mtt colorimetric assay . pc3 cell lines treated with the extracts were analyzed for the induction of caspases by cell death detection (elisa) and ldh assay . dye viability analysis was performed for viability rate . our results demonstrated that the punica granatum l . var . spinosa extract dose dependently suppressed the proliferation of pc3 cells (ic (50) = 250 . 21 μg / ml) when compared with a chemotherapeutic anticancer drug (toxol) (vesper pharmaceuticals) with increased apoptosis death from apoptotic cells . the punica granatum l . var . spinosa extract attenuated the human prostate cell proliferation in vitro possibly by inducing p53 . the punica granatum l . var . spinosa is likely to be valuable for the treatment of some forms of human prostatic cell line . $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 1.0 **Cosine Similarity** = 0.44 $SemS$ = 0.84	the punica granatum l . var . granatum (pomegranate) has been demonstrated to exert antitumor effects on various types of cancer cells . the present study aimed to evaluate the medicinal properties punica granatum l . var . spinosa (citrus punice) that are native to india . this study was determined to test the possible cytotoxic activity and induction of differentiation on human tumor cell lines . the effect of ethanol extracts of the fruit on the inhibition of cell proliferation was assessed by mtt colorimetric assay . pc3 cell lines treated with the extracts were analyzed for the induction of apoptosis by cell death detection (elisa) and immunoblot assay . dye exclusion analysis was performed for viability rate . our results demonstrated that the punica granatum l . var . spinosa extract dose dependently suppressed the proliferation of pc3 cells (ic (50) = 250 . 21 μg / ml) when compared with a chemotherapeutic anticancer drug (toxol) (vesper pharmaceuticals) with increased apoptosis death from apoptotic cells . the punica granatum l . var . spinosa extract attenuated the human pc3 cell proliferation in vitro possibly by inducing apoptosis . the punica granatum l . var . spinosa is likely to be valuable for the treatment of some forms of human pc3 cell line . $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 1.0 **Cosine Similarity** = 0.54 $SemS$ = 0.8

	VANILLA	ADVERSARIAL	FAR-IG
Original	objective although downregulation of neural cell adhesion molecule (ncam) has been correlated with poor prognosis in colorectal cancer (crc) , it is also possible that colon cancer spreading comes from reducing tumor cell adhesion through ncam polysialylation , as occurs in lung carcinoma or wilms ' tumor . methods to prove this hypothesis , we have performed a prospective study on tumor and control specimens from 39 crc patients , which were immunostained for ncam and psa (polysialic acid) expression . results tumor versus control expression of ncam and psa epitopes in tissue specimens , as well as correlation between tumor expression and clinicopathological features , were statistically analyzed . results showed a low constitutive expression of ncam and psa (psa - ncam) in control tissue , which reached a statistically significant increase in the tumor tissue . likewise , the presence and number of lymph node metastases at surgery were correlated with ncam expression and psa / ncam coexpression . conclusions these data highlight the importance of taking into account psa - associated epitopes when dealing with ncam cell expression studies in tumor development and progression . the analysis of psa and ncam expression in crc suggests a new way , other than downregulation of ncam , in order to escape contact inhibition and promote cell tumor spreading in colorectal cancer . $F(\boldsymbol{s}, \hat{l}$ ="<multilabel>") = 1.0	objective although downregulation of neural cell adhesion molecule (ncam) has been correlated with poor prognosis in colorectal cancer (crc) , it is also possible that colon cancer spreading comes from reducing tumor cell adhesion through ncam polysialylation , as occurs in lung carcinoma or wilms ' tumor . methods to prove this hypothesis , we have performed a prospective study on tumor and control specimens from 39 crc patients , which were immunostained for ncam and psa (polysialic acid) expression . results tumor versus control expression of ncam and psa epitopes in tissue specimens , as well as correlation between tumor expression and clinicopathological features , were statistically analyzed . results showed a low constitutive expression of ncam and psa (psa - ncam) in control tissue , which reached a statistically significant increase in the tumor tissue . likewise , the presence and number of lymph node metastases at surgery were correlated with ncam expression and psa / ncam coexpression . conclusions these data highlight the importance of taking into account psa - associated epitopes when dealing with ncam cell expression studies in tumor development and progression . the analysis of psa and ncam expression in crc suggests a new way , other than downregulation of ncam , in order to escape contact inhibition and promote cell tumor spreading in colorectal cancer . $F(\boldsymbol{s}, \hat{l}$ ="<multilabel>") = 1.0	objective although downregulation of neural cell adhesion molecule (ncam) has been correlated with poor prognosis in colorectal cancer (crc) , it is also possible that colon cancer spreading comes from reducing tumor cell adhesion through ncam polysialylation , as occurs in lung carcinoma or wilms ' tumor . methods to prove this hypothesis , we have performed a prospective study on tumor and control specimens from 39 crc patients , which were immunostained for ncam and psa (polysialic acid) expression . results tumor versus control expression of ncam and psa epitopes in tissue specimens , as well as correlation between tumor expression and clinicopathological features , were statistically analyzed . results showed a low constitutive expression of ncam and psa (psa - ncam) in control tissue , which reached a statistically significant increase in the tumor tissue . likewise , the presence and number of lymph node metastases at surgery were correlated with ncam expression and psa / ncam coexpression . conclusions these data highlight the importance of taking into account psa - associated epitopes when dealing with ncam cell expression studies in tumor development and progression . the analysis of psa and ncam expression in crc suggests a new way , other than downregulation of ncam , in order to escape contact inhibition and promote cell tumor spreading in colorectal cancer . $F(\boldsymbol{s}, \hat{l}$ ="<multilabel>") = 0.99

