UNIVERSITY OF AMSTERDAM

MSc Artificial Intelligence

Master Thesis

Global Aggregations of Local Explanations for Black Box Models

by

Ilse van der Linden

10093990

April 9, 2019

36 EC

April 2018 - April 2019

Supervisors:

Prof. Dr. E. Kanoulas

Prof. Dr. H. Haned

Assessor:

Prof. Dr. M. de Rijke

Abstract

The increased capabilities of machine learning methods in recent years have stimulated real-life applications in which consequential decisions are made based on model predictions. However, there is always a divergence between optimization goals and requirements in real-life application. Therefore, we cannot assume that the right rule is applied by the model, that the model takes in all - and only - the relevant information, and that all the data is accurate. Typically, the decision-making process of many of these models is inherently inscrutable to the extent that it is impossible for a human to interpret the model directly: they are black box models. This has led to a call for research on explaining black box models, for which there are two main approaches. Global explanations that aim to explain a model's decision making process in general, and local explanations that aim to explain a single prediction. Since it remains challenging to establish fidelity to black box models in globally interpretable approximations, much attention is put on local explanations. The open challenge for local explanations is the gap between local explanations and global model behavior. To what extent do local explanations reliably represent a model's behavior and how can the explanations best be used to gain global insights on a black box model? To answer these questions, we present Global Aggregations of Local Explanations (GALE). The objective of GALE is to provide insights in a model's global decision making process. Overall, our findings indicate that global aggregations of local explanations have the potential to gain global insights from local explanations. Furthermore, our results reveal that the choice of aggregation matters regarding the ability to gain reliable and useful global insights on a black box model. We find that the global importance introduced for LIME does not reliably represent the model's global behavior. Our proposed aggregations are better able to represent how features affect the model's predictions and performance, and to provide global insights by identifying distinguishing features.

Acknowledgements

I would like to express my great appreciation to Evangelos Kanoulas for his willingness to be my supervisor. Even though it was a challenge to dive into a relatively new field of research, you have made it possible for me to pursue a topic that greatly interested me. The generous bestowal of your time has been very much appreciated, as was your patience and impatience, both of which were present exactly when I needed them.

Also, I am particularly grateful for the assistance provided by Hinda Haned. Every step of the way, you made me aware that I was closer than I recognized. Your enthusiastic encouragement, useful critiques and constructive advice enabled me to focus my broad journey of exploration into the concrete contributions presented in this work.

I would like to thank my friends and family for their support and patience. In particular, I wish to acknowledge Alexander van Someren and Casper Thuis, for all the collaboration over the course of our master and for being great company throughout my study. Furthermore, I offer my special thanks to my aunt Mirjam van der Linden and my uncle Erik Wilms for the generous offering of a study home throughout the writing of this thesis.

Lastly, I would like to express my deep gratitude to Ruben Seggers. There are many reasons why this work would not exist if it were not for you. From the many mentions of entropy that I endured over the years, to all your patience and support over the last year. Thank you so much.

Contents

List of Symbols

N	number of instances, i.e. documents in test set	$\mathbb{N}$
M	number of features	$\mathbb{N}$
D	document length	$\mathbb{N}$
V	set of feature ids	$\mathbb{N}$
L	set of class labels	$\mathbb{N}$
S_c	set of data instances classified as class c	
x	input	V^D
y	output class	L
f	black box prediction function	$x \rightarrow y$
x'	interpretable representation of input x	$\{0,1\}^D$
z'	sampled data point around x'	$\{0,1\}^D$
z	sampled data point in original representation	V^D
g	explanation function	$z' \rightarrow y$
W	attribution matrix	$\mathbb{R}^{N \times M}$
ξ	locality-aware squared loss	
π_x	similarity measure	$z \rightarrow \mathbb{R}$

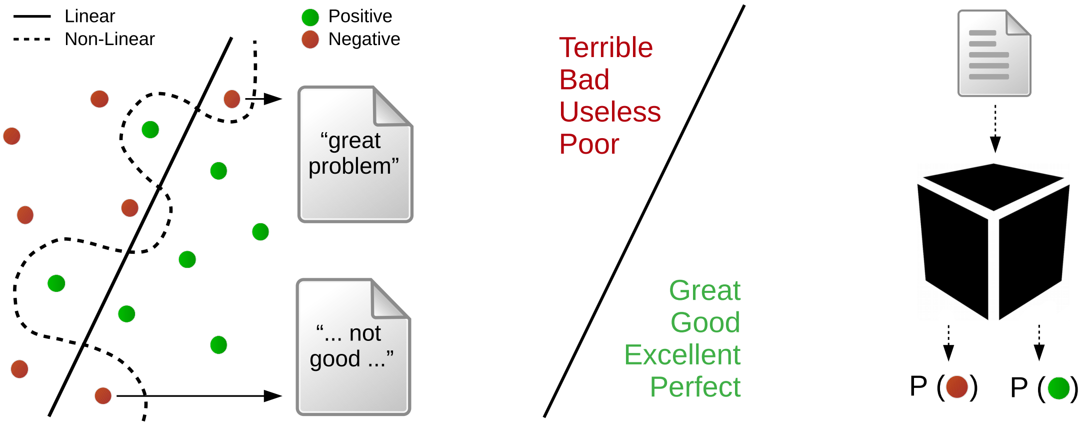
$I_j^{\textsc{Lime}}$	global LIME importance of feature j
$I_j^{\textsc{Avg}}$	global average importance of feature j
I_j^{H}	global homogeneity-weighted importance of feature j
H_j	entropy of the distribution over classes of feature importance I_j
$I_{cj}^{\textsc{Lime}}$	global LIME class importance of feature j for class c
$I_{cj}^{\textsc{Avg}}$	global average class importance of feature j for class c
I_{cj}^{H}	global homogeneity-weighted class importance of feature j for class c

1 | Introduction

The interpretability of machine learning models has gained significant attention in recent years, at least partially in response to a societal demand for it (Boyd & Crawford, 2012; Goodman & Flaxman, 2017). Attempts to meet this demand have followed one of two approaches: developing transparent models (Kim, Rudin, & Shah, 2014; Lakkaraju, Bach, & Leskovec, 2016), or providing explanations for inscrutable models (Simonyan, Vedaldi, & Zisserman, 2013; Karpathy, Johnson, & Fei-Fei, 2015; Ribeiro, Singh, & Guestrin, 2016b). In this thesis we focus on so-called black box architectures, meaning architectures that are inherently inscrutable to the extent that it is impossible for a human to interpret the model directly. Prior work on black box models has focused on either global or local explanations, where global explanations aim to explain a model's decision making process in general, while local explanations aim to explain a single prediction specifically (Ribeiro et al., 2016b). Global explanations suffer from the trade-off between interpretability of the explanation model and fidelity to the black box model, i.e. the more comprehensible a simplified explanation is, the less faithful it can be to the complexity of the black box model. Local explanations solve this by being restricted to local fidelity: fidelity to the black box model in the vicinity of the instance examined. The drawback is that it is unclear in what way the inspected instance is representative of the global behavior of the model. This thesis intends to fill this gap by presenting Global Aggregations of Local Explanations (GALE), to understand to what extent local explanations are representative of the model's decision rules and provide reliable and useful global insight.

Over the last decade, many machine learning breakthroughs occurred, spiking widespread interest in the development of advanced machine learning methods, most specifically in the field of deep learning (Krizhevsky, Sutskever, & Hinton, 2012; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014; Greff, Srivastava, Koutník, Steunebrink, & Schmidhuber, 2017; Goodfellow et al., 2014). By using many layers of non-linear operations and abstractions, these complex models make it possible to make more accurate predictions than simpler methods can achieve. The increased capabilities of these machine learning models have stimulated the development of real-life applications in which consequential decisions are made based on model predictions. Examples of such applications include autonomous driving (Bojarski et al., 2016), clinical diagnosis (Esteva et al., 2017), and financial forecasting (Korczak & Hemes, 2017).

The downside of these complex models is that the decision-making process of such models is not comprehensible to a human. An intuitive example of this drawback is shown in the form of a sentiment analysis task in Figure 1.1. Sentiment analysis is a binary classification task in which documents are labeled as expressing an overall positive or negative opinion (Pang, Lee, et al., 2008). As Figure 1.1a suggests, a simpler linear model might be less able to approximate the actual decision boundary than a more complex model that is able to grasp non-linear patterns. For example, in documents where a more complex interaction of features occurs, such as “great problems" or “not good", a linear classifier might be unable to classify the document correctly. On the other hand, as illustrated in 1.1b and 1.1c, a linear model can provide direct insight in how particular words influence its decisions, while the decisions of a complex model are inscrutable to humans; the model is a black box. Although such black box models can provide higher accuracy in their decisions in many cases, there use is limited by a lack of trust in their decision-making process due to the inscrutability of the model (Passi & Jackson, 2018).



(a) Decision boundaries of a linear and a non-linear model for a sentiment analysis classification task.

(b) Linear model provides interpretable insight of decision making process.

(c) Black box model provides only probabilities and no further insight.

Figure 1.1: Comparison of complexity and performance between a linear and a non-linear model in Figure 1.1a. Illustration of interpretability of a linear model in Figure 1.1b. Illustration of interpretability of a non-linear model in Figure 1.1c.

From a theoretical perspective, there is reason to doubt our models’ decisions. The need to understand what a decision is based on arises from the divergence between optimization of test set performance and requirements of application in a real-life environment (Lipton, 2016). Possible causes for this divergence are that the optimization goal is a simplification of the actual problem or that the offline training data does not resemble the dynamic deployment environment. Because of the inscrutability of state-of-the-art models, we are increasingly less able to assess this divergence directly (Selbst & Barocas, 2018). For this reason, we cannot assume that the right rule is applied by the model, that the model takes in all - and only - the relevant information, and that all the data is accurate.

The inscrutability of black box models, combined with the potential application for consequential decisions that might affect our safety, our economy, or our opportunities, makes the societal demand for insight in their decision-making process pertinent. This has recently led to EU regulation on the right to "meaningful information" about the logic involved in automated decision-making[1]. Although there is discussion on the legal implication of the terminology, admittedly it implies a right to explanation (Selbst & Powles, 2017). The difficulty lies in questions regarding what makes for a valid, reliable and useful explanation; what are the desiderata?

Currently proposed explanation approaches differ in the kind of insight they provide. Global explanations provide insight on a global level by explaining a model's decision making process in general. The open challenge for these approaches is to remain faithful to the original model while providing a simple and therefore interpretable explanation. This is inherently difficult: if a simple global model could easily grasp the task at hand, no complex model would be needed to begin with. Local explanations provide insight on the level of a single prediction by explaining the model's decision making process in the vicinity of that prediction. Instead of aiming to explain the implicit decision rules a sentiment analysis model uses to classify any input as either positive or negative, a local explanation explains why one particular document is classified as it is. Such an explanation provides local fidelity to the model, i.e. it can be expected to be representative of highly similar documents, but not in general. Although local explanations are reliable by restricting their explaining power to local model behavior, a shortcoming is that it is unclear how this relates to global model behavior.

