Limitations of Feature Attribution in Long Text Classification of Standards

Katharina Beckh[1,2], Joann Rachel Jacob[1], Adrian Seeliger[3], Stefan Rüping[1], Najmeh Mousavi Nejad[1]

[1]Fraunhofer IAIS
[2]Lamarr Institute for Machine Learning and Artificial Intelligence
[3]Deutsches Institut für Normung e. V. (DIN)
katharina.beckh@iais.fraunhofer.de

Abstract

Managing complex AI systems requires insight into a model's decision-making processes. Understanding how these systems arrive at their conclusions is essential for ensuring reliability. In the field of explainable natural language processing, many approaches have been developed and evaluated. However, experimental analysis of explainability for text classification has been largely constrained to short text and binary classification. In this applied work, we study explainability for a real-world task where the goal is to assess the technological suitability of standards. This prototypical use case is characterized by large documents, technical language, and a multi-label setting, making it a complex modeling challenge. We provide an analysis of approx. 1000 documents with human-annotated evidence. We then present experimental results with two explanation methods evaluating plausibility and runtime of explanations. We find that the average runtime for explanation generation is at least 5 minutes and that the model explanations do not overlap with the ground truth. These findings reveal limitations of current explanation methods. In a detailed discussion, we identify possible reasons and how to address them on three different dimensions: task, model and explanation method. We conclude with risks and recommendations for the use of feature attribution methods in similar settings.

Introduction

Explaining the decision process of machine learning (ML) models is ever more important amidst the implementation of regulatory policies. However, due to the complexity of language models in terms of parameters, they are inherently difficult to interpret. To better understand model predictions, various methods have been developed to provide more insight into model behavior (Burkart and Huber 2021; Danilevsky et al. 2021). For this, explainability is said to play a key role in increasing transparency. Explainability refers to the ability of a ML model to explain or present model output in human-understandable terms (Doshi-Velez and Kim 2017). In the field of natural language processing (NLP), use of explainability is extensively studied and the most common method is feature attribution which provides importance scores for each feature in the input (Danilevsky et al. 2021; Bastings and Filippova 2020; Lu et al. 2024). However, the focus is commonly constrained to toy problems with the following characteristics: (a) short text, often only one sentence or a paragraph, and (b) binary classification. This setting is appropriate for certain use cases, but often, the cognitive demands in document processing emerge with longer text. In applied settings, the aim is to ease time-consuming tasks, e.g., in health care to reduce documentation time (Hou et al. 2024), and in legal and administrative sectors to categorize documents and extract relevant information (Valvoda and Cotterell 2024). So far, the applicability of feature attribution methods and their evaluation has not been sufficiently studied for long documents. In addition, there is too little discussion about the consequences of resource demands for post-hoc explanations.

In this applied work, we investigate explainability for a real-world task concerned with assessing the AI readiness of standards and specifications. AI readiness refers to the question of whether standards enable the use of AI or whether modifications are necessary to fully leverage the potential (Görge et al. 2024). This setting involves long documents, technical language, and a multi-label setting, which makes it a complex modeling and explanation task. It represents a prototypical application scenario in enterprises.

As a data basis, we have approx. 1000 annotated documents with an average length of roughly 6000 tokens. The documents are considerably longer compared to typical datasets in the field. The dataset contains both document labels and annotated explanations. We analyze the human-annotated explanations with respect to length, position, and number per document. We find that the explanations are dispersed throughout the document, which requires processing the documents in their entirety.

In experiments, we evaluate a language model and two common explainability methods. For the explanation component, we examine runtime and compute the degree of overlap with the ground truth annotations provided by human experts. Our findings indicate that the average runtime is high, exceeding practical limits for real-time applications. We show that the explainability methods match the ground truth only barely or not at all. Overall, our results suggest that current explainability methods are ill-suited for the explanation needs in this and possibly similar settings. Along the dimensions of task, model and explanation method we discuss possible reasons and implications.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Related Work