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	objective although downregulation of neuronal neural cell molecule (ncam) has been correlated with poor prognosis in colorectal cancer (crc) , it is also possible that colon cancer aggressiveness comes from reducing tumor cell adhesion through ncam polysialylation , as occurs in lung carcinoma or wilms ' tumor . methods to prove this hypothesis , we have performed a prospective study on tumor and control specimens from 39 crc patients , which were immunostained for ncam and pa (polysialic acid) expression . results tumor versus control expression of ncam and psa epitopes in tissue specimens , as well as correlation between tumor expression and clinicopathological features , were statistically analyzed . results showed a low constitutive expression of ncam and psa (anti - ncam) in control tissue , which reached a statistically significant increase in the tumor tissue . likewise , the presence and number of regional node metastases at surgery were correlated with ncam expression and psa / ncam coexpression . conclusions these data highlight the importance of taking into account psa - associated epitopes when dealing with ncam cell expression studies in tumor development and progression . the analysis of psa and ncam expression in crc suggests a new way , other than downregulation of ncam , in order to escape nc inhibition and thus cell cell spread in colorectal cancer .	objective although downregulation of neuronal cell adhesion molecule (ncam) has been correlated with poor prognosis in colorectal cancer (crc) , it is also possible that colon cancer spreading comes from aberrant tumor cell adhesion through nc adhesion , as occurs in lung carcinoma or wilms ' tumor . methods to prove this hypothesis , we have performed a prospective study on tumor and control specimens from 39 crc patients , which were immunostained for ncam and psma (polysialic acid) expression . results tumor versus control expression of ncam and specific epitopes in tissue specimens , as well as correlation between tumor expression and clinicopathological features , were statistically analyzed . results showed a low constitutive expression of ncam and psa (serine - ncam) in control tissue , which reached a statistically significant increase in the tumor tissue . likewise , the presence and number of lymph node metastases at surgery were correlated with ncam expression and psa / ncam coexpression . conclusions these data highlight the importance of taking into account cell - associated epitopes when dealing with ncam cell expression studies in tumor development and progression . the analysis of pca and ncam expression in crc suggests a new way , other than downregulation of ncam , in order to escape the metastasis and promote cell tumor spreading in colorectal cancer .	objective although downregulation of cell - activation molecule (psa) has been correlated with poor prognosis in colorectal cancer (crc) , it is also possible that colon cancer metastasis comes from reducing tumor cell proliferation through antigen upregulation , as occurs in lung carcinoma or wilms ' tumor . methods to prove this hypothesis , we have performed a prospective study on tumor and control specimens from 39 crc patients , which were immunostained for ncam and protein (polysialic acid) expression . results tumor versus control expression of ncam and psa epitopes in tissue specimens , as well as correlation between tumor expression and clinicopathological features , were statistically analyzed . results showed a low constitutive expression of ncam and psa (psa - ncam) in control tissue , which reached a statistically significant increase in the tumor tissue . likewise , the presence and number of lymph node metastases at surgery were correlated with ncam expression and psa / ncam coexpression . conclusions these data highlight the importance of taking into account psa - associated epitopes when dealing with ncam cell expression studies in tumor development and progression . the analysis of cd44 and ncam expression in crc suggests a new way , other than downregulation of ncam , in order to escape contact inhibition and promote cell tumor growth in colorectal cancer .
	$F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 1.0	$F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 0.98	$F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 0.99
	Cosine Similarity = -0.16	**Cosine Similarity** = 0.19	**Cosine Similarity** = 0.37
	$SemS = 0.74$	$SemS = 0.7$	$SemS = 0.7$

	VANILLA	ADVERSARIAL	FAR-IG
Original	objective to analyze histological factors not routinely assessed as potential prognostic factors in renal cell carcinoma , such as tumor necrosis , microscopic vascular invasion , and sinus fat invasion . materials and methods a retrospective , analytical study was conducted of surgical specimens from 139 patients with localized renal cell carcinoma who underwent nephrectomy from 1993 to 2005 . tumor necrosis , microscopic vascular invasion , and sinus fat invasion were analyzed and compared to the classical factors : tnm classification , fuhrman grade , and tumor size . for statistical analysis , variables analyzed were categorized as pt1 , 2 vs pt3 , 4 ; fuhrman grade 1 , 2 vs 3 , 4 ; tumor size < 7 cm vs > or = 7cm ; tumor necrosis vs no tumor necrosis ; microvascular invasion of sinus fat vs no invasion . cancer - specific survival probability and disease - free survival were calculated . a descriptive and analytical statistical analysis was performed using logistic regression for univariate and multivariate analyses . dependent variables were used to analyze cancer - specific survival rates . disease - free survival was estimated using a cox regression model and kaplan - meier curves . results in the univariate analysis , all variables analyzed had a significant influence on death for renal cell carcinoma . in the multivariate analysis , the variable having the greatest influence was fuhrman grade (p = 0 , 032) . $F(\boldsymbol{s}, \hat{l} =$ "<multilabel>"$) = 1.0$	objective to analyze histological factors not routinely assessed as potential prognostic factors in renal cell carcinoma , such as tumor necrosis , microscopic vascular invasion , and sinus fat invasion . materials and methods a retrospective , analytical study was conducted of surgical specimens from 139 patients with localized renal cell carcinoma who underwent nephrectomy from 1993 to 2005 . tumor necrosis , microscopic vascular invasion , and sinus fat invasion were analyzed and compared to the classical factors : tnm classification , fuhrman grade , and tumor size . for statistical analysis , variables analyzed were categorized as pt1 , 2 vs pt3 , 4 ; fuhrman grade 1 , 2 vs 3 , 4 ; tumor size < 7 cm vs > or = 7cm ; tumor necrosis vs no tumor necrosis ; microvascular invasion of sinus fat vs no invasion . cancer - specific survival probability and disease - free survival were calculated . a descriptive and analytical statistical analysis was performed using logistic regression for univariate and multivariate analyses . dependent variables were used to analyze cancer - specific survival rates . disease - free survival was estimated using a cox regression model and kaplan - meier curves . results in the univariate analysis , all variables analyzed had a significant influence on death for renal cell carcinoma . in the multivariate analysis , the variable having the greatest influence was fuhrman grade (p = 0 , 032) . $F(\boldsymbol{s}, \hat{l} =$ "<multilabel>"$) = 1.0$	objective to analyze histological factors not routinely assessed as potential prognostic factors in renal cell carcinoma , such as tumor necrosis , microscopic vascular invasion , and sinus fat invasion . materials and methods a retrospective , analytical study was conducted of surgical specimens from 139 patients with localized renal cell carcinoma who underwent nephrectomy from 1993 to 2005 . tumor necrosis , microscopic vascular invasion , and sinus fat invasion were analyzed and compared to the classical factors : tnm classification , fuhrman grade , and tumor size . for statistical analysis , variables analyzed were categorized as pt1 , 2 vs pt3 , 4 ; fuhrman grade 1 , 2 vs 3 , 4 ; tumor size < 7 cm vs > or = 7cm ; tumor necrosis vs no tumor necrosis ; microvascular invasion of sinus fat vs no invasion . cancer - specific survival probability and disease - free survival were calculated . a descriptive and analytical statistical analysis was performed using logistic regression for univariate and multivariate analyses . dependent variables were used to analyze cancer - specific survival rates . disease - free survival was estimated using a cox regression model and kaplan - meier curves . results in the univariate analysis , all variables analyzed had a significant influence on death for renal cell carcinoma . in the multivariate analysis , the variable having the greatest influence was fuhrman grade (p = 0 , 032) . $F(\boldsymbol{s}, \hat{l} =$ "<multilabel>"$) =$ 0.93