This shortcoming is illustrated in Figure 1.2 by an example explanation for a sentence classified as expressing a positive sentiment. Although it provides insights in this particular example, it raises questions about the general patterns that underlie this decision. A user might be inclined to expect similar behavior in other situations. However, due to the locality of these explanations and its undefined coverage, there would be no way of saying something valid about even a quite similar sentence. Let us take the sentence "I have great problems working with this software!" for example. Even though it includes similar and even the same words as the sentence in Figure 1.2, there is no way of knowing what to expect from the model.

[1] EU General Data Protection Regulation: https://gdpr-info.eu/

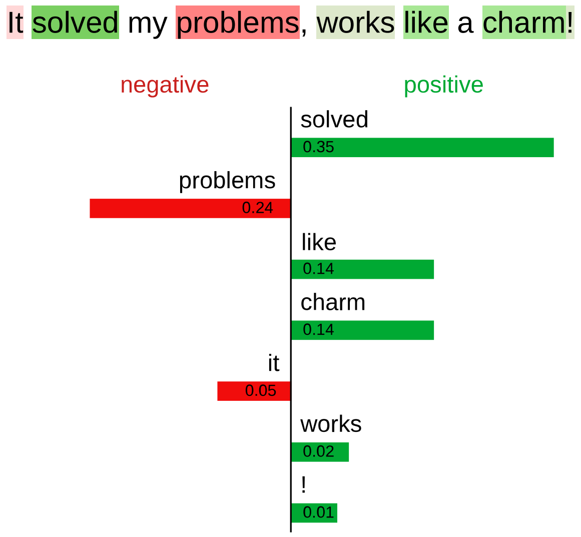


Figure 1.2: Example local explanation for a sentence classified as expressing a positive sentiment. The green color indicates attributions supporting the positive sentiment, while red represents attributions that support a negative sentiment. The attribution value for each feature is displayed in the bar under the feature.

Therefore, we need to know more about the way in which local explanations are representative of global model behavior. A user cannot assume an explanation or parts of it generalize to other instances, because this approach only provides a degree of local fidelity with undefined coverage. For these explanations to be useful and reliable, a user needs a clear understanding of how to interpret them. A local explanation relates in some way to the global logic of the model. Increasing our understanding of the representativeness of local explanations, the emerging patterns they reflect, and the limitations of their locality is necessary in order to make proper use of them.

In this thesis we present our approach and evaluation of gaining global insight from local explanations. For this task we analyze explanations obtained through Local Interpretable Model-agnostic Explanations (LIME) on models trained for a multiclass document classification task and a binary sentiment analysis task. We present several approaches to aggregate a set of local explanations in order to gain global insights on a model's decision making process. Our main research question is:

Research Question

How can local explanations be aggregated to provide global insights on a black box model?

The following section will provide background on the field of interpretability which will also clarify the scope of this work: local explanations, specifically additive feature attribution methods. Subsequently, we provide a further review of previous work, discussing desiderata for local explanations, attempts to gain global insight from local explanations and evaluation of such work. Thereafter, we include an elaborate description of LIME in Section 3, which is the additive feature attribution method used in this thesis. Section 4 will present our main contribution: Global Aggregations of Local Explanations. This is followed by a description of the experimental design for our evaluation in Section 5. Lastly, we present our conclusion and a discussion of future work in Section 7.

2 | Related work

> *A model is interpretable if it provides explanations for its predictions in a form humans can understand; an explanation provides reliable information about the model's implicit decision rules for a given prediction. A model is accurate if most of its predictions are correct, but only right for the right reasons if the implicit rules it has learned generalize well and conform to the domain experts' knowledge about the problem.*
>
> Ross, Hughes, and Doshi-Velez, *2017*

Due to the novelty and broadness of the field of interpretability, many different purposes and definitions can be found across the literature. Interpretability and explanations are terms often used interchangeably, while these are distinct concepts; in this thesis we follow the formulation by Ross, Hughes, and Doshi-Velez (2017) as shown in the preamble of this section. We consider interpretability as a goal, and explanations as a means to that goal. Explanations are not the only form in which interpretability can be achieved, and interpretability - although it might be the core motivation for our efforts to explain - is not the only requirement for reliable and useful explanations.

Firstly, this section aims to clearly position this thesis within the field. For this purpose, we elaborate on methods for interpretability and distinguish the local explanations that are within the scope of our research. Subsequently, the desiderata for such explanations and previous work on gaining global insight from local explanations is discussed.

2.1 The scope of this thesis

The authors of acclaimed position papers argue that there is a lack of foundation in much of the published work on interpretability (Lipton, 2017) and that the lack of consensus leads to an unrigorous evaluation of proposed approaches (Doshi-Velez & Kim, 2017). Therefore, we aim to clearly specify the scope of our contributions within the field. For this purpose, we follow the distinctions made by Lipton (2016), Ribeiro et al. (2016b), Guidotti et al. (2018) and Lundberg and Lee (2017), to distinguish the set of post-hoc local explanations that are the focus of this thesis.

2.1.1 Post-hoc explanations

Lipton (2016) makes a broad distinction within approaches aimed at increasing model interpretability: those that aim to offer transparency on how the model works and those that put forward post-hoc explanations on what the model does. In some cases it might be possible to obtain a transparent model with high performance, yet in many cases decreasing the complexity of the model used would yield a lower performance on the task.

The focus of this thesis is on models that are inherently not transparent in the sense that they are not comprehensible to a human, i.e. *black box models*. This lack of transparency can arise from two causes: the size of the model - the amount of parameters and computations, or the complexity of the computations it performs (Lipton, 2016), both of which we consider to be too large in state-of-the-art neural architectures. Even though we could provide a user with all trained parameters of a neural network, this will not lead to understanding due to the amount and complexity of computations involved.

Post-hoc explanations intend to explain what the model does instead of how it works. This point of view can be compared with a human explaining their decision making processes. A human is not able to explain the workings of their brain, and this would probably not be useful either. Instead, a human will explain, to their ability, what their decision was based on. It offers an explanation of what has influenced their decision and what considerations were being made. Likewise, we can interpret a model post-hoc - after training, by providing explanations for its predictions.

2.1.2 Local explanations

Post-hoc explanations can be further divided into global and local explanations (Ribeiro et al., 2016b). A global explanation offers insight in the model's decision making process; it aims to explain the model. It does so by specifying an interpretable model that provides insight through approximating the original model. This approach trades fidelity to the original model for interpretability (Guidotti et al., 2018). Since it remains challenging to establish fidelity to black box models in globally interpretable approximations, attention has been put on an alternative first proposed by Ribeiro et al. (2016b). They propose a local explanation model that offers insight in a model's prediction on a single instance; it aims to explain a single prediction. It does so by approximating the model's decision making in a specific region of the input space. By approximating a model locally, local fidelity to the original model is optimized. The aim of such an approach is to obtain an explanation that complies to the model's behavior in the vicinity of the instance under examination. (Lundberg & Lee, 2017).

2.1.3 Additive feature attribution methods

The work of Lundberg and Lee (2017) identifies a family of approaches that provides a local explanation model in the form of a linear function of binary variables. By representing the features as binary variables, the weights of the linear model w can be directly interpreted as feature attributions. This ensures that the explanation model is interpretable to humans, even though the original model might use complex features as input. For this reason, the methods require a mapping between x, the feature input of the model, and x', an interpretable representation of that input. The explanation model $g(x')$ is a linear function of the interpretable representation, and approximates a model's prediction on an instance $f(x)$:

$$g\left(x'\right) = w_0 + \sum_{j=1}^{D} w_j x'_j \tag{2.1}$$

where D is the input dimension of the instance explained. Local feature attribution methods provide the user with a local explanation through the weights of this linear model. The weights are considered feature attributions that reflect the influence on a prediction per feature. Attributions can either be supporting or opposing, and the higher the attribution, the higher the explanation considers a feature's influence on the prediction. The sum of the attributions for all features in the input approximates the model output $f(x)$.

Lundberg and Lee (2017) unify a family of approaches that provides an explanation model that is a linear function of binary variables as shown in Equation 2.1. LIME, the local explanation method used in this thesis, is one such approach. Other methods that adhere to Equation 2.1 include Layer-wise Relevance Propagation (LRP), DeepLift and SHapley Additive exPlanation Values (SHAP).

LIME offers a local explanation by approximating the model in the vicinity of the instance being explained. It gathers local information by sampling instances from the instance being explained and is therefore considered a perturbation-based approach. The linear approximation model around a specific prediction offers an interpretable explanation that adheres to Equation 2.1 (Ribeiro et al., 2016b; Lundberg & Lee, 2017). LIME is discussed elaborately in Section 3.

Layer-wise Relevance Propagation is a backpropagation-based approach. It computes a relevance score, i.e. attribution, for each input neuron by redistributing the output of the prediction function backwards layer-by-layer. At each layer l the relevance per neuron is determined by that neuron's activation in layer l, the magnitude of the weights connecting it to layer $l+1$ and the relevance of the neurons in layer $l+1$. The exact redistribution function can differ depending on the classifier, as long as it satisfies the *relevance conservation* property: at each layer the total amount of relevance, i.e. the sum of relevance of all neurons in the layer, equals the output of the prediction function (Bach et al., 2015).

DeepLift improves over LRP by introducing the notion of a *reference value*. The reference value of a neuron is defined as the neuron's activation to a reference input. The reference input is defined per task for the input neurons and is then propagated through the network to obtain reference values for all neurons. Subsequently, DeepLift obtains contribution scores, i.e. attributions, by expressing the difference from reference value of the output neuron in terms of the difference from reference value of the input neuron. Since the contribution scores are propagated from the output layer to the input layer, DeepLift is also considered a backpropagation-based approach. In fact, it is equivalent to LRP with all reference values set to zero (Shrikumar, Greenside, & Kundaje, 2017; Lundberg & Lee, 2016).

SHapley Additive exPlanation values Apart from unifying the additive feature attribution methods as a family of approaches with the same explanation model, Lundberg and Lee (2017) reveal the relationship between these methods and the concept of *Shapley values*, from the field of game theory, and propose them as a measure to quantify feature importance. The Shapley value of a feature is the averaged marginal contribution of that feature to all possible subsets of features, i.e. meaning the average difference in prediction with or without the feature included for each subset. Lundberg and Lee (2017) present several approaches for approximating the Shapley values to obtain an explanation model.

While this thesis mostly uses the more general description *local explanation*, it will refer to this specific family of approaches. The global aggregations that we present in Section 4.2 are applicable to local explanations that adhere to Equation 2.1.