Classification for Long Documents

Document classification is the task of assigning one or more categories to a text. Recent work studied long document classification in medical and legal domains (Valvoda and Cotterell 2024; Hou et al. 2024). An important decision is the model choice because not every method can process the complete document. Three different approaches can be found in the literature. The baseline approach combines TF-IDF (Term Frequency-Inverse Document Frequency) with either a linear classifier (Mamakas et al. 2022) or a neural network (Tuteja and González Juclà 2023). In terms of classification performance, this approach competes with others while requiring little compute resources. The more advanced approaches rely on transformer variants (Vaswani et al. 2017). A version of BERT (Devlin et al. 2019; Pappagari et al. 2019) was used, which creates chunks and encodes them together in a hierarchical way. Longformer models (Beltagy, Peters, and Cohan 2020), which can handle longer sequences than BERT, performed best overall for corpora with texts exceeding 700 tokens, although they require longer training times (Park, Vyas, and Shah 2022; Mamakas et al. 2022; Tuteja and González Juclà 2023). Considering the datasets in prior work, the token range is typically between 500 to 5000. In this work, the average token count is roughly 6000, exceeding typical experimental settings.

Explainability

Explainable AI, or explainability, refers to approaches presenting ML output in a human-comprehensible form and is an active field of research (Danilevsky et al. 2021; Atanasova et al. 2020; Burkart and Huber 2021; Nauta et al. 2023; Lu et al. 2024). The study focuses on post-hoc explanations, which are applied to a trained model.

There are a number of studies that explore explainability for text classifiers. However, these have been performed primarily on short text and binary classification, with the majority of datasets coming from social media and product reviews (Wiegreffe and Marasović 2021; Mendez Guzman, Schlegel, and Batista-Navarro 2024; Herrewijnen et al. 2024). A few works have investigated explainability for longer text classification, but they were limited to binary settings or lacked ground truth (Bhambhoria, Dahan, and Zhu 2021; Stremmel et al. 2022). Often, ground truth is not available (Gurrapu et al. 2023; Nauta et al. 2023). Standard evaluation methods for measuring how well model explanations agree with those marked by humans are Intersection-Over-Union and F1 scores on a token level (DeYoung et al. 2020). For long texts, these metrics may still be too strict, so we consider more lenient evaluation metrics which treat a token in proximity to the ground truth as a match.

In this study, we utilize ground truth annotations to explore the applicability of post-hoc explanation methods on long, technical text in a multi-label setting.

Use Case Description

Explainability is important for the task of assessing standards towards their technological readiness, which we motivate and describe in the following. Standards establish quality benchmarks and promote interoperability. They form the basis for the successful integration of technology into existing and new production processes and applications.

As AI emerges in almost every sector, it is crucial to ensure that existing standards are *ready* for AI integration. This means that standards should not unjustifiably restrict AI use or fail to address specific risks and requirements of novel technologies. Since it is not easy to predetermine which standards are *AI ready*, assessing existing standards is as important as creating new ones.

Prior work formalized AI readiness by building on existing AI definitions, analyzing several standards and consulting experts (Görge et al. 2024). Key steps in the process included defining *AI relevance*, whether a standard is related to the application of AI, and several levels of *AI readiness*. This resulted in the following five classes: Not AI relevant, AI ready, almost AI ready, marginally AI ready and not AI ready. In addition to these classes, a flowchart was created to assist the label process. Decisions are made along the nodes in the flow chart. For example, one node refers to the question of whether the use of AI is discouraged. This is true if, for instance, a visual inspection by a human is required. With a tool to support the classification of standards into appropriate AI readiness levels, the overall goal is to identify documents which require attention and possible revision. The AI-based tool is intended to be used in human-AI decision making, reducing the assessment burden for the reviewer. For this task, the comprehensibility and verifiability of the model's decision play a crucial role.

Domain	Documents
Mechanical Engineering	294
Automobile	377
Medicine	439
Artificial Intelligence	14

Table 1: Document counts according to domains

Methods

In the following, the data characteristics, choice of methods, classification procedure, and evaluation are described.

Dataset

The dataset is a collection of 1124 documents in the German language, comprising DIN, DIN EN and DIN EN ISO standards, technical reports and specifications.[1] The average token count is roughly 6000. The documents cover three domains: Mechanical engineering, automobile and medicine. Additionally, standards from the field of AI were included to represent standards that are AI ready. Table 1 shows the respective document counts.