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	objective to analyze clinicopathological factors not routinely assessed as potential prognostic factors in clear cell carcinoma , such as tumor necrosis , microscopic venous permeation , and sinus fat involvement . materials and methods a retrospective , analytical study was conducted of surgical specimens from 139 patients with localized renal cell carcinoma who underwent nephrectomy from 1993 to 2005 . tumor invasion , microscopic tumor invasion , and sinus fat invasion were analyzed and compared to the classical factors : tnm classification , fuhrman grade , and tumor size . for statistical analysis , variables analyzed were categorized as pt1 , 2 vs pt3 , 4 ; fuhrman grade 1 , 2 vs 3 , 4 ; tumor size < 7 cm vs > or = 7cm ; tumor thrombus vs no microscopic invasion ; invasion presence of sinus fat vs no invasion . cancer - specific survival probability and disease - free survival were calculated . a descriptive and analytical statistical analysis was performed using logistic regression for univariate and multivariate analyses . dependent variables were used to analyze cancer - specific survival rates . disease - free survival was estimated using a cox regression model and kaplan - meier curves . results in the univariate analysis , all variables analyzed had a significant influence on death for renal cell carcinoma . in the multivariate analysis , the variable having the greatest influence was fuhrman grade (p = 0 , 032) . $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 0.99 **Cosine Similarity** = 0.09 $SemS$ = 0.66	objective to analyze several factors not routinely assessed as potential prognostic factors in renal cell carcinoma , such as tumor necrosis , microscopic vascular invasion , and lymph fat invasion . materials and methods a retrospective , analytical study was conducted of surgical specimens from 139 patients with localized renal cell carcinoma who underwent nephrectomy from 1993 to 2005 . tumor invasion , microscopic vascular permeation , and microscopic vascular invasion were analyzed and compared to the classical factors : histological classification , fuhrman grade , and tumor size . for statistical analysis , variables analyzed were categorized as pt1 , 2 vs pt3 , 4 ; fuhrman grade 1 , 2 vs 3 , 4 ; tumor size < 7 cm vs > or = 7cm ; tumor infiltration vs no tumor permeation ; invasion lymph of vascular fat vs no invasion . cancer - specific survival probability and disease - free survival were calculated . a descriptive and analytical statistical analysis was performed using logistic regression for univariate and multivariate analyses . dependent variables were used to analyze cancer - specific survival rates . disease - free survival was estimated using a cox regression model and kaplan - meier curves . results in the univariate analysis , all variables analyzed had a significant influence on death for renal cell carcinoma . in the multivariate analysis , the variable having the greatest influence was fuhrman grade (p = 0 , 032) . $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 0.97 **Cosine Similarity** = 0.28 $SemS$ = 0.73	objective to analyze histological factors not routinely assessed as potential prognostic factors in renal cell carcinoma , such as tumor necrosis , microscopic vascular invasion , and lymph lymphatic invasion . materials and methods a retrospective , analytical study was conducted of surgical specimens from 139 patients with localized renal cell carcinoma who underwent lymphadenectomy from 1993 to 2005 . tumor metastasis , microscopic vascular invasion , and lymph fat metastases were analyzed and compared to the classical factors : histological classification , fuhrman grade , and tumor size . for statistical analysis , variables analyzed were categorized as pt1 , 2 vs pt3 , 4 ; fuhrman grade 1 , 2 vs 3 , 4 ; tumor size < 7 cm vs > or = 7cm ; tumor necrosis vs no tumor necrosis ; neph lymph of vascular necrosis vs no necrosis . cancer - specific survival probability and disease - free survival were calculated . a descriptive and analytical statistical analysis was performed using logistic regression for univariate and multivariate analyses . dependent variables were used to analyze cancer - specific survival rates . disease - free survival was estimated using a cox regression model and kaplan - meier curves . results in the univariate analysis , all variables analyzed had a significant influence on death for renal cell carcinoma . in the multivariate analysis , the variable having the greatest influence was fuhrman grade (p = 0 , 032) . $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 0.96 **Cosine Similarity** = 0.62 $SemS$ = 0.67

	Vanilla	Adversarial	FAR-IG
Original	dna - protein cross - links (dpcs) are formed upon exposure to a variety of chemical and physical agents and pose a threat to genomic integrity . in particular , acrolein and related aldehydes produce dpcs , although the chemical linkages for such cross - links have not been identified . here , we report that oligodeoxynucleotides containing 1 , n (2) - deoxyguanosine adducts of acrolein , crotonaldehyde , and trans - 4 - hydroxynonenal can form cross - links with the tetrapeptide lys - trp - lys - lys . we concluded that complex formation is mediated by a schiff base linkage because dna - peptide complexes were covalently trapped following reduction with sodium cyanoborohydride , and pre - reduction of adducted dnas inhibited complex formation . a previous nmr study demonstrated that duplex dna catalyzes ring opening for the acrolein - derived gamma - hydroxy - 1 , n (2) - propanodeoxyguanosine adduct to yield an aldehydic function (de los santos , c . , zaliznyak , t . , and johnson , f . (2001) j . biol . chem . 276 , 9077 - 9082) . consistent with this earlier observation , the adducts under investigation were more reactive in duplex dna than $F(\boldsymbol{s}, \hat{l}=$"<multilabel>"$) = 1.0$	dna - protein cross - links (dpcs) are formed upon exposure to a variety of chemical and physical agents and pose a threat to genomic integrity . in particular , acrolein and related aldehydes produce dpcs , although the chemical linkages for such cross - links have not been identified . here , we report that oligodeoxynucleotides containing 1 , n (2) - deoxyguanosine adducts of acrolein , crotonaldehyde , and trans - 4 - hydroxynonenal can form cross - links with the tetrapeptide lys - trp - lys - lys . we concluded that complex formation is mediated by a schiff base linkage because dna - peptide complexes were covalently trapped following reduction with sodium cyanoborohydride , and pre - reduction of adducted dnas inhibited complex formation . a previous nmr study demonstrated that duplex dna catalyzes ring opening for the acrolein - derived gamma - hydroxy - 1 , n (2) - propanodeoxyguanosine adduct to yield an aldehydic function (de los santos , c . , zaliznyak , t . , and johnson , f . (2001) j . biol . chem . 276 , 9077 - 9082) . consistent with this earlier observation , the adducts under investigation were more reactive in duplex dna than $F(\boldsymbol{s}, \hat{l}=$"<multilabel>"$) = 1.0$	dna - protein cross - links (dpcs) are formed upon exposure to a variety of chemical and physical agents and pose a threat to genomic integrity . in particular , acrolein and related aldehydes produce dpcs , although the chemical linkages for such cross - links have not been identified . here , we report that oligodeoxynucleotides containing 1 , n (2) - deoxyguanosine adducts of acrolein , crotonaldehyde , and trans - 4 - hydroxynonenal can form cross - links with the tetrapeptide lys - trp - lys - lys . we concluded that complex formation is mediated by a schiff base linkage because dna - peptide complexes were covalently trapped following reduction with sodium cyanoborohydride , and pre - reduction of adducted dnas inhibited complex formation . a previous nmr study demonstrated that duplex dna catalyzes ring opening for the acrolein - derived gamma - hydroxy - 1 , n (2) - propanodeoxyguanosine adduct to yield an aldehydic function (de los santos , c . , zaliznyak , t . , and johnson , f . (2001) j . biol . chem . 276 , 9077 - 9082) . consistent with this earlier observation , the adducts under investigation were more reactive in duplex dna than $F(\boldsymbol{s}, \hat{l}=$"<multilabel>"$) = 1.0$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	dna - protein cross - links (dpcs) are formed upon exposure to a variety of chemical and physical agents and pose a threat to genomic integrity . in particular , acrolein and related molecules produce dpcs , although the chemical linkages for such cross - links have not been identified . here , we report that oligodeoxynucleotides containing 1 , n (2) - diol esters of acrolein , crotonaldehyde , and trans - 4 - hydroxynonenal can form cross - linking with the tetrapeptide lys - trp - lys - lys . we concluded that complex formation is mediated by a schiff base linkage because dna - peptide complexes were selectively trapped following reduction with sodium cyanoborohydride , and pre - reduction of adducted dnas inhibited complex formation . a previous nmr study demonstrated that duplex dna catalyzes ring opening for the acrolein - derived gamma - hydroxy - 1 , n (2) - propanodeoxyguanosine radical to yield an aldehydic function (de los santos , c . , zaliznyak , t . , and johnson , f . (2001) j . biol . chem . 276 , 9077 - 9082) . consistent with this earlier observation , the peptides under investigation were more efficiently in this buffer than $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"<multilabel>"}) = 1.0$ **Cosine Similarity** = 0.33 $SemS = 0.62$	dna - protein cross - links (dpcs) are formed upon exposure to a variety of chemical and physical agents and pose a threat to genomic integrity . in particular , acrolein and related molecules produce dpcs , although the chemical linkages for such cross - links have not been identified . here , we report that oligodeoxynucleotides containing 1 , n (2) - deoxyguanosine analogues of acrolein , crotonaldehyde , and trans - 4 - hydroxynonenal can form cross - linkages with the sequences lys - trp - lys - lys . we concluded that complex formation is mediated by a single base linkage because dna - peptide complexes were not trapped following reduction with sodium azide , and pre - reduction of adducted dnas inhibited complex formation . a previous nmr study demonstrated that duplex dna catalyzes the opening for the acrolein - derived gamma - hydroxy - 1 , n (2) - propanodeoxyguanosine molecule to yield an aldehydic function (de los santos , c . , zaliznyak , t . , and johnson , f . (2001) j . biol . chem . 276 , 9077 - 9082) . consistent with this earlier observation , the bases under investigation were more reactive in duplex dna than $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"<multilabel>"}) = 1.0$ **Cosine Similarity** = 0.54 $SemS = 0.64$	dna - protein cross - links (dpcs) are formed upon exposure to a variety of chemical and physical agents and pose a threat to genomic integrity . in particular , acrolein and related adducts produce dpcs , although the chemical linkages for such cross - links have not been identified . here , we report that oligodeoxynucleotides containing 1 , n (2) - deoxyguanosine conjugates of acrolein , crotonaldehyde , and trans - 4 - hydroxynonenal can form cross - links with the sequences lys - lys - lys - lys . we concluded that dpc formation is mediated by a dna base linkage because dna - adduct sites were covalently trapped following reduction with sodium cyanoborohydride , and pre - reduction of adducted dnas inhibited complex formation . a previous nmr study demonstrated that duplex dna catalyzes ring opening for the acrolein - derived gamma - keto - 1 , n (2) - propanodeoxyguanosine adduct to yield an aldehydic function (de los santos , c . , zaliznyak , t . , and johnson , f . (2001) j . biol . chem . 276 , 9077 - 9082) . consistent with this earlier observation , the linkages under investigation were more reactive in duplex dna than $F(\boldsymbol{s}_{\text{adv}}, \hat{l} = \text{"<multilabel>"}) = 1.0$ **Cosine Similarity** = 0.66 $SemS = 0.67$