2.2 Desiderata for local explanations

The general trade-off for post-hoc explanation models is that of fidelity versus interpretability (Guidotti et al., 2018; Ribeiro et al., 2016b), in other literature sometimes referred to as completeness versus comprehensibility or complexity. Fidelity entails the faithfulness of the explanation model in relation to the original model. This can be assessed either at a global or a local level, depending on the method providing global or local explanations. Methods designed to provide local fidelity generally do so in order to increase interpretability in comparison to their global counterparts. Interpretability is the extent to which the explanation for a model's decision making is comprehensible to a human; it requires the complexity of an explanation to not exceed human understanding (Guidotti et al., 2018; Gilpin et al., 2018). Although balancing this trade-off remains challenging, the desiderata of fidelity and interpretability are quite widely accepted. Nonetheless, there remains discussion on whether they constitute sufficient requirements to provide useful and reliable explanations. This need for other dimensions to be accounted for is described differently in a range of papers, some of which we will highlight specifically.

It is clear the choice for local feature explanations limits the fidelity of resulting explanation models to local fidelity. As with this choice for local explanations, the trade-off between fidelity and interpretability underlies many other design choices in post-hoc explanation approaches. The risk of focusing mainly on interpretability is that of favoring explanations that make sense to humans at the expense of their fidelity to the actual model being interpreted (Herman, 2017). The case for interpretable explanations that confirm user expectations is that the aim of interpretability in the first place is to increase user trust and expectations. However, the result of this might be persuasive models that lack in fidelity as opposed to descriptive models with higher fidelity to the original model. Therefore, we side with Gilpin et al. (2018), in that it is unethical to provide users with a simplified explanation if the limitations of these explanations cannot be understood.

Ribeiro et al. (2016b) stipulate LIME suffers from the fact that it is unclear what the coverage of an explanation is, i.e. to what extend it generalizes to other situations. Correspondingly, Mittelstadt, Russell, and Wachter (2018) claim a three way trade-off, adding the size of the domain described as a third desiderata. In another recent review, it was argued that explanations should be required to show relevance, meaning they must provide insights for users into a chosen domain problem (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). In general, this requires a diversity in approaches, varying the balance in fidelity and interpretability, to meet the need of distinct domain problems. Particularly in the case of post-hoc explanations, Murdoch et al. (2019) point out the gap between local explanations and global model behavior, leading to a call for future work to answer to what extent post-hoc methods fully capture a model's behavior and how the explanations can best be used. Our work intends to increase the understanding of the limitations of post-hoc local explanations and through this improve the usefulness of such explanations.

2.3 Global insight from local explanations

As mentioned, the authors of the LIME paper acknowledge the gap between their method and the model's global behavior (Ribeiro et al., 2016b). It is with this limitation in mind that they propose their submodular pick algorithm for selecting a set of data instances for which to present explanations to a user in order to provide global insight.

The submodular pick algorithm selects explanations to show by picking instances that contain features with a high global importance score. They aim to define global importance such that features that explain many different instances have a higher global importance score, than features that explain less instances. In the case of text classification they propose the global importance I_j of a feature j as the square root of the sum of its attributions, as shown and further elaborated on in section 4.2.1. Additionally, they aim to pick a subset S out of N instances, such that there is little redundancy in the features shown, arguing that if a feature appears in multiple instances this would lead to similar explanations shown to a user. Both these intuitions are formalized as a coverage function c. Let W be the $N \times M$ matrix containing the attributions per instance for each feature out of M unique features. Then, given a set of instances S, the explanation matrix W and the importance vector I, the coverage of the set S is defined by:

$$c\left(S, W, I\right) = \sum_{j=1} \left[\exists i \in S : W_{ij} > 0\right] I_j \tag{2.2}$$

Maximizing this function would yield the optimal set of instances covering important features. However, since this problem is NP-hard, a greedy algorithm is proposed that iteratively adds instances to the set S to approximate the optimal set. The instance i added at each iteration is the one with the highest marginal gain, which is defined as $c\left(S \cup \{i\}, W, I\right) - c\left(S, W, I\right)$ (Ribeiro et al., 2016b).

In the article where this approach is presented, only limited evaluation on the submodular pick is provided, comparing their approach only to providing randomly selected instances to the user. It is shown that providing explanations for specific instances chosen by the submodular pick algorithm makes users more able to compare between models and features and make choices that positively affect performance, than when they are shown a random set of instances. It is unclear to what extent the chosen instances are representative for the global behavior of the model, and the importance function is not further evaluated. Therefore, the gap between the local explanations and the global model behavior remains.

3 | Preliminaries

3.1 Local Interpretable Model-agnostic Explanations

The Local Interpretable Model-agnostic Explanations (LIME) approach is to locally approximate the model with an interpretable model that is then considered the local explanation (Ribeiro et al., 2016b). In order to make no assumption about the original model, it samples data points around the instance for which it aims to explain the model's predictions. The LIME explanation is the result of fitting an interpretable model on the sampled data points considering as target labels the predictions of the original model on these data points. It does so by minimizing the locality-aware squared loss while penalizing for the complexity of the explanation model.

To obtain a local approximation of the model's behavior, LIME first samples data points in the vicinity of the data instance being explained. Since models can use complex features as input, this often requires a mapping between the feature input of the model x and an interpretable representation of that input x'. Subsequently, data points are sampled from x' by selecting a subset of the interpretable representation uniformly at random (where the size of this subset is also drawn uniformly at random), which results in an interpretable perturbation z'. After recovering these samples in the original representation z, it uses the models predicted probabilities on these samples $f(z)$ to fit an interpretable model. This model $g(z')$ is fit on the interpretable sampled input z' and therefore provides an interpretable explanation model that has local fidelity to the original model.

The explanation model is obtained by minimizing the locality-aware squared loss, because the influence of a sample z' should depend on its similarity to the interpretable input x' being examined. In the case of text input the authors propose a squared exponential kernel on the cosine similarity (Ribeiro et al., 2016b). The choice for this distance metric is not further justified by the authors. It has the effect of exponential decay of the weights when cosine similarity decreases. This results in

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \sum_{z, z' \in Z} \pi_x(z) \left(f\left(z\right) - g\left(z'\right)\right)^2 + \Omega(g) \tag{3.1}$$

where $\pi_{x'}(z')$ is the similarity measure of sampled data points z' with respect to the interpretable representation x' and $\Omega(g)$ penalizes for the complexity of the explanation model.

4 | Methodology

With our work we intend to fill the gap between local explanations and global model behavior. This gap limits the reliability and usefulness of local explanations, because it is unclear to what extend these explanations capture a model's behavior and how they can be interpreted. We formalize our perspective as an additional desideratum specific to local explanations in Section 4.1. Building on this we present our approach to gain global insight from local explanations in Section 4.2. Rather than validating local explanations independently, we assess their usefulness to gain overall insight on the model. This does not mean a global understanding in the sense of a global explanation model. Instead we assess the emerging insight from aggregating multiple local explanations. For this purpose, we propose a set of Global Aggregations of Local Explanations (GALE).

4.1 Desideratum of representativeness

In order for local explanations to provide reliable and useful global insight, we argue they need not only be interpretable and show local fidelity to the underlying model, but they also need to be *representative* of the global model behavior. To be representative means to be able to be used as a typical example of something[1]. Being representative relates to, yet is different from generalizing. A local explanation is specific to the locality in which the model is making its decision and therefore does not necessarily need to be generalizable. Instead it needs to be representative of the model as a typical example of its behavior, either by showcasing generalizable model behavior, or by being a specific example of a case in which the model's decision making differs from the general. Reconsidering the quote from Ross et al. (2017) at the preamble of Section 2: as with any model, explanation models need also be assessed on whether or not they are right for the right reasons, meaning the explanations need to generalize to some extend and conform to the users expectations of them being representative of the model's decision rules. We will show that different aggregation functions inhabit distinct assumptions that may or may not hold. Inappropriate assumptions can

[1] https://en.oxforddictionaries.com

lead to a misleading representation of the model's global decision making process.

4.2 Global aggregations

The way to aggregate local explanations is not straightforward since the attribution scores are not determined in relation to other data instances. It is unclear how attributions between different instances or in support of different classifications relate to each other. This is further complicated when applied to a textual task, given the sparsity of features in this domain. Any choice of aggregation function inhabits assumptions about the way in which local explanations are representative of the global model behavior. In our discussion of GALE we aim to make these assumptions explicit.

4.2.1 Global LIME importance

Ribeiro et al. (2016b) propose their submodular pick algorithm to select a set of instances to show a user in order to provide global insight. In order to select a representative and informative subset of instances, they propose a global aggregation function I to assess global feature importance. Specifically for the text domain, Ribeiro et al. (2016b) define the global feature importance I_j^{LIME} as the square root of the sum of attributions W_{ij} of the feature j over all data instances $i \in \mathbb{N}$:

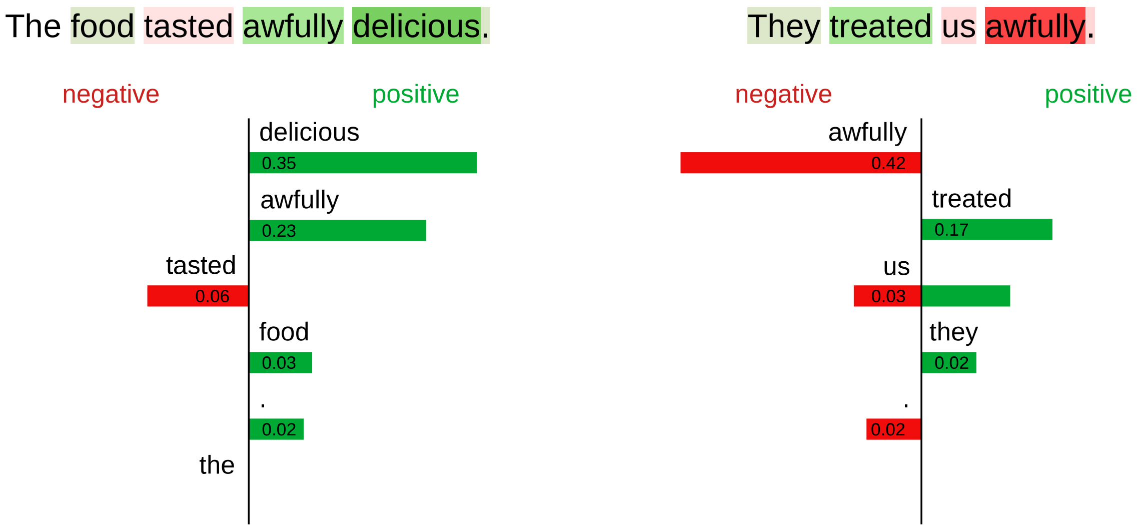
$$I_j^{\text{LIME}} = \sqrt{\sum_{i=1}^{N} |W_{ij}|} \tag{4.1}$$

Two assumptions underlie this aggregation function:

1. ***Features with higher attributions are expected to have a larger effect on model predictions than features with lower attributions.***
2. ***Features that occur more often are expected to have a larger effect on model predictions.***

In the aggregation over separate instances, these assumptions will not always hold. Although assumption 1 seems reasonable amongst the features within one instance, this is less certain for feature attributions from various instances. Since the explanation model is a linear function of attribution values, the magnitude of attributions are affected by the amount of features per explanation. Similarly, the magnitude of attributions are affected by the prediction value that the explanation model is approximating. However, in comparison between attributions from different instances, the absolute value of the attribution might be less informative than its relative importance within the instance.