We differentiate between final labels and path labels. Final labels are the five classes: not AI relevant, not AI suitable, marginally AI ready, almost AI ready, AI ready. Path labels

[1]Due to its proprietary nature, the dataset cannot be made publicly available.

Path label description	Nodes
Legal restrictions	E1, K4
Use of AI for the ”main topic” of the document is feasible	K5.1a, K5.2a
High-level / technology-unspecific	E3, K5.1, K5.2
Use of AI is discouraged or implicitly excluded	EK2.1, EK2.2, K6
Standard developed from ISO/IEC JTC 1/SC 42, CEN/CENELEC JTC 21 or Joint Committee on AI	A1
AI or AI related topics is explicitly mentioned in the standard document	A2
Special requirements for the use of AI (or the AI technology mentioned) are listed	E2
Tasks or the required quality assurance of systems or processes can potentially be taken over by AI	K1
AI use is relevant in practice	K2
Document has an impact on the life cycle of an AI system, new requirements or new risks	K3
Document refers to other horizontal AI standards that cover risks	EK1

Table 2: Description of overarching topics and node labels. The nodes that comprise a topic are taken together to form a label.

derive from the nodes in the flowchart, e.g. A1, K1 etc. (see Görge et al. (2024) for details on the flowchart). Table 2 lists all path labels. Some of the nodes answer the same underlying question. For example, EK1.2, EK2.2, and K6 all refer to the topic that the use of AI is not recommended or excluded. In an effort to reduce redundancy, the overarching topics serve as path labels. This way, some labels contain several nodes.

Each document received a final label and, if applicable, one or more path labels. Hence, the problem is modeled as a multi-label classification task. All path labels are treated as document labels. This refinement is aimed at implementing a classification module, where the information captured by individual annotations contributes to a characterization of the entire document. In addition to the document labels, explanations were annotated wherever possible, i.e., textual evidence for a respective path label. Figure 2 shows an example of textual evidence for path label K1.

Due to restricted resources, the annotation was performed in a *sufficient* manner. That means that annotators were instructed to find and annotate evidence that is enough to justify a label. Consequently, it is not the case that all relevant text for a label is annotated. With long documents, this is a necessary means to handle available annotation resources. At the same time, this poses an evaluation challenge which we address in more detail in the discussion.

Annotation Analysis

The characteristics of human-annotated explanations, i.e., annotation spans for a specific label, inform modeling decisions. In particular, the goal was to find out whether the annotations are clustered in specific locations. This would allow us to focus on these sections and, thus, reduce the complex problem to a simpler, well-studied problem.

As a first step, we aimed to understand the annotation characteristics of the explanations, in particular the average length, position in the document, and number of annotations per document. For each label, we calculated the average character count to gain insight into the length of annotations. The average character count is roughly between 10 and 500. Since the average German word is around 10 characters long, we can deduce that annotation length ranges from a single word to several sentences.

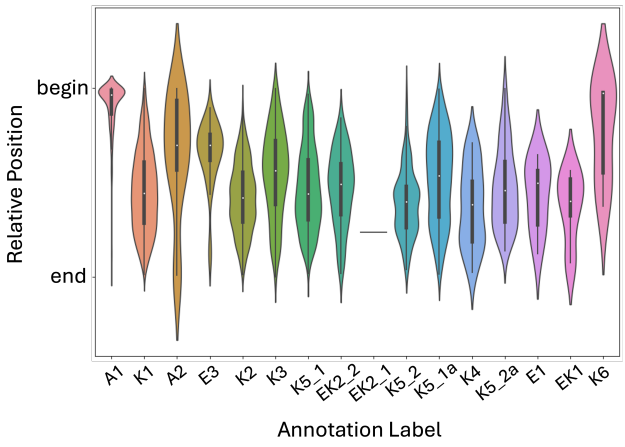

Figure 1: Distribution of annotations w.r.t. relative position

Regarding the occurrence of annotations, certain labels exhibit less evidence. Roughly half of the labels have below 60 textual annotations. Labels E2 and EK1 do not or not often occur, which is consistent with their text annotation frequency. However, A2, E1, K4, EK2.1, EK2.2 and K6 are assigned more than 100 times as document labels without textual annotations. This suggests that the information needed for the label decision lies outside of the document, such as in a reference or as external knowledge. Consequently, these labels are more difficult to model and also to explain.