	VANILLA	ADVERSARIAL	FAR-IG
Original	verbascum thapsus commonly known as ' mullein ' is part of a large family of scrophulariaceae consisting of more than 360 species . from antiquity verbascum thapsus has been used as a medicinal herb , it contains diverse polysaccharides , iroid glycosides , flavonoids , saponins , volatile oils and phenylentanoids . inducible nitric oxide synthase (inos) represents one of the three isoforms that produce nitric oxide using l - arginine as a substrate in response to an increase in superoxide anion activated by nf - kb . it is implicated in different pathophysiological events and its expression increases greatly during an inflammatory process , due to oxidative stress and the activation of the enzymes of the antioxidant network such as sod , cat and gpx . in this study an inflammatory state was reproduced by treating thp - 1 cells (human myelomonocytic leukaemia) with pro - inflammatory stimuli , such as lps and ifn - gamma , obtaining an up - regulation both in the expression and in the activity of inos . the aim of the work was to investigate the antiinflammatory action of verbascoside using a concentration of 100 mum . the results show a significant decrease of the expression and activity of inos , extracellular o (2) (-) production , sod , cat and gpx activity when the cells were treated	verbascum thapsus commonly known as ' mullein ' is part of a large family of scrophulariaceae consisting of more than 360 species . from antiquity verbascum thapsus has been used as a medicinal herb , it contains diverse polysaccharides , iroid glycosides , flavonoids , saponins , volatile oils and phenylentanoids . inducible nitric oxide synthase (inos) represents one of the three isoforms that produce nitric oxide using l - arginine as a substrate in response to an increase in superoxide anion activated by nf - kb . it is implicated in different pathophysiological events and its expression increases greatly during an inflammatory process , due to oxidative stress and the activation of the enzymes of the antioxidant network such as sod , cat and gpx . in this study an inflammatory state was reproduced by treating thp - 1 cells (human myelomonocytic leukaemia) with pro - inflammatory stimuli , such as lps and ifn - gamma , obtaining an up - regulation both in the expression and in the activity of inos . the aim of the work was to investigate the antiinflammatory action of verbascoside using a concentration of 100 mum . the results show a significant decrease of the expression and activity of inos , extracellular o (2) (-) production , sod , cat and gpx activity when the cells were treated	verbascum thapsus commonly known as ' mullein ' is part of a large family of scrophulariaceae consisting of more than 360 species . from antiquity verbascum thapsus has been used as a medicinal herb , it contains diverse polysaccharides , iroid glycosides , flavonoids , saponins , volatile oils and phenylentanoids . inducible nitric oxide synthase (inos) represents one of the three isoforms that produce nitric oxide using l - arginine as a substrate in response to an increase in superoxide anion activated by nf - kb . it is implicated in different pathophysiological events and its expression increases greatly during an inflammatory process , due to oxidative stress and the activation of the enzymes of the antioxidant network such as sod , cat and gpx . in this study an inflammatory state was reproduced by treating thp - 1 cells (human myelomonocytic leukaemia) with pro - inflammatory stimuli , such as lps and ifn - gamma , obtaining an up - regulation both in the expression and in the activity of inos . the aim of the work was to investigate the antiinflammatory action of verbascoside using a concentration of 100 mum . the results show a significant decrease of the expression and activity of inos , extracellular o (2) (-) production , sod , cat and gpx activity when the cells were treated
	$F(\boldsymbol{s}, \hat{l} =$"<multilabel>"$) = 1.0$	$F(\boldsymbol{s}, \hat{l} =$"<multilabel>"$) = 1.0$	$F(\boldsymbol{s}, \hat{l} =$"<multilabel>"$) = 1.0$