With respect to assumption 2, the amount of occurrences of a feature might be a misleading notion for several reasons, especially in text classification. Firstly, the assumption is that occurrences across different instances amount to a higher influence of the feature. Common words, such as “the”, “and”, or “is”, are then likely to be ranked as very important due to many occurrences in different instances, even when their attributions are low. Secondly, as mentioned in Ribeiro, Singh, and Guestrin (2016a), local explanations for different instances can also be inconsistent with each other. An example of an inconsistency is shown in Figure 4.1 where the feature “awfully” is explained as being supportive for predicting a positive sentiment in the instance shown in Figure 4.1a. From the example in Figure 4.1b we know that the feature also - and maybe to a greater extend - influences the model in favor of negative sentiment predictions. An explanation in the form of feature attributions does not show a feature’s influence in other cases when providing a single example. This issue is further amplified in case of a multiclass classification task, because a single explanation neither shows a feature’s possible relationship to other classes when providing a single example. Features that occur in documents of different classes will have a higher global LIME importance, independently of what classes the individual attributions support.



(a) Explanation for a positive sentiment.

(b) Explanation for a negative sentiment.

Figure 4.1: Example of inconsistency in feature attributions. The feature “awfully” is attributed differently in either sentence.

4.2.2 Global average importance

As described above, the LIME importance makes the assumption that features that occur more often in the data are more important. However, this clearly depends on the domain. In case of textual data we often deal with sparse features; common words will occur often, while most other words will only occur in few instances. For this reason, we expect the global LIME importance to be unreasonably biased towards common words.

Therefore, the first alternative aggregation we propose is the average importance, which is computed as the sum of attributions W_{ij} averaged over the feature's occurrences in the dataset:

$$I_j^{\textsc{Avg}} = \frac{\sum_{i=1}^{N} |W_{ij}|}{\sum_{i:W_{ij}\neq 0} \mathbb{1}} \tag{4.2}$$

Although the global average importance addresses the second assumption that is made in global LIME importance, it also makes its own assumption:

3. Features are expected to have a similar effect in all of their occurrences.

To understand why this assumption might not hold, imagine the case of a feature being important for some class predictions and less important when appearing in documents unrelated to that class. The occurrence in other documents will strongly lower its average importance, even though the feature was highly important for some classes.

4.2.3 Global homogeneity-weighted importance

The global homogeneity-weighted importance is designed to address assumptions 2 and 3. The idea is to determine the homogeneity of a feature's influence on the model in order to deal with multiple occurrences and potential inconsistencies between occurrences. To quantify the homogeneity per feature, the spread of attributions over different classes is determined by Shannon entropy. First, we define p_j as the vector of normalized LIME importance per class:

$$p_{cj} = \frac{\sqrt{\sum_{i\in S_c} |W_{ij}|}}{\sum_{c\in L} \sqrt{\sum_{i\in S_c} |W_{ij}|}} \tag{4.3}$$

where S_c is the set of all instances i classified as class c and L is the set of class labels. The normalized LIME importance p_j represents the distribution of feature j's importance over all classes $c \in L$. The Shannon entropy of this distribution is defined by:

$$H_j = -\sum_{c \in L} p_{cj} \log\left(p_{cj}\right) \tag{4.4}$$

The entropy score is used to assess the degree of homogeneity with which the feature attributions of a feature are distributed over multiple classes. Low entropy indicates most of the attributions point to one particular class, as opposed to the case of high entropy in which attributions point to many classes. However, since entropy will be equally low for all features that only occur a single time in the test set, the entropy score does not discriminate in these cases. Therefore we propose to derive from this a homogeneity-weighted importance I_j^H. For this purpose, the entropy score is normalized and subtracted from 1 to obtain a weighting factor that is close to 1 if the feature is homogeneous and close to 0 when its attributions are spread over many different classes. The homogeneity weighted importance is the LIME importance of a feature weighted by this weighting factor:

$$I_j^H = \left(1 - \frac{H_j - H_{min}}{H_{max} - H_{min}}\right) I_j^{\text{LIME}} \tag{4.5}$$

where H_{min} and H_{max} are the minimum and maximum entropy measured across all features.

5 | Experimental design

This section describes the experiments done in order to answer our Research Question. Firstly, we elaborate on the data and models used in our experiments. Secondly, we discuss our quantitative evaluation approach, devised to assess how reliable each global aggregation represents the model's global behavior. Lastly, we present our qualitative evaluation, demonstrating the global insights provided by GALE through visualization of the most important features according to each of the global aggregations.

5.1 Datasets and models

Experiments are carried out on two distinct datasets: a relatively small sentiment analysis dataset and a larger document classification dataset. This makes us able to evaluate to what extend our approach is influenced by the complexity of the task and the amount of data available. For each task, a model with a task specific architecture is trained on a subset of the data. Parameter tuning of these models is done based on a validation set that is not further used in our experiments. Subsequently, LIME explanations of the model predictions on a separate test set are obtained and used to gather the global aggregations over all instances in the test set.

5.1.1 Sentiment analysis task

The aim of sentiment analysis is to classify documents based on the expression of sentiment in them; it is a binary classification task in which documents are labeled as expressing either an overall positive or negative sentiment (Pang, Lee, et al., 2008). The sentiment analysis dataset used for this thesis consists of 3000 sentences with labels evenly distributed over a positive or negative sentiment[1]. Kotzias, Denil, De Freitas, and Smyth (2015) selected these instances from

[1] Retrieved from https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

larger datasets originating from the websites of IMDB, Amazon and Yelp, incorporating 1000 sentences per source. The LSTM architecture[2] used for this task consist of one LSTM layer with tanh activation function and both input and recurrent dropout at 0.2, followed by one fully connected softmax layer. The neural network is optimized with Adam over a run of 10 epochs with a batch size of 32. The input features are pretrained 100-dimensional GloVe word embeddings[3], which are not further fine-tuned during training (Pennington, Socher, & Manning, 2018).

5.1.2 Document classification task

The 20 Newsgroups dataset is a collection of approximately 20,000 newsgroup post, almost evenly distributed over 20 different classes, which are depicted in Table 5.1.

comp.graphics comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.os.ms-windows.misc comp.windows.x	sci.crypt sci.electronics sci.med sci.space	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey
politics.mideast politics.guns politics.misc	alt.atheism religion.christian religion.misc	misc.forsale

Table 5.1: 20 Newsgroup classification topics.

The CNN architecture[4] consists of three convolutional layers with ReLU activation functions, max pooling after each convolutional layer and dropout at 0.2. The neural network has a final fully connected softmax layer, is optimized with Adam and run for 10 epochs with batch size at 32. The input features are pretrained 100-dimensional GloVe word embeddings[3], which are not further fine-tuned during training (Pennington et al., 2018). With this setup we obtain an accuracy of around 0.75.

[2] LSTM architecture is based upon the model described in https://towardsdatascience.com/sentiment-analysis-through-lstms-3d6f9506805c

[3] Retrieved from https://nlp.stanford.edu/projects/glove/

[4] CNN architecture is based upon the model described in https://blog.keras.io/using-pre-trained-word-embeddings-in-a-keras-model.html

5.2 Quantitative evaluation of GALE

In order to determine which global aggregation best represents the global model behavior, we propose an adaptation of the Area Over the Perturbation Curve (AOPC) evaluation. The AOPC was proposed by Samek, Binder, Montavon, Lapuschkin, and Müller (2017) as an evaluation metric for local feature attribution methods. It defines a good local explanation as one that is able to identify the features that have the largest effect on model prediction. Expanding on this, we evaluate to what extend the aggregations are able to identify the features that have the largest global effect on model prediction and performance. This effect is measured by progressively removing features per document in the order of their global ranking by GALE. The models' decisions on the resulting documents are evaluated by the Averaged Cumulative Probability Drop (ACPD), and the Averaged Cumulative Accuracy Drop (ACAD). The aggregations are compared against a random baseline. With this approach we intend to address the following subquestions:

Quantitative Research Questions

1. ***To what extent can global LIME importance represent how features affect the models' predictions and performance?***
2. ***To what extent can global average and homogeneity-weighted importance represent how features affect the models' predictions and performance, i.e. can the proposed aggregations improve on global LIME importance?***
3. ***To what extent do the quantitative results differ between a binary and a multiclass text classification task?***

When the AOPC evaluation is applied to local feature attribution methods, features are removed from a particular explanation in the order of their attribution value for that explanation. However, for the purpose of evaluating global aggregations, features need to be removed in the order of highest global importance first according to the global aggregations. The way to do this is not trivial due to the sparsity of features in text documents. One option would be to iteratively remove the highest ranked features according to global aggregations from all documents in the test set at once. However, in that case, different aggregation methods might remove radically different percentages of words from the collection, for example, the removal of frequently occurring features would affect far more documents than the removal of features with low frequency. This evaluation would favor global aggregations that highly rank features that occur often even if those features are not considered important in local explanations. Therefore, in our approach, we remove the same amount of features per document.

For each aggregation of features, the features in a document x_i are ranked according to the global aggregation. Subsequently, features are iteratively removed from the original data point x_i in order of that ranking. Let r be the vector of feature indices in document x_i ranked in the order of the global aggregation. Then, the perturbed instance x_i^k is the result of recursively removing the k highest ranking features defined as follows:

$$\begin{aligned} x_i^0 &= x_i \\ x_i^k &= s\left(x_i^{k-1},\ r_k\right) \end{aligned} \tag{5.1}$$

where the function s removes the k^{th} highest ranking feature in document x_i according to the global aggregation from index r_k in the data point x_i.

Since we assess the *global* importance of features, not only their influence on predicted class probability, but also their influence on overall accuracy is of interest. Therefore, two metrics are proposed: the Averaged Cumulative Probability Drop (ACPD) and the Averaged Cumulative Accuracy Drop (ACAD). These metrics represent the average effect of feature removals at K, with K being the number of removed features. At K, ACPD and ACAD are defined as follows:

- *ACPD@K - the averaged cumulative sum up to K of the drop in predicted class probability averaged over all instances in the test set:*

  $$ACPD@K = \frac{1}{K+1}\left\langle \sum_{k=0}^{K} f(x_i^0) - f(x_i^k) \right\rangle_{avg}^{i\in N} \tag{5.2}$$

  where $\langle\cdot\rangle_{avg}^{i\in N}$ denotes the averaging over all instances in the test set and the black box prediction function f returns the probability of the predicted class.