Furthermore, we examined the positions of annotations per label across documents. We were especially interested in this because we considered shortening the documents if most of the relevant information is contained in specific sections. Reducing text has the advantage to reduce computational resources and opens up more modeling options.

Figure 1 shows the typical position of label annotations across all documents for each label. The positions are calculated based on the average of begin and end of an annotation, divided by the document length. The line for EK2.1 indicates that only one annotation exists.

In general, annotations are distributed throughout the entire document. However, there is a clear tendency for A1 to be annotated at the start. This is because the standards

"[X.X.4] Inkubation und Auszählen des Prüfgemischs
Das Verfahren zur Inkubation und zum Auszählen des Prüfgemischs ist folgendermaßen:
a) Die Platten sind für 20h bis 24h zu bebrüten (5.3.2.3). Alle (aus einem beliebigen Grund) nicht auszählbaren Platten sind zu verwerfen. Die Platten sind durch das Bestimmen koloniebildender Einheiten (KBE) auszuzählen . Die Platten sind für weitere 20h bis 24h zu bebrüten. Platten, die keine gut voneinander abgetrennten Kolonien mehr zeigen, sind nicht neu auszuzählen. Die verbleibenden Platten sind neu auszuzählen. Sofern die Anzahl angestiegen ist, ist die für die weitere Auswertung lediglich die höhere Anzahl zu verwenden."

"[X.X.4] Incubation and counting of the test mixture
The procedure for incubation and counting of the test mixture is as follows:
a) The plates are to be incubated for 20h to 24h (5.3.2.3). All plates that cannot be counted (for any reason) are to be discarded. The plates are to be counted by determining the colony-forming units (CFUs) . The plates are to be incubated for another 20h to 24h. Plates that no longer show well-separated colonies are not to be recounted. The remaining plates are to be recounted. If the number has increased, only the higher number is to be used for further evaluation."

Figure 2: Source: DIN EN 12791:2018-01. The top shows the original text in German and the bottom is an English translation. Example of textual explanation for path label K1 signifying that tasks or the required quality assurance of systems or processes can potentially be assumed by AI.

committee is mentioned directly in the beginning, and the annotators were instructed to annotate at that position. A2, E3 and K6 also tend to be annotated in the first half of the document. In contrast, evidence for K4 seems to occur more towards the end of the document.

From these analyses, we summarize the takeaways: Explanations range from one word to several sentences. Some labels are assigned less often or have less evidence, making modeling and explanation for these labels more challenging. No overall position pattern emerged, which necessitates keeping the complete document for further processing.

Classification

Choice of Model The requirements for the language model were to handle long text in German and English, preferably an encoder architecture to ensure precision, and availability as pre-trained model. Most transformer-based models are constrained regarding input length as the self-attention scales quadratically with input length. A common choice is BERT (Devlin et al. 2019) which has a token limit of 512 tokens which on average is less than 400 words and, thus, less than a page of text. Therefore, BERT is less suitable for processing standard documents. We considered recent decoder models, however, decided against them for two reasons: (1) It would have considerably changed the modeling strategy, and (2) the applicability of available models, given the requirements for the down-stream task, was inconclusive. Driven by the findings of prior work (Park, Vyas, and Shah 2022; Mamakas et al. 2022), we opted for a Longformer model (Beltagy, Peters, and Cohan 2020) as it fulfilled nearly all requirements. This modified transformer architecture is suited for processing long text without sacrificing computational efficiency. A pre-trained version supporting the German language was selected.[2] Fine-tuning was performed on 80% of the data [3]

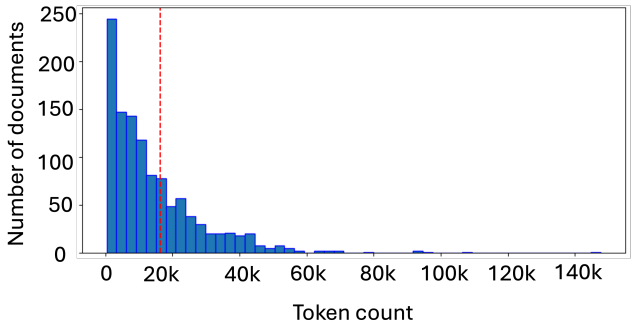

Figure 3: Document length measured by token count. The red vertical line depicts the length limit; all documents to the right of that line had to be truncated.