	VANILLA	ADVERSARIAL	FAR-IG
Adversarial	verbascum thapsus commonly known as ' mullein ' is part of a large family of scrophulariaceae consisting of more than 360 species . from antiquity verbascum thapsus has been used as a medicinal herb , it contains diverse polysaccharides , iroid glycosides , flavonoids , saponins , volatile oils and phenylentanoids . inducible nitric oxide synthase (its) represents one of the three isoforms that produce nitric oxide using l - arginine as a substrate in response to an increase in radical anion activated by nf - kb . it is implicated in different pathophysiological events and its expression increases greatly during an activation process , due to redox stress and the activation of the enzymes of the antioxidant network such as sod , cat and gpx . in this study an activation model was reproduced by treating l - 1 cells (human myelomonocytic leukaemia) with pro - oxidant stimuli , such as pma and ifn - gamma , obtaining an up - regulation both in the expression and in the activity of inos . the aim of the work was to investigate the inhibitory action of verbascoside using a concentration of 100 mum . the results show a significant decrease of the expression and activity of inos , extracellular o (2) (-) production , sod , cat and gpx activity when the stimulation were treated $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 0.9 **Cosine Similarity** = -0.01 $SemS$ = 0.62	verbascum thapsus commonly known as ' mullein ' is part of a large family of scrophulariaceae consisting of more than 360 species . from antiquity verbascum thapsus has been used as a medicinal herb , it contains diverse polysaccharides , iroid glycosides , flavonoids , saponins , volatile oils and phenylentanoids . inducible no oxide synthase (enos) represents one of the three isoforms that produce adenosine oxide using l - tyrosine as a substrate in response to an increase in inflammatory , activated by nf - kb . it is implicated in different pathophysiological events and its expression increases greatly during an inflammatory process , due to oxidative stress and the activation of the enzymes of the antioxidant network such as sod , cat and gpx . in this study an inflamed state was reproduced by treating thp - 1 cells (human myelomonocytic leukaemia) with pro - inflammatory stimuli , such as lps and ifn - gamma , obtaining an up - regulation both in the expression and in the activity of is . the aim of the work was to investigate the antioxidant action of verbascoside using a concentration of 100 mum . the results show a significant decrease of the expression and activity of nnos , extracellular no (2) (-) production , sod , cat and gpx activity when the cells were treated $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 0.98 **Cosine Similarity** = 0.4 $SemS$ = 0.62	verbascum thapsus commonly known as ' mullein ' is part of a large family of scrophulariaceae consisting of more than 360 species . from antiquity verbascum thapsus has been used as a medicinal herb , it contains diverse polysaccharides , iroid glycosides , flavonoids , saponins , volatile oils and phenylentanoids . inducible inducible monoxide synthase (ias) represents one of the three isoforms that produce nitric oxide using l - tyrosine as a substrate in response to an increase in hydroxyl anion activated by nf - kb . it is implicated in different pathophysiological events and its expression increases greatly during an inflammatory process , due to oxidant stress and the activation of the enzymes of the oxidative network such as sod , cat and gpx . in this study an activation state was reproduced by treating thp - 1 cells (human myelomonocytic leukaemia) with pro - inflammatory stimuli , such as lps and ifn - gamma , obtaining an up - regulation both in the expression and in the activity of enos . the aim of the work was to investigate the antioxidant action of verbascoside using a concentration of 100 mum . the results show a significant decrease of the expression and activity of nnos , extracellular o (2) (-) production , sod , cat and gpx activity when the cells were treated $F(\boldsymbol{s}_{\text{adv}}, \hat{l} =$"<multilabel>") = 0.99 **Cosine Similarity** = 0.53 $SemS$ = 0.62

Bibliography

[1] A. Ghorbani, A. Abid and J. Zou, “Interpretation of Neural Networks is Fragile”, *AAAI Conference on Artificial Intelligence*, vol. 33, pp. 3681–3688, 2019.

[2] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation”, *Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[3] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf and M. Funtowicz, “Transformers: State-of-the-Art Natural Language Processing”, *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

[4] P. E. Brown, V. J. Della Pietra, S. A. Della Pietra and J. C. Lai, “An Estimate of an Upper Bound for the Entropy of English”, *Computational Linguistics*, vol. 18, no. 1.

[5] A. Das and P. Rad, “Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey”, *arXiv e-prints*, pp. arXiv–2006, 2020.

[6] T. Hosaka, “Bankruptcy Prediction Using Imaged Financial Ratios and Convolutional Neural Networks”, *Expert Systems with Applications*, vol. 117:pp. 287–299, 2019.

[7] D. Pessach and E. Shmueli, “A Review on Fairness in Machine Learning”, *ACM Computing Surveys (CSUR)*, vol. 55, no. 3:pp. 1–44, 2022.

[8] M. Sundararajan, A. Taly and Q. Yan, “Axiomatic Attribution for Deep Networks”, *International Conference on Machine Learning*, vol. 70, pp. 3319–3328, 2017.

[9] K. Simonyan, A. Vedaldi and A. Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, *arXiv preprint arXiv:1312.6034*, 2013.

[10] M. T. Ribeiro, S. Singh and C. Guestrin, “"Why Should I Trust You?" Explaining the Predictions of any Classifier”, *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[11] M. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks”, *European Conference on Computer Vision*, pp. 818–833, Springer, 2014.

[12] A. Nguyen, J. Yosinski and J. Clune, "Understanding Neural Networks via Feature Visualization: A Survey", *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 55–76, 2019.

[13] D. Bahdanau, K. H. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", *International Conference on Learning Representations*, 2015.

[14] D. Alvarez-Melis and T. S. Jaakkola, "Towards Robust Interpretability with Self-Explaining Neural Networks", *International Conference on Neural Information Processing Systems*, pp. 7786–7795, 2018.

[15] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt and B. Kim, "Sanity Checks for Saliency Maps", *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[16] C. Tan, "On the Diversity and Limits of Human Explanations", *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2173–2188, 2022.

[17] A. F. Markus, J. A. Kors and P. 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*, vol. 113:p. 103655, 2021.

[18] M. Rizzo, A. Veneri, A. Albarelli, C. Lucchese and C. Conati, "A Theoretical Framework for AI Models Explainability", *arXiv e-prints*, pp. arXiv–2212, 2022.

[19] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?", *Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020.

[20] A. Jacovi and Y. Goldberg, "Aligning Faithful Interpretations with their Social Attribution", *Transactions of the Association for Computational Linguistics*, vol. 9:pp. 294–310, 2021.

[21] S.-M. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

[22] D. Alvarez-Melis and T. S. Jaakkola, "On the Robustness of Interpretability Methods", *arXiv preprint arXiv:1806.08049*, 2018.

[23] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan and B. Kim, "The (un)reliability of Saliency Methods", *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280, Springer, 2019.

[24] N. Bansal, C. Agarwal and A. Nguyen, "SAM: The Sensitivity of Attribution Methods to Hyperparameters", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8670–8680, IEEE, 2020.

[25] L. Engstrom, A. Ilyas, H. Salman, S. Santurkar and D. Tsipras, "Robustness (Python Library)", 2019.

[26] X. Zhang, J. Zhao and Y. Lecun, "Character-Level Convolutional Networks for Text Classification", *Advances in Neural Information Processing Systems*, vol. 2015:pp. 649–657, 2015.

[27] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato and P. Frossard, "Robustness via Curvature Regularization, and vice versa", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

[28] C. Etmann, S. Lunz, P. Maass and C. Schoenlieb, "On the Connection Between Adversarial Robustness and Saliency Map Interpretability", *International Conference on Machine Learning*, pp. 1823–1832, Pmlr, 2019.

[29] M. Singh, N. Kumari, P. Mangla, A. Sinha, V. N. Balasubramanian and B. Krishnamurthy, "On the Benefits of Attributional Robustness", *arXiv preprint arXiv:1911.13073*, 2019.

[30] Z. Wang, H. Wang, S. Ramkumar, M. Fredrikson, P. Mardziel and A. Datta, "Smoothed Geometry for Robust Attribution", *International Conference on Neural Information Processing Systems*, pp. 13623–13634, 2020.