- *ACAP@K - the averaged cumulative sum up to K of the drop in accuracy over all instances in the test set:*

  $$ACAD@K = \frac{1}{K+1}\sum_{k=0}^{K} \left\langle x_i^0 \right\rangle_{acc}^{i\in N} - \left\langle x_i^k \right\rangle_{acc}^{i\in N} \tag{5.3}$$

  where $\langle\cdot\rangle_{acc}^{i\in N}$ denotes the accuracy over all instances in the test set.

These metrics are computed over a consecutive range of removals per document K. Both resulting curves are evaluated on two aspects. Firstly, the overall height of the curve is assessed. Higher curves indicate a better ranking of features in the order of global influence on predictions and performance respectively. Secondly, we assess the initial slope of the curve. The steeper the slope of the curve for the first features removed, the stronger their influence on the model. A good aggregation is expected to demonstrate a positive decreasing slope for the ACPD curve. The

ACAD is expected to strongly correlate with the ACPD, although the shape of ACAD curve may vary more per task.

This evaluation is used to answer the quantitative research questions. Specifically, we examine the global effect of removing features per document in the order of each of the respective GALE rankings, i.e. global LIME importance, global average importance and global homogeneity-weighted importance. This effect is measured on prediction, i.e. the drop in predicted class probability, and performance, i.e. drop in accuracy, and carried out on both previously discussed datasets. The evaluation for each aggregation is compared against a random baseline, in which randomly selected features per document are removed.

5.3 Qualitative visualization of GALE

In addition to the quantitative evaluation, a qualitative visualization is presented. This enables the examination of whether GALE provides useful insights into the models' global decision making process by demonstrating the most influential features per class. In this perspective, a good global aggregation is one that considers features important if they distinguish between classes. Our qualitative research addresses the following questions:

Qualitative research questions

1. ***To what extent does global LIME importance identify distinguishing features?***
2. ***To what extent does global average importance identify distinguishing features?***
3. ***To what extent does global homogeneity-weighted importance identify distinguishing features?***

In order to answer these questions, a visualization of the most important features per class is demonstrated for each aggregation. To that end, a slightly adapted version of each aggregation function is used to compute per class global aggregations. Let S_c be the set of all instances $i \in \mathbb{N}$ classified as class c. Then, the global LIME class importance for feature j and class c is defined as:

$$I_{cj}^{\textsc{Lime}} = \sqrt{\sum_{i \in S_c} |W_{ij}|} \tag{5.4}$$

The global average class importance for feature j and class c is defined as:

$$I_{cj}^{\textsc{Avg}} = \frac{\sum_{i \in S_c} |W_{ij}|}{\sum_{i \in S_c : W_{ij} \neq 0} \mathbb{1}} \tag{5.5}$$

And lastly, the global homogeneity-weighted class importance for feature j and class c is defined as:

$$I_{cj}^{H} = \left(1 - \frac{H_j - H_{min}}{H_{max} - H_{min}}\right) I_{cj}^{LIME} \tag{5.6}$$

Notice that for the homogeneity-weighted class importance, the homogeneity weighting factor is still computed over all classes.

These global class importance functions are used to visualize the most important features per class, as determined by each of the aggregation methods. The amount of features visualized is chosen per task and aggregation, since the maximum amount for which the visualization remains comprehensible could differ per case. The selected features are plotted using t-SNE for dimensionality reduction on the word embeddings (Maaten & Hinton, 2008). More than just illustrating the most influential features, these visualizations present clusters of words that share a similarity in being indicative of a particular class. A global aggregation that can better identify distinguishing features than other aggregations, is expected to demonstrate more distinct clusters of words. Furthermore, it is presumed that a good global aggregation considers substantive words important, and common words not important. However, it is evident that the visualization depends not merely on the global aggregation method, but also on the model being explained. For instance, if a model in fact considers common words important, the global aggregation would be right to indicate this behavior. Rejecting the global aggregation based solely on the qualitative evaluation, would favour persuasive explanations over fidelity to the model, as discussed in Section 2. Therefore, our qualitative results are interpreted while taking into account the quantitative evaluation proposed in Section 5.2. A prerequisite for useful global insights gained from GALE is that the global aggregation reliably represents a model's global behavior.

6 Results

This section aims to answer the subquestions described in Section 5. Firstly, we will assess how well the global aggregations proposed in Section 4.2 represent global model behavior to answer our Quantitative Research Questions. Secondly, visualizations are presented to demonstrate global insights on important and distinguishing features, which will answer our Qualitative Research Questions.

6.1 Quantitative evaluation of GALE

We intend to evaluate if features that are considered globally important according to the aggregations, indeed have a large impact on model predictions and performance. For each of the global aggregations, features are removed in order of global importance, and compared against a random baseline for which the removed words are selected at random. The random baseline is averaged over five runs; the variance is also shown in the result plot. The evaluation is carried out for both classification tasks described in Section 5.1. The results for the sentiment analysis task are shown in Figure 6.1, where up to 20 features are removed per sentence. Figure 6.2 presents the results for the 20 Newsgroup text classification task. Since this task entails larger documents, results are shown for up to 50 features removed per document.

As expected, the ACPD and the ACAD generally display corresponding results. In both Figures 6.1 and 6.2, it can be seen that the global LIME aggregation obtains only slightly higher ACPD and ACAD results than the random baseline. In particular it is found that the initial steepness of all LIME importance curves is equal to the initial steepness of the baseline. This indicates that the global LIME aggregation especially fails to correctly identify the most important features; the first features removed in the evaluation affect the models' predictions and performance no more than average. Nonetheless, ACPD and ACAD for the global LIME importance are higher than the random baseline over the whole range of feature removals. This reflects that on average the aggregation is able to represent the global influence of features on the model to some extent. This finding is in accordance with the evaluation of the submodular pick algorithm presented by Ribeiro et al. (2016b), which is also based on global LIME importance. Here, presenting users with particular explanations based on this global aggregation was compared against random selection of presented explanations; this was shown to improve users' ability to make choices to positively affect model performance. As we expected, this evaluation did not really grasp how well the aggregation represents the model. Our findings indicate that global insights through aggregation can be improved by selecting an aggregation function that better represents the local explanations with respect to the global model behavior.

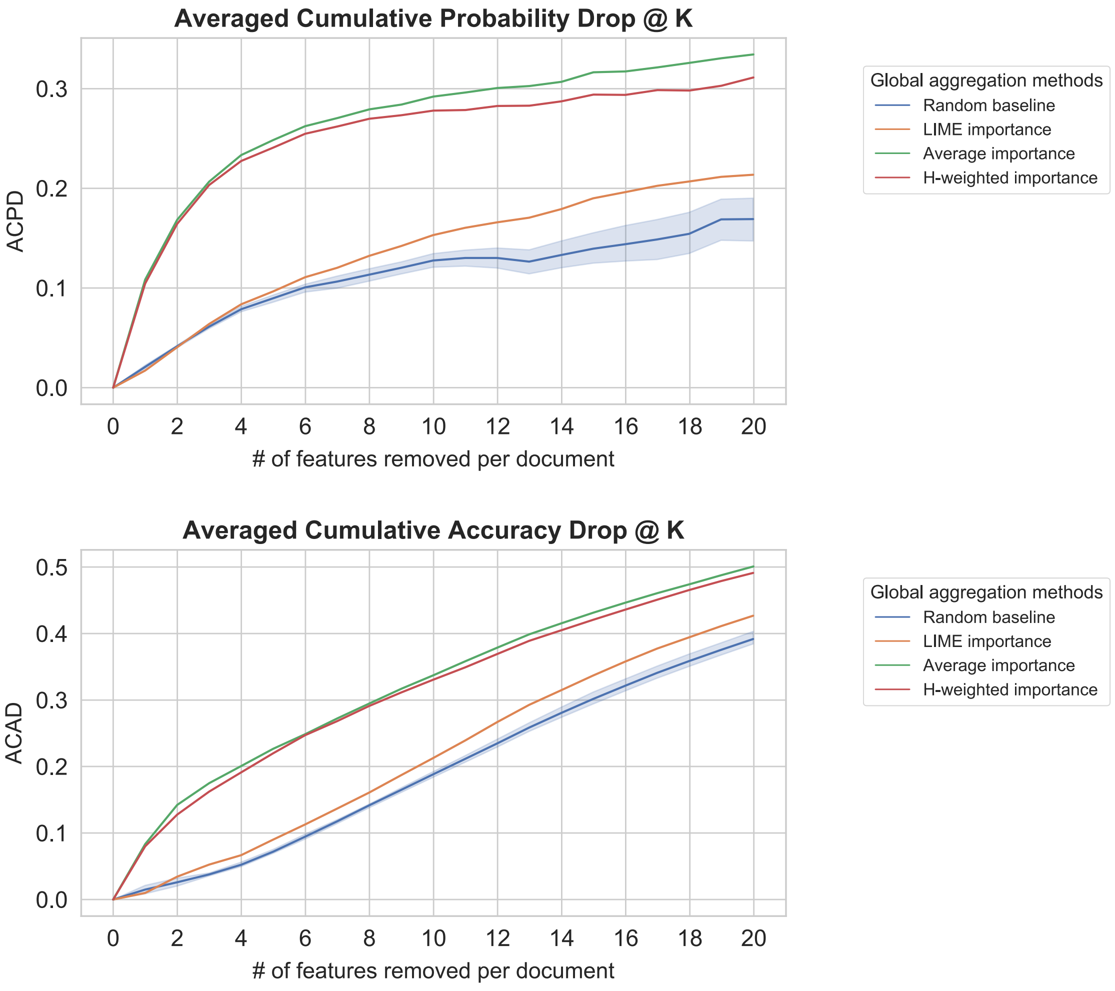
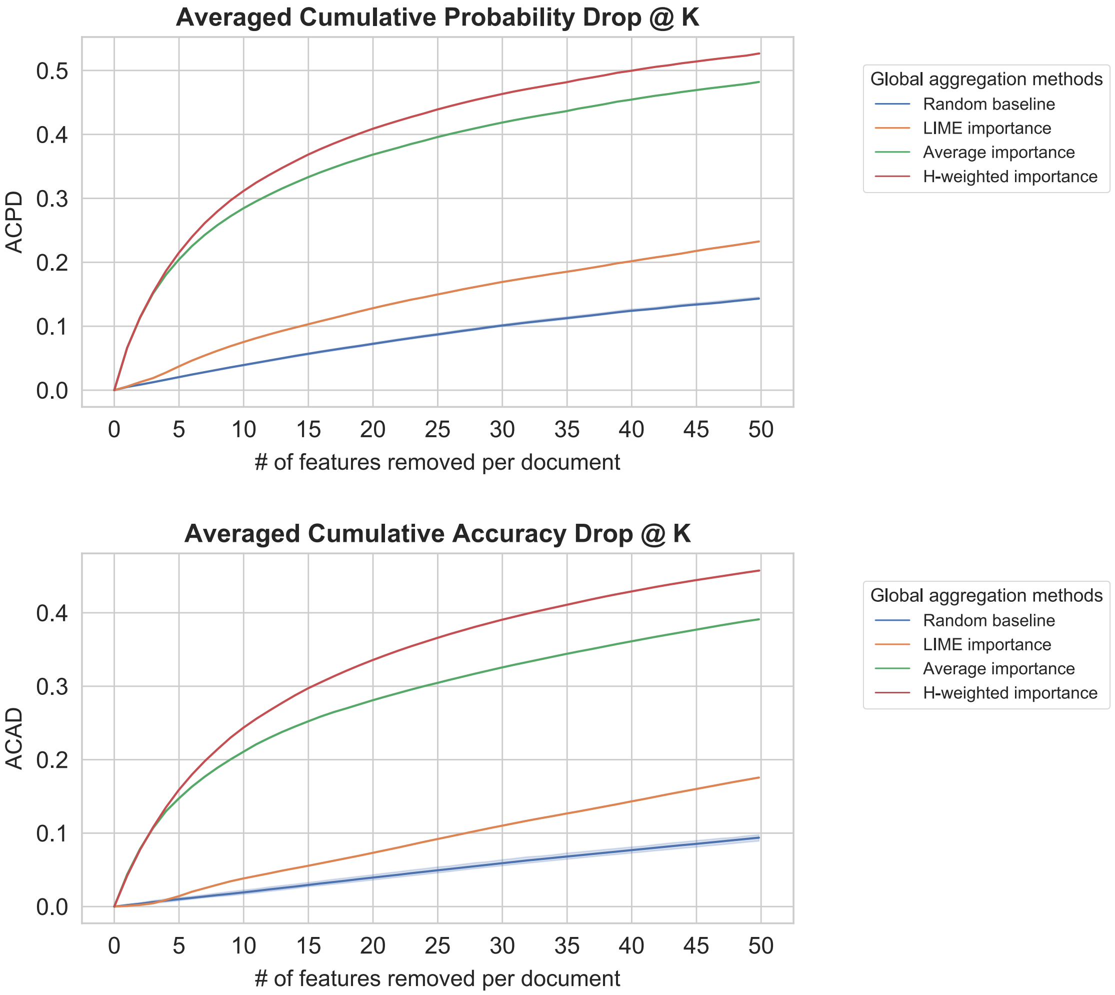