Text Processing With an input length of up to roughly 16000 tokens, the model was able to process the complete text of most of the documents. One limitation was that roughly one-third of the documents were above the token limit and had to be truncated. Figure 3 shows the token count sorted from lowest to highest. For clarity, several documents are depicted in one bin. The vertical red line shows the maximum token length of the model at 16384. All documents to the right of the line were shortened to the maximum length. In addition to truncation, simple text cleaning was performed removing words over 40 characters and punctuation chains ("..." and "___") if they occur more than twice in succession. This was done because PDF parsing in some cases resulted in undesired character sequences. According to the Duden dictionary, the longest German word is 44 characters long, the others are below 40 (Duden n.d.). Therefore, we chose a threshold of 40 characters.

Classifier Evaluation An 80-10-10 split was used for train, validation, and test. The split was performed such that the proportion of each final label in the train, validation and

[2]https://huggingface.co/severinsimmler/xlm-roberta-longformer-base-16384

[3]Hyperparameter settings: learning rate 5e-5, weight decay 0.01, 30 epochs, batch size of 2 and gradient accumulation steps set to 4 with gradient checkpointing enabled

test sets was the same. Labels which were not assigned positive or negative at least once were removed from the data.

Due to the multi-label setting with imbalanced label distribution, several evaluation metrics are used. We report subset accuracy and micro and macro versions of precision, recall, and F1 score. Subset accuracy only counts a prediction as correct if all labels are correctly predicted. Micro F1 treats classes according to their frequency, while macro F1 assigns equal importance to each class.

Explainability

Choice of Explanation Method As it is desired to better understand the model prediction for a particular standard document, the focus for the explainability component was on local post-hoc methods based on feature importance. These methods can be divided into two types: gradient-based and perturbation-based methods. We selected one method for each type. Integrated Gradients (Sundararajan, Taly, and Yan 2017) was chosen because of its completeness and additive properties, and because it was found to perform best in human-grounded evaluation (Lu et al. 2024). There are two popular perturbation-based methods, namely LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP, which is an approximation of Shapley values (Lundberg and Lee 2017). SHAP was preferred in user evaluation (Jesus et al. 2021) and was therefore chosen as the second method. We do not consider attention because it is more a measure of *sensitivity* as laid out by Bastings and Filippova (2020). For the implementation, the Captum library was used.[4] The parameters for both methods were set to 50, i.e., step size for Integrated Gradients and sample size in the case of Kernel SHAP.

Explanation Evaluation The evaluation of explanations is not trivial. As a qualitative measure, we perform a sanity check by checking the top-k tokens of correctly predicted classes. Regarding quantitative measures, we report runtime and overlap with the ground truth. Due to the possibility of partial overlap or a span that is close to the ground truth span, a context window w is introduced. This is a lenient measure which can capture whether a model explanation is near important information. Matches are calculated in the following way.

$$\text{match} = \text{True} \quad \text{if} \quad \big(\max(\text{begin}_{gt} - w, \text{begin}_{pred}) < \min(\text{end}_{gt} + w, \text{end}_{pred})\big)$$

Here, begin_{gt} and end_{gt} represent the ground truth annotation's begin and end positions, and begin_{pred} and end_{pred} represent the prediction annotation's begin and end positions. Context window w is either 0 (exact match) or 200 characters (proximity match).

We report results for top-1 and top-5 features with the highest attribution scores by summing up the respective match counts. For the remainder of the paper, we refer to the proximity measures as prox top-1 and prox top-5.

A random permutation baseline shuffles the importance, is repeated five times, and the average value is rounded down to the nearest integer.