[31] A. Ivankay, I. Girardi, C. Marchiori and P. Frossard, "Fooling Explanations in Text Classifiers", *International Conference on Learning Representations*, 2022.

[32] S. Sinha, H. Chen, A. Sekhon, Y. Ji and Y. Qi, "Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing", *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 420–434, 2021.

[33] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[34] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI.", *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[35] G. Ras, N. Xie, M. van Gerven and D. Doran, "Explainable Deep Learning: A Field Guide for the Uninitiated", *Journal of Artificial Intelligence Research*, vol. 73:pp. 329–396, 2022.

[36] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina and R. Benjamins, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", *Information Fusion*, vol. 58:pp. 82–115, 2020.

[37] D. Smilkov, N. Thorat, B. Kim, F. Viégas and M. Wattenberg, "SmoothGrad: Removing Noise by Adding Noise", *arXiv preprint arXiv:1706.03825*, 2017.

[38] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh and D. Batra, "Grad-CAM: Why Did You Say That?", *arXiv preprint arXiv:1611.07450*, 2016.

[39] D. Marcos, S. Lobry and D. Tuia, "Semantically Interpretable Activation Maps: What-Where-How Explanations within CNNs", *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4207–4215, IEEE, 2019.

[40] A. Ghorbani, J. Wexler, J. Y. Zou and B. Kim, "Towards Automatic Concept-Based Explanations", *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[41] A. Shrikumar, P. Greenside and A. Kundaje, "Learning Important Features through Propagating Activation Differences", *International Conference on Machine Learning*, pp. 3145–3153, PMLR, 2017.

[42] M. Ancona, E. Ceolini, C. Öztireli and M. Gross, "Gradient-Based Attribution Methods", *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 169–191, 2019.

[43] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", *International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.

[44] G. Montavon, A. Binder, S. Lapuschkin, W. Samek and K.-R. Müller, "Layer-Wise Relevance Propagation: An Overview", *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 193–209, 2019.

[45] L. M. Zintgraf, T. S. Cohen, T. Adel and M. Welling, "Visualizing Deep Neural Network Decisions: Prediction Difference Analysis", *International Conference on Learning Representations*, 2017.

[46] J. Li, W. Monroe and D. Jurafsky, "Understanding Neural Networks through Representation Erasure", *arXiv preprint arXiv:1612.08220*, 2016.

[47] M. T. Ribeiro, S. Singh and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations.", *Aaai*, vol. 18, pp. 1527–1535, 2018.

[48] G. Montavon, S. Lapuschkin, A. Binder, W. Samek and K.-R. Müller, "Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition", *Pattern Recognition*, vol. 65:pp. 211–222, 2017.

[49] Y. Sawada and K. Nakamura, "C-SENN: Contrastive Self-Explaining Neural Network", *arXiv preprint arXiv:2206.09575*, 2022.

[50] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana and G. E. Hinton, "Neural Additive Models: Interpretable Machine Learning with Neural Nets", *Advances in Neural Information Processing Systems*, vol. 34:pp. 4699–4711, 2021.

[51] W. Nie, Y. Zhang and A. Patel, "A Theoretical Explanation for Perplexing Behaviors of Backpropagation-Based Visualizations", *International Conference on Machine Learning*, pp. 3809–3818, PMLR, 2018.

[52] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez and J. Boyd-Graber, "Pathologies of Neural Models Make Interpretations Difficult", *Conference on Empirical Methods in Natural Language Processing*, pp. 3719–3728, 2018.

[53] S. Jain and B. C. Wallace, "Attention is not Explanation", *Proceedings of NAACL-HLT*, pp. 3543–3556, 2019.

[54] S. Wiegreffe and Y. Pinter, "Attention is not not Explanation", *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pp. 11–20, Association for Computational Linguistics, 2020.

[55] S. Serrano and N. A. Smith, "Is Attention Interpretable?", *Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, 2019.

[56] J. Bastings and K. Filippova, "The Elephant in the Interpretability Room: Why use Attention as Explanation when we have Saliency Methods?", *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 149–155, 2020.

[57] C. Meister, S. Lazov, I. Augenstein and R. Cotterell, "Is Sparse Attention more Interpretable?", *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 122–129, 2021.

[58] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and Harnessing Adversarial Examples", *arXiv preprint arXiv:1412.6572*, 2014.

[59] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks", *IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, IEEE Computer Society, 2017.

[60] A. Modas, S.-M. Moosavi-Dezfooli and P. Frossard, "SparseFool: A Few Pixels Make a Big Difference", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9087–9096, 2019.

[61] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi and P. Frossard, "Universal Adversarial Perturbations", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765–1773, 2017.

[62] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks", *International Conference on Learning Representations*, 2018.

[63] C.-C. Tu, P. Ting, P.-Y. Chen, S. Liu, H. Zhang, J. Yi, C.-J. Hsieh and S.-M. Cheng, "Auto-ZOOM: Autoencoder-Based Zeroth Order Optimization Method for Attacking Black-Box Neural Networks", *AAAI Conference on Artificial Intelligence*, vol. 33, pp. 742–749, 2019.

[64] P. Yang, J. Chen, C.-J. Hsieh, J.-L. Wang and M. I. Jordan, "Greedy Attack and Gumbel Attack: Generating Adversarial Examples for Discrete Data", *Journal of Machine Learning Research*, vol. 21:pp. 1–36, 2020.

[65] L. Li, R. Ma, Q. Guo, X. Xue and X. Qiu, "BERT-ATTACK: Adversarial Attack Against BERT Using BERT", *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6193–6202, Association for Computational Linguistics, 2020.

[66] J. Ebrahimi, A. Rao, D. Lowd and D. Dou, "HotFlip: White-Box Adversarial Examples for Text Classification", *Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 31–36, 2018.

[67] D. Jin, Z. Jin, J. T. Zhou and P. Szolovits, "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment", *AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8018–8025, 2020.

[68] L. Sun, K. Hashimoto, W. Yin, A. Asai, J. Li, P. Yu and C. Xiong, "Adv-BERT: BERT is not Robust on Misspellings! Generating Nature Adversarial Samples on BERT", *arXiv preprint arXiv:2003.04985*, 2020.

[69] H. Gao and T. Oates, "Universal Adversarial Perturbation for Text Classification", *arXiv preprint arXiv:1910.04618*, 2019.

[70] B. Liang, H. Li, M. Su, P. Bian, X. Li and W. Shi, "Deep Text Classification can be fooled", *International Joint Conference on Artificial Intelligence*, pp. 4208–4215, 2018.

[71] J. Buckman, A. Roy, C. Raffel and I. Goodfellow, "Thermometer Encoding: One Hot Way to Resist Adversarial Examples", *International Conference on Learning Representations*, 2018.

[72] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin and N. Usunier, "Parseval Networks: Improving Robustness to Adversarial Examples", *International Conference on Machine Learning*, pp. 854–863, Pmlr, 2017.

[73] D. Slack, S. Hilgard, E. Jia, S. Singh and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation methods", *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

[74] A. Subramanya, V. Pillai and H. Pirsiavash, "Fooling Network Interpretation in Image Classification", *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2020–2029, IEEE, 2019.