Figure 6.1: Quantitative evaluation of GALE on the sentiment analysis task. On the top the Averaged Cumulative Probability Drop @ K is presented over a range of feature removal K up to 50. Similarly, the Averaged Cumulative Accuracy Drop @ K is displayed on the bottom.

Quantitative Finding 1

Global LIME importance performs only slightly better than the random baseline, indicating that the global LIME importance aggregation is not able to reliably represent how features affect a model's global decision making process.



Averaged Cumulative Probability Drop @ K

Averaged Cumulative Accuracy Drop @ K

Figure 6.2: Quantitative evaluation of GALE on the 20 Newsgroups classification task. On the top the Averaged Cumulative Probability Drop @ K is presented over a range of feature removal K up to 50. Similarly, the Averaged Cumulative Accuracy Drop @ K is displayed on the bottom.

Both the global average importance and the global homogeneity-weighted importance surpass the ACPD and ACAD values obtained by global LIME importance and the random baseline, demonstrating that these aggregations are more able to reliably represent the global model behavior. Additionally, for both aggregations, both the ACPD and ACAD curve obtained, display a much steeper initial slope of the curve. This finding implies that the global average importance and global homogeneity-weighted importance more adequately identify the most important features; these aggregations are better able to rank the features based on their global influence on model predictions and performance. This is evidence for our hypothesis that the global LIME importance is based on misleading assumptions. More general, it reveals that the choice of aggregation function matters regarding its ability to represent a model's global decision making process.

Quantitative Finding 2

Global average importance and global homogeneity-weighted importance perform better than global LIME importance, showing they are better able to rank features in the order of their global influence on the model. This indicates that these aggregations better represent the models' global decision making process.

The average importance aggregation performs slightly better on ACPD than the homogeneity-weighted importance in case of the sentiment analysis task. On the contrary, the homogeneity-weighted importance displays a steeper curve for both ACPD and ACAD in the document classification task. A possible explanation is that this is because the sentiment analysis task is a binary classification task, while the 20 Newsgroup classification is a multiclass classification task. In a binary classification a local explanation for a particular instance informs about the feature influence for all possible class predictions - there are only 2 classes. The attribution for a feature is either in support of the predicted class or against it, in the latter case this signifies support for the opposing class. In the case of multiclass classification, local explanations only provide an explanation for the influence of features in light of the predicted class. The global average importance of a feature that is influential for some classes would be significantly lowered due to low attributions in explanations for other classes. The global homogeneity weighting factor is more appropriate when explaining a multiclass classification model, because the weighting factor is affected by the spread of a feature's attributions over different classes, specifically to the degree of uniformity of that distribution. The effect is that global importance is reduced more for features that obtain high attributions for multiple classes than for features with high attributions for one class and low attributions in explanations for other classes.

Quantitative Finding 3

Global homogeneity-weighted importance best represents the model's global decision making process for the multiclass document classification task. There is only a small difference in performance between global average importance and global homogeneity-weighted importance in the binary sentiment analysis task.

6.2 Qualitative visualization of GALE

To deepen our understanding of which global aggregation for local explanations best provides global insight in a complex model, several visualizations are presented as described in Section 5.3.

Figure 6.3 demonstrates the 25 most influential features per class according to global LIME importance - 50 features in total. On the upper left corner a cluster of positive and a cluster of negative words can be observed. Further along towards the lower right corner, words seem to become more common and the distinction between positive and negative clusters of words is less evident. Similar observations can be made for Figure 6.6, which presents the 15 most influential features per class for the 20 Newsgroups data - 300 features in total, as obtained by taking the highest scoring features in LIME importance per class. Clusters of features with the same color can be noticed, indicating those features are attributed a high global feature importance for the same class. The words in these clusters are similar in their word embeddings and seem reasonably related to the topic of a class (e.g. "space", "shuttle" and "solar" for the class sci_space). On the other hand some clusters consist of the same word considered important for multiple classes (e.g. "the", "and", ")"). Based on human intuition we would not consider this latter set of words to be indicative of a particular class. This in itself is not the right reason to dismiss the explanation or aggregation approach, since the model might in fact be influenced by these features. However, combined with our quantitative findings indicating global LIME importance is not able to reliably represent how features affect a model's behavior, it is concluded that global LIME importance is not a reliable and useful way to aggregate local explanations for global insight.

Qualitative Finding 1

The presented visualizations for global LIME importance contain class-specific clusters of distinguishing features, as well as less substantive features, e.g. common words and punctuation, that do not appear in clearly distinct clusters. Both features that are likely and unlikely to distinguish between classes, are deemed important by the global LIME aggregation.

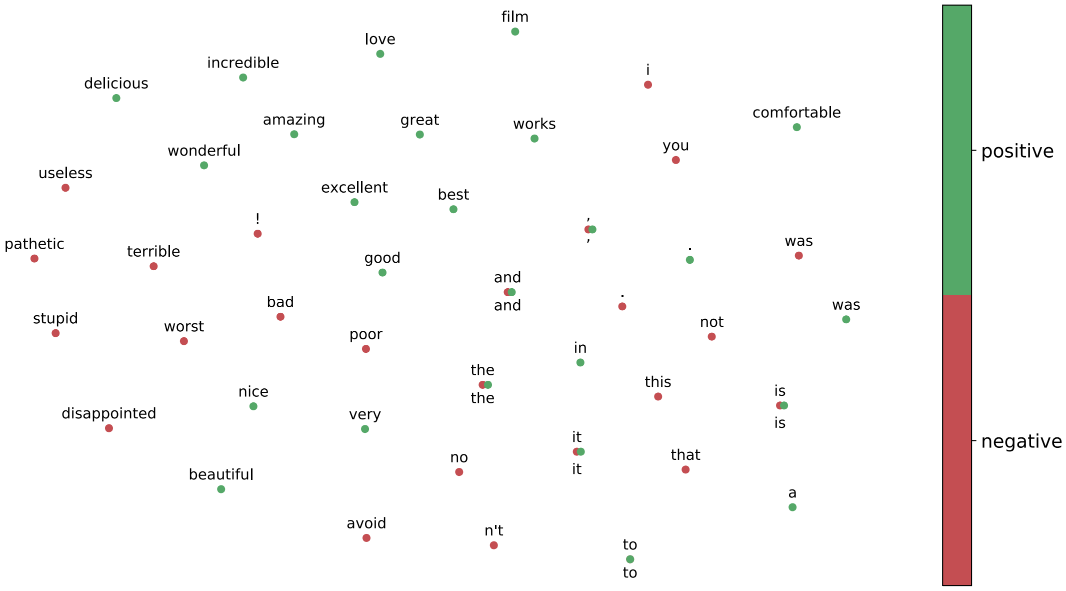


Figure 6.3: Top 25 features per class according to global LIME importance on the sentiment analysis task.

As is concluded in Section 6.1, global average importance and global homogeneity-weighted importance better represent the models' global decision making process than global LIME importance. The open question is whether or not this leads to more useful global insights; specifically we intend to examine if global average importance and global homogeneity-weighted importance are better able to identify distinguishing features. Since there are similarities in the visualizations of these aggregations for the sentiment analysis task, we will discuss these first, after which the visualizations for the document classification task are discussed separately.

In both Figure 6.4 and Figure 6.5, displaying the top 25 features per class for the global average and homogeneity-weighted aggregations, two major clusters appear: a negative cluster on the left and a positive cluster on the right. This is considered a strong indication of distinguishing

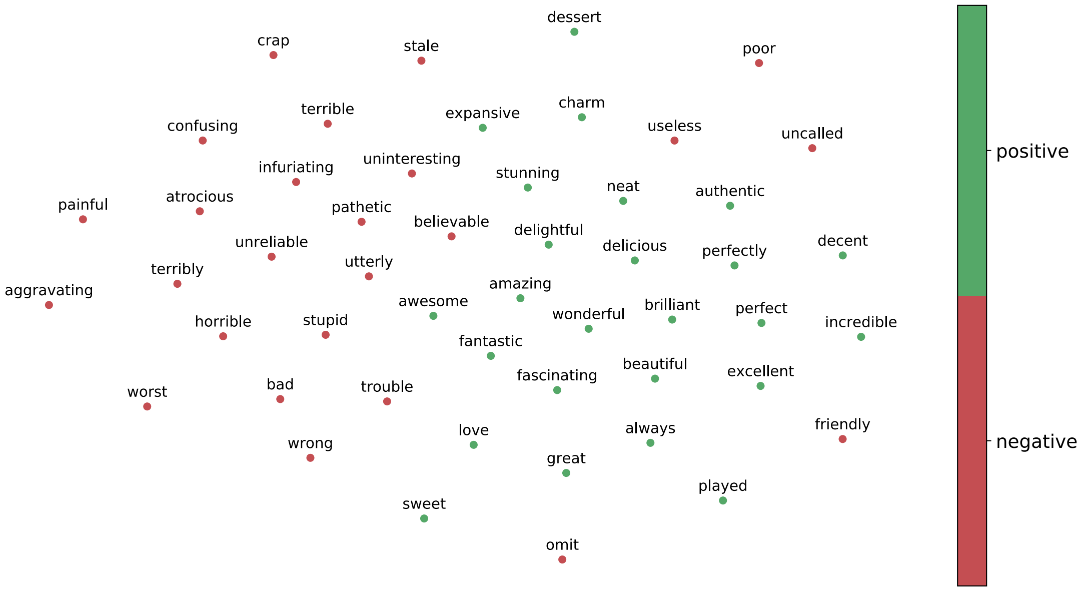


Figure 6.4: Top 25 features per class according to global average importance on the sentiment analysis task.

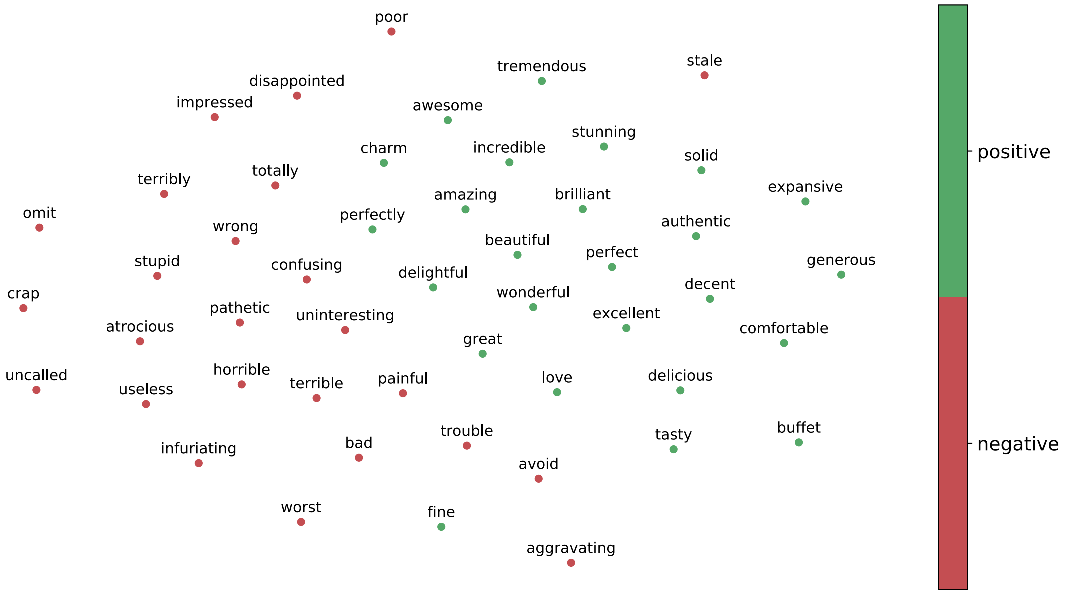


Figure 6.5: Top 25 features per class according to global homogeneity-weighted importance on the sentiment analysis task.

features. Moreover, the visualization in Figure 6.4 and Figure 6.5 display more substantive words; they do not contain the punctuation marks and common words, e.g. “the” and “of”, that appear in Figure 6.3. Some of the features are the same as in Figure 6.3, such as “terrible”, “bad”, “great” and “amazing”. These are probably the most likely words to appear in documents that express a sentiment. Figure 6.4 and Figure 6.5 also contain features that probably occur less often but clearly express a sentiment, such as “atrocious”, “generous” and “infuriating”.

Figure 6.7 presents the top 10 features in the document classification task according to global average importance - 200 features in total. As with the sentiment analysis task, there are still clusters of features than seem reasonably related to the topic, and there no longer are clusters of common word features that are considered to be important for multiple classes. However, the visualization for global average importance does not demonstrate substantially more clusters of distinguishing features. Some of the most distinct clusters are the sci_med and and talk_politics_guns classes on the left, and the sci_space class on the bottom right of Figure 6.7. Overall, these findings agree with the results presented in Section 6.1 in that the global average importance better represents the models' global behavior than global LIME importance. The visualization demonstrates that this indeed leads to more useful insights.

Qualitative Finding 2

The presented visualizations for global average importance contain class-specific clusters of distinguishing features for both classes in the sentiment analysis task, and for a minority of the classes in the document classification task. The global average aggregation considers substantive features important, i.e. no common words and punctuation, which more evidently distinguish between classes in the sentiment analysis task than in the document classification task.

Lastly, let us compare Figure 6.6 and 6.7 to Figure 6.8, which shows the 10 most influential features per class according to the homogeneity-weighted importance. Again, the visualized features are more substantive compared to global LIME importance, and there are no punctuation marks and common words among the features considered most important. Furthermore, we are able to distinguish approximately 15 class-specific clusters of features. It is concluded that in comparison to the other global aggregations, global homogeneity-weighted importance is best able to identify distinguishing features. This is in line with the results presented in Section 6.1, where global homogeneity-weighted importance is shown to better represent the model's global behavior in the document classification task. These findings indicate that global homogeneity-weighted importance is the best aggregation of local explanations to obtain global insight on a complex model.

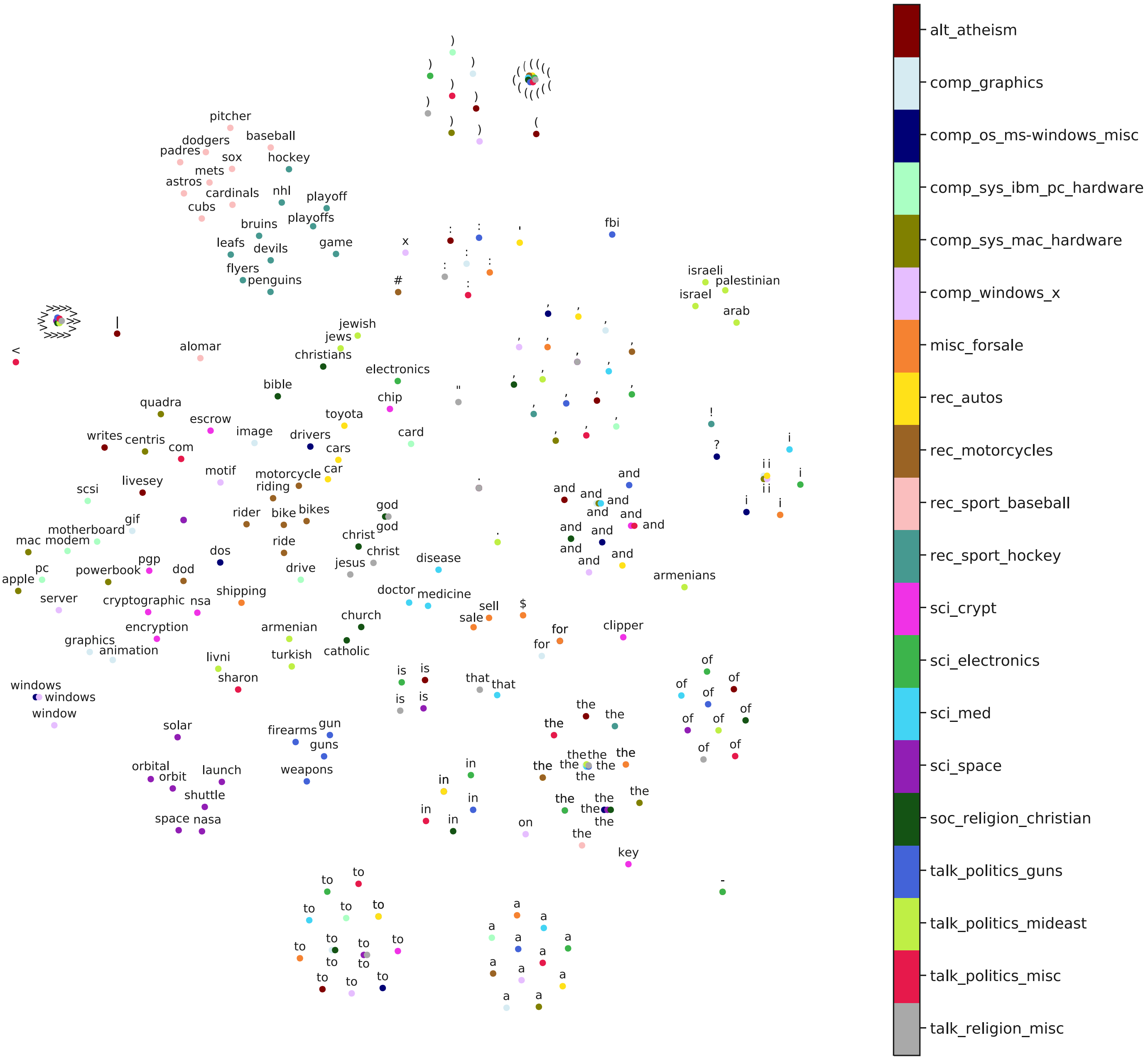


Figure 6.6: Top 15 features per class according to global LIME importance on the 20 Newsgroups document classification task.

Qualitative Finding 3

The presented visualizations for global homogeneity-weighted importance contain class-specific clusters of distinguishing features for a majority of the classes in both tasks. Features deemed important by the global homogeneity-weighted aggregation are substantive features that distinguish between classes.