[75] J. Heo, S. Joo and T. Moon, "Fooling Neural Network Interpretations via Adversarial Model Manipulation", *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[76] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller and P. Kessel, "Explanations can be Manipulated and Geometry is to blame", *Advances in Neural Information Processing Systems*, pp. 13589–13600, 2019.

[77] R. Tang, H. Chen and Y. Ji, "Identifying the Source of Vulnerability in Explanation Discrepancy: A Case Study in Neural Text Classification", *arXiv preprint arXiv:2212.05327*, 2022.

[78] A. Binder, M. Bockmayr, M. Hägele, S. Wienert, D. Heim, K. Hellweg, M. Ishii, A. Stenzinger, A. Hocke and C. Denkert, "Morphological and Molecular Breast Cancer Profiling through Explainable Machine Learning", *Nature Machine Intelligence*, vol. 3, no. 4:pp. 355–366, 2021.

[79] I. Girardi, P. Ji, A.-P. Nguyen, N. Hollenstein, A. Ivankay, L. Kuhn, C. Marchiori and C. Zhang, "Patient Risk Assessment and Warning Symptom Detection Using Deep Attention-Based Neural Networks", *International Workshop on Health Text Mining and Information Analysis*, pp. 139–148, 2018.

[80] J. Chen, X. Wu, V. Rastogi, Y. Liang and S. Jha, "Robust Attribution Regularization", *Advances in Neural Information Processing Systems*, pp. 14300–14310, 2019.

[81] L. Rieger and L. K. Hansen, "A Simple Defense Against Adversarial Attacks on Heatmap Explanations", *Annual Workshop on Human Interpretability in Machine Learning*, 2020.

[82] A. Levine, S. Singla and S. Feizi, "Certifiably Robust Interpretation in Deep Learning", *arXiv preprint arXiv:1905.12105*, 2019.

[83] A. Liu, X. Chen, S. Liu, L. Xia and C. Gan, "Certifiably Robust Interpretation via Rényi Differential Privacy", *Artificial Intelligence*, vol. 313:p. 103787, 2022.

[84] M. Wicker, J. Heo, L. Costabello and A. Weller, "Robust Explanation Constraints for Neural Networks", *arXiv preprint arXiv:2212.08507*, 2022.

[85] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart and J. Sun, "RETAIN: An Interpretable Predictive Model for Healthcare using Time Attention Mechanism", *arXiv preprint arXiv:1608.05745*, 2016.

[86] L. A. Hendricks, R. Hu, T. Darrell and Z. Akata, "Generating Counterfactual Explanations with Natural Language", *arXiv preprint arXiv:1806.09809*, 2018.

[87] A. Adadi and M. Berrada, "Explainable AI for Healthcare: From Black Box to Interpretable Models", *Embedded Systems and Artificial Intelligence*, pp. 327–337, Springer, 2020.

[88] R. Ghaeini, X. Fern and P. Tadepalli, "Interpreting Recurrent and Attention-Based Neural Models: a Case Study on Natural Language Inference", *Conference on Empirical Methods in Natural Language Processing*, pp. 4952–4957, 2018.

[89] M. Honnibal, I. Montani, S. Van Landeghem and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python", 2020.

[90] A. Ivankay, I. Girardi, C. Marchiori and P. Frossard, "FAR: A General Framework for Attributional Robustness", *The 32nd British Machine Vision Conference*, 2021.

[91] J. Pennington, R. Socher and C. D. Manning, "Glove: Global Vectors for Word Representation", *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[92] N. Mrkšic, D. OSéaghdha, B. Thomson, M. Gašic, L. Rojas-Barahona, P.-H. Su, D. Vandyke, T.-H. Wen and S. Young, "Counter-fitting Word Vectors to Linguistic Constraints", *NAACL-HLT*, pp. 142–148, 2016.

[93] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng and C. Potts, "Learning Word Vectors for Sentiment Analysis", *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, 2011.

[94] W. Lifferth, "Fake News", 2018.

[95] N. Asghar, "YELP Dataset Challenge: Review Rating Prediction", *arXiv preprint arXiv:1605.05362*, 2016.

[96] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", *arXiv preprint arXiv:1907.11692*, 2019.

[97] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", *International Conference on Neural Information Processing Systems*, pp. 5753–5763, 2019.

[98] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein and L. Antiga, "PyTorch: An Imperative Style, High-Performance Deep Learning Library", *International Conference on Neural Information Processing Systems*, pp. 8026–8037, 2019.

[99] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya and S. Yan, "Captum: A Unified and Generic Model Interpretability Library for PyTorch", *arXiv preprint arXiv:2009.07896*, 2020.

[100] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest and A. M. Rush, "Transformers: State-of-the-Art Natural Language Processing", *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Association for Computational Linguistics, Online, 2020.

[101] K. Pearson, "Notes on Regression and Inheritance in the Case of Two Parents", *Proceedings of the Royal Society of London*, vol. 58, no. 347-352:pp. 240–242, 1895.

[102] M. G. Kendall, "A New Measure of Rank Correlation", *Biometrika*, vol. 30, no. 1/2:pp. 81–93, 1938.

[103] L. Myers and M. J. Sirois, "Spearman Correlation Coefficients, Differences between", *Wiley StatsRef: Statistics Reference Online*, 2014.

[104] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, "Language Models are Unsupervised Multitask Learners", *OpenAI blog*, vol. 1, no. 8:p. 9, 2019.

[105] D. Cer, Y. Yang, S.-Y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan and C. Tar, "Universal Sentence Encoder", *arXiv preprint arXiv:1803.11175*, 2018.

[106] J. Vig, "BertViz: A Tool for Visualizing Multihead Self-Attention in the BERT Model", *ICLR Workshop: Debugging Machine Learning Models*, 2019.

[107] G. Navarro, "A Guided Tour to Approximate String Matching", *ACM Computing Surveys (CSUR)*, vol. 33, no. 1:pp. 31–88, 2001.

[108] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation", *International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, 2017.

[109] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers", *Advances in Neural Information Processing Systems*, vol. 33:pp. 5776–5788, 2020.

[110] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", *arXiv preprint arXiv:1910.01108*, 2019.

[111] J. Devlin, M.-W. Chang, L. Kenton and L. K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding", *NAACL-HLT*, pp. 4171–4186, 2019.

[112] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks", *International Conference on Learning Representations*, 2018.

[113] F. Gräßer, S. Kallumadi, H. Malberg and S. Zaunseder, "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning", Association for Computing Machinery, 2018.

[114] S. Baker, A.-L. Korhonen and S. Pyysalo, "Cancer Hallmark Text Classification using Convolutional Neural Networks", 2016.

[115] D. Hanahan, "Hallmarks of Cancer: New Dimensions", *Cancer Discovery*, vol. 12, no. 1:pp. 31–46, 2022.

[116] M. Yasunaga, J. Leskovec and P. Liang, "LinkBERT: Pretraining Language Models with Document Links", *Association for Computational Linguistics (ACL)*, 2022.

[117] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi and R. G. Mark, "MIMIC-III, a Freely Accessible Critical Care Database", *Scientific Data*, vol. 3, no. 1:pp. 1–9, 2016.