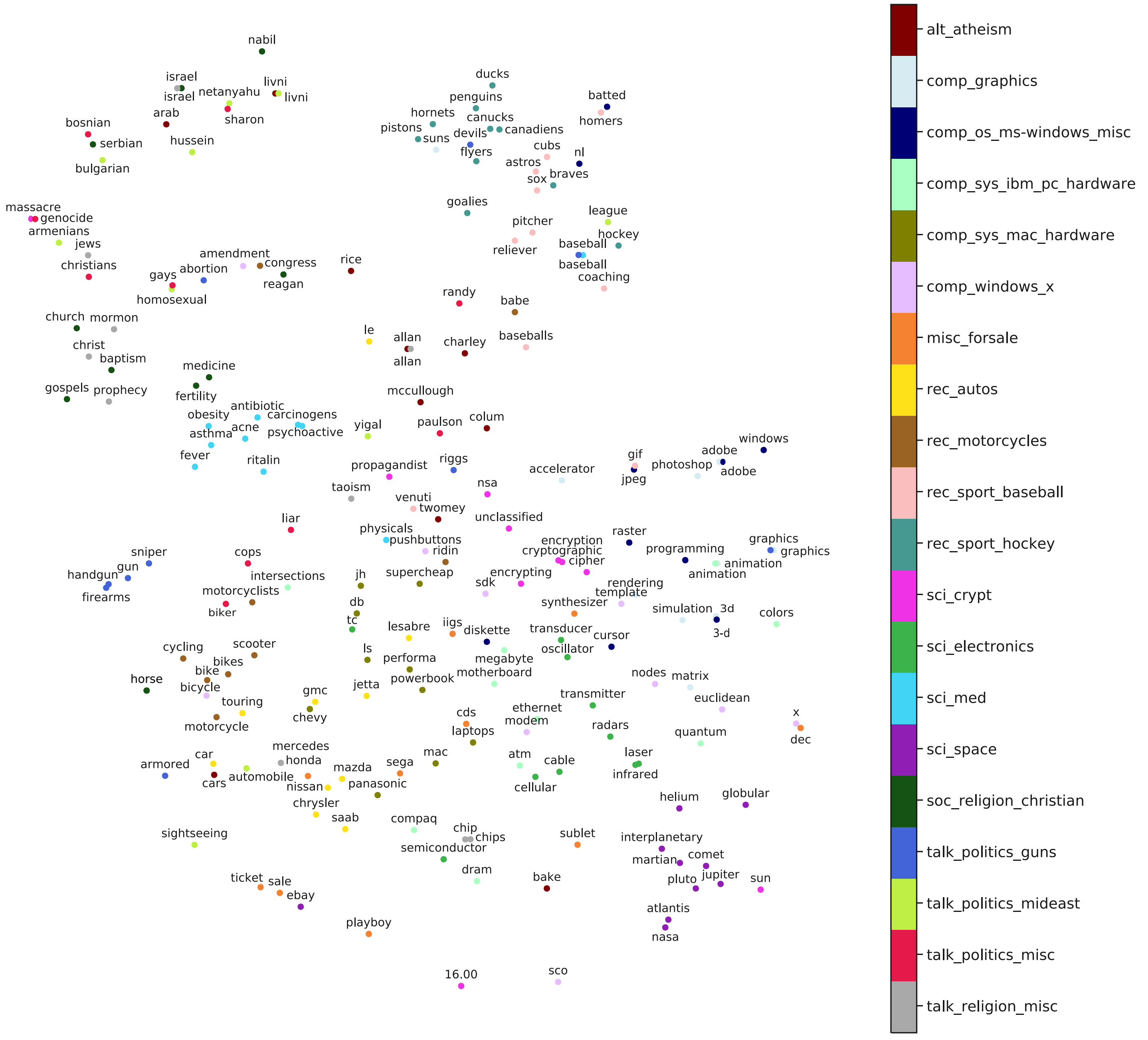


Figure 6.7: Top 10 features per class according to global average importance on the 20 Newsgroups document classification task.

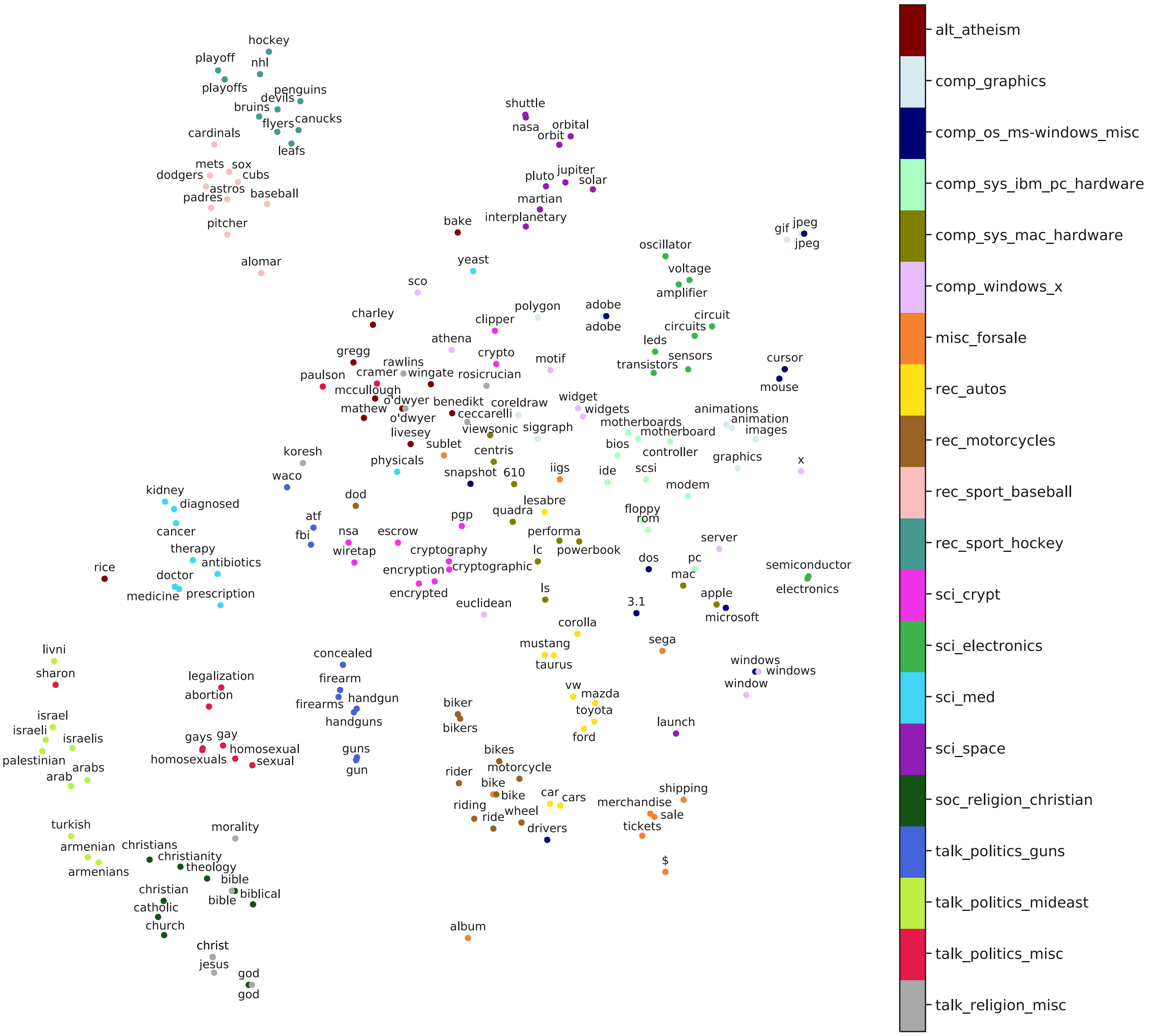


Figure 6.8: Top 10 features per class according to global homogeneity-weighted importance on the 20 Newsgroups document classification task.

7 | Conclusion & Discussion

In this thesis Global Aggregations of Local Explanations are proposed. The primary objective of GALE is to provide insights in a model's global decision making process. In our quantitative evaluation in Section 6.1 we assess to what extent each of the proposed global aggregations is able to represent how features affect the models' predictions and performance. The qualitative visualizations presented in Section 6.2 demonstrate to what extent each of the aggregations is able to identify distinguishing features. Our final conclusions are based on the combined quantitative and qualitative findings. These are shared first, followed by a discussion of implications and future work.

Our first conclusion, is that global LIME importance is not a reliable and useful way to aggregate local explanations for global insights on a black box model. This follows from both the quantitative findings, showing the aggregation performs only slightly better than the random baseline on both ACPD and ACAD, and the qualitative findings, indicating that the global LIME aggregation considers features important that are unlikely to distinguish between classes. Both global average importance and global homogeneity-weighted importance demonstrate better results on all accounts. The quantitative results show these aggregations are able to better represent how features affect the models' predictions - as measured by ACPD, and performance - as measured by ACAD. This is also reflected by the visualizations of most important features per class according to each of the aggregations, in the distinguishing features that are considered important and the appearance of class-specific clusters of features. Specifically, global homogeneity-weighted importance outperforms global average importance in the document classification task, and is therefore considered the best aggregation approach. Overall, we conclude that Global Aggregations of Local Explanations have the potential to provide global insights from local explanations. In addition to this, the findings reveal that the choice of aggregation matters regarding the ability to gain reliable and useful global insights on a black box model.

Apart from providing insights in the globally important features as demonstrated by the visualizations in this thesis, GALE opens up other possibilities to gain insights. GALE could potentially increase reliability and usefulness of local explanations by increasing understanding of their limitations. Information about the representativeness of individual explanations might help to comprehend and mitigate the gap between local explanations and global model behavior. For instance, the global homogeneity-weighted aggregation incorporates what part of an explanation is generalizable to other instances, i.e. high homogeneity, or what part has a varying influence on the model, i.e. low homogeneity. Alternatively, the global average importance could reflect what part of an explanation is typical to the example and diverges from the general, i.e. large difference between average importance and local feature attribution. Future work could asses how useful GALE is for these types of insights.

Furthermore, our work offers opportunities for future work to assess GALE on different tasks, different explanation methods, and experiment with different aggregation functions. As our results show, the choice of aggregation matters in regard to the ability to gain global insights on a black box model. In Section 4.2.1 two underlying assumptions of the global LIME importance are discussed. Our aggregations are derived to address the shortcomings of the second assumption, i.e. a feature that occurs more often considered more important, since we deem this assumption the most problematic for text classification tasks. Nonetheless, we already explained why the first assumption, i.e. features with higher attributions expected to have a larger effect, might not hold in all occasions either. Evidently, the aggregations proposed in this thesis also make assumptions about the way in which local explanations are representative of the global model behavior. These are open ends that could inspire other global aggregation functions. Future work could follow the procedure outlined in our methodology. Determine which assumptions are likely or unlikely to hold given the domain of the task, and design global aggregations accordingly. Apart from finding better aggregation approaches, this will help increase our understanding of the representativeness of local explanations, the emerging patterns they reflect, and the limitations of their locality.

Bibliography

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), e0130140.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., … Zhang, J., et al. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Boyd, D. & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662–679.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Goodman, B. & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222–2232.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 93.

Herman, B. (2017). The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414.*

Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078.*

Kim, B., Rudin, C., & Shah, J. A. (2014). The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in neural information processing systems* (pp. 1952–1960).

Korczak, J. & Hemes, M. (2017). Deep learning for financial time series forecasting in a-trader system. In *2017 federated conference on computer science and information systems (fedcsis)* (pp. 905–912). IEEE.

Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015). From group to individual labels using deep features. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 597–606). ACM.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1675–1684). ACM.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490.*

Lipton, Z. C. (2017). The doctor just won't accept that! *arXiv preprint arXiv:1711.08037.*

Lundberg, S. & Lee, S.-I. (2016). An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478.*

Lundberg, S. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4768–4777).

Maaten, L. v. d. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Mittelstadt, B., Russell, C., & Wachter, S. (2018). Explaining explanations in ai. *arXiv preprint arXiv:1811.01439.*

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: Definitions, methods, and applications. *arXiv preprint arXiv:1901.04592.*

Pang, B., Lee, L. et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135.

Passi, S. & Jackson, S. J. (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 136.

Pennington, J., Socher, R., & Manning, C. D. (2018). Glove: Global vectors for word representation. 2014. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014).*

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM.

Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, *28*(11), 2660–2673.

Selbst, A. D. & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev. 87*, 1085.

Selbst, A. D. & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 3145–3153). JMLR. org.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.