[118] World Health Organization, "International Classification of Diseases — Ninth revision (ICD-9)", *Weekly Epidemiological Record = Relevé épidémiologique hebdomadaire*, vol. 63, no. 45:pp. 343–344, 1988.

[119] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang and Y. Luo, "Clinical-Longformer and Clinical-BigBird: Transformers for Long Clinical Sequences", *arXiv preprint arXiv:2201.11838*, 2022.

[120] I. Beltagy, M. E. Peters and A. Cohan, "Longformer: The Long-Document Transformer", *arXiv preprint arXiv:2004.05150*, 2020.

[121] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter", *ArXiv*, vol. abs/1910.01108, 2019.

[122] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing", *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1:pp. 1–23, 2021.

[123] J. Y. Yoo and Y. Qi, "Towards Improving Adversarial Training of NLP Models", *Findings of the Association for Computational Linguistics: EMNLP*, pp. 945–956, 2021.

[124] X. Yang, X. He, H. Zhang, Y. Ma, J. Bian and Y. Wu, "Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models.", *JMIR Medical Informatics*, vol. 8, no. 11:pp. e19735–e19735, 2020.

[125] Y. Wang, N. Afzal, S. Fu, L. Wang, F. Shen, M. Rastegar-Mojarad and H. Liu, "MedSTS: A Resource for Clinical Semantic Textual Similarity", *Language Resources and Evaluation*, vol. 54, no. 1:pp. 57–72, 2020.

[126] W. Falcon and The PyTorch Lightning Team, "PyTorch Lightning", 2019.

[127] A. Kurakin, I. J. Goodfellow and S. Bengio, "Adversarial Examples in the Physical World", *Artificial Intelligence Safety and Security*, p. 99, 2018.

[128] Y. LeCun, "The MNIST Database of Handwritten Digits", *http://yann.lecun.com/exdb/mnist/*, 1998.

[129] H. Xiao, K. Rasul and R. Vollgraf, "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms", *arXiv preprint arXiv:1708.07747*, 2017.

[130] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images", Tech. rep., Department of Computer Science, University of Toronto, 2009.

[131] J. Stallkamp, M. Schlipsing, J. Salmen and C. Igel, "The German Traffic Sign Recognition Benchmark: A Multi-Class Classification Competition", *International Joint Conference on Neural Networks*, pp. 1453–1460, IEEE, 2011.

[132] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[133] K. He, X. Zhang, S. Ren and J. Sun, "Surpassing Human-Level Performance on ImageNet Classification", *IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

[134] X. Glorot and Y. Bengio, "Understanding the Difficulty of Training Deep Feedforward Neural Networks", *International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[135] Y. LeCun, "LeNet-5, Convolutional Neural Networks", *URL: http://yann. lecun. com/exdb/lenet*, 2015.

ADAM DANIEL IVANKAY

I am a research and software engineer in explainable deep learning for natural language processing, passionate about the intersection of machine learning and real-world problems. I try to solve the hurdles of deploying cutting-edge algorithms in critical real-life domains, such as healthcare. I have 5+ years of experience working in big tech and have been a key member of several client projects, such as the Medgate medical triage application.

CONTACT

- adam@ivankay.de
- +41 78 676 74 35
- Zurich, Switzerland (C-Permit)
- Adam Ivankay
- Publications

SKILLS

Deep Learning

Python
PyTorch/TensorFlow
Natural Language Processing
Computer Vision
Explainability
Robustness
NLP Frameworks
(Hugging Face, SpaCy, NLTK, Gensim)
Large Language Models
Acceleration
(GPUs, distributed, HF Accelerate)
Scientific Coding
(Numpy, Pandas, Scipy, Scikit-learn)
Data Processing

Software Development

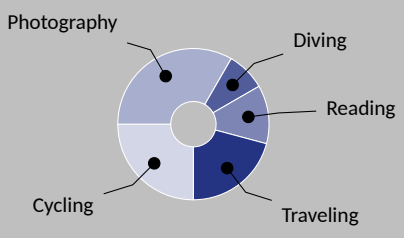
Docker
Git
CI/CD
MLOps
Testing
Databases
(PostgreSQL, MySQL, MongoDB)
Coding Standards
(PyFlake, Pytest, Sphinx)
Mobile App Development

5 - Very Confident (used in several projects)
1 - Basic Knowledge

Languages

German	**Native/Bilingual**
English	**Full Professional Proficiency**
Hungarian	**Native/Bilingual**
French	**Elementary Knowledge**

LEISURE



EXPERIENCE

Pre-Doctoral Researcher

IBM Research, *Zurich*
2019-2023

Worked on several industrial, client and research projects:

- Developed an automated triage system for a major European telemedicine provider. Responsible for several parts of the development, such as entity extraction from free text EHR data, design and implementation of a knowledge graph-based data representation for fast retrieval and implementation of a Q&A between app and patient.
- Developed a mobile-health virtual coaching and monitoring station for chronic COPD patients, increasing their treatment adherence and well-being. Responsible for development of backend for sensor data collection, creating and maintaining SQL database for health data and implementing aggregated data view for patients and clinicians.
- Trained vision machine learning models for industrial fault and crack detection on bridges. Validated best performing models through downstream task performance. Created a 3D model for easy navigation of crack annotations.
- Maintained industry coding standards, such as modularization, documentation and unit testing.

Python | Deep Learning | PyTorch | NLP | Data Processing | Docker | Git | Testing | CI/CD | MLOps | Databases | Computer Vision | Client Interaction | Coding Standards

Development Intern

IBM Research, *Zurich*
2017-2019

Developed and delivered an interactive patient Q&A module for an automated medical triage system to one of Europe's largest telemedicine provider.

Python | Bash | Numpy | MongoDB | Information Retrieval | Docker | Testing | CI/CD

Student Researcher

Supercomputing Systems AG, *Zurich*
2016-2017

Assessed the feasibility of using several types of autoencoders to perform automated male fertility tests based on images.

Python | TensorFlow | Transfer Learning | Unsupervised Learning | Bash

EDUCATION

Doctor of Philosophy (Ph.D.) in Deep Learning for Natural Language Processing

EPF Lausanne, *Lausanne*
2019 - 2023

Doctoral dissertation in robust explainable deep learning in medical text applications.
Thesis: "Estimating and Improving the Robustness of Attributions in Text"
Supervisor: Prof. Dr. Pascal Frossard

Master of Science (M.Sc.) in Electrical Engineering and Information Technology

ETH Zurich, *Zurich*
2015 - 2018

Graduated with majors in deep learning, statistical inference and computer vision.
Thesis: "Data-driven Medical Triage: Confidence and interactive Q&A for triage decision support"

Bachelor of Science (B.Sc.) in Electrical Engineering and Information Technology

Technical University of Munich, *Munich*
2012 - 2015

Graduated with majors in computational intelligence, algorithms and programming.
Thesis: "Development of Wireless Infrared Communication between LDVbots"

ACHIEVEMENTS, HONOURS AND AWARDS

- Several papers at top AI conferences, such as ICLR, AMIA and BMVC
- IBM Invention Plateau for published patents
- IBM A-Level Accomplishment for medical decision support
- National Physics Educational Championship (Hungary) - 5th place