[4]https://captum.ai

	Precision	Recall	F1
micro	0.66	0.61	0.63
macro	0.37	0.34	0.35
accuracy	0.56		

Table 3: Performance metrics calculated using both final labels and path labels

	top-1	prox top-1	top-5	prox top-5
IG	0	1	0	8
kSHAP	0	0	5	8
random	0	2	2	10

Table 4: Overlap counts for the two explanation methods, Integrated Gradients and Kernel SHAP, and a random baseline

Results

Classification Results

Table 3 reports subset accuracy, micro and macro precision, recall and F1 score from the test set including final labels and path labels. The micro scores are higher than macro scores and subset accuracy is slightly below the micro scores.

Explainability Results

Overlap with ground truth In total, the number of model predictions that agree with the ground truth class is 73, which presents the upper bound of the overlap count. Table 4 shows the aggregated counts over all labels for the two explanation methods and the random baseline.

There is no overlap between the ground truth and the token with the highest importance score. Kernel SHAP shows a slightly higher top-5 count compared to the random baseline. However, for the soft top-5, the random baseline has a higher count. The overlap counts demonstrate that the explanation methods show no gains over a random baseline.

Qualitative Results As a sanity check we checked the top-5 tokens and report three examples in Table 5. The examples illustrate that explanations may contain special characters or numbers.

Runtime The experiments were performed on one NVIDIA Tesla P40 GPU. Runtime for the explanation generation was 10 hours for Kernel SHAP and 30 hours for Integrated Gradients. This led to an average computation time of roughly 5min per document for Kernel SHAP and 15min for Integrated Gradients.

Label	top-5 token
Not AI relevant	“)”, “5.5”, “Pro”, “Sch”, “Auto”
K1 (in doc 1)	“den”, “App”, “ter”, “.”, “nachhaltige”
K1 (in doc 2)	“ur”, “en”, “Bed”, “be”, “Flu”

Table 5: Explanation examples

Discussion

The results revealed that feature attribution methods do not overlap with ground truth more than a random baseline. High computational demands for the system, Longformer model combined with feature attribution methods, prevent interactive use. These findings raise questions about the applicability of feature attribution methods for similar settings.

As it is a complex real-world setting, various factors play a role that need to be unraveled. In the following, we discuss them across three dimensions: task, model and explanation method. We address the limitations of our work and highlight the risks and implications for each dimension.

Task

Assessing standard documents is generally a difficult endeavor, which is exactly the reason why we seek support from AI tools in the first place. Different types of knowledge are required to perform an assessment, including domain, technological, and regulatory knowledge. This became especially clear in the annotation process. In the early stages of the annotation, the same standard document was annotated by several annotators, and the classification was slightly different. The reason for these initial differences is that information may not be given as hard facts, but an interpretation of the text is necessary. The initial differences in annotations were resolved and aligned during the annotation process. Although the matter was addressed, it clearly showed that an assessment is not unambiguous.

Another aspect to take into account is the annotation style. It was not feasible to annotate all occurrences of relevant information for a label. The approach was to annotate in a "sufficient" manner. This affects the evaluation of the explanation method, which we discuss below.

A takeaway is that transforming the label flow chart into information usable for an ML model was not trivial. Several steps, such as the aggregation of labels into overarching topics, were performed to generate machine-readable data. We therefore recommend investing adequate time in translating human classification systems into a machine-readable classification system.

Model

Assessment of AI readiness is characterized by long documents, rendering it an extreme case of text classification. Due to limitations in the input length of the Longformer, roughly one-third of the documents had to be truncated. To ensure processing of the whole text, approaches with TF-IDF and also hierarchical BERT are possible. However, in prior work, these approaches yielded lower performance scores compared to Longformer variants (Mamakas et al. 2022; Tuteja and González Juclà 2023; Park, Vyas, and Shah 2022). At the same time, the resource requirements for these approaches are lower, resulting in reduced runtime which can be an option if the performance drop is tolerable. A disadvantage of document splitting is the need of a logic to reassemble the segments, which complicates explanation generation. Technically, it is also possible to use decoder architectures by reformulating the task to information retrieval or question answering, which has the advantage of processing even longer text. However, this opens up other challenges, such as stability and hallucinations (Bang et al. 2023). Nevertheless, this is a promising direction to take for solving the task in creative ways, e.g., by interacting with a document.

Important information for the AI readiness assessment is sometimes a single phrase hidden in a long document that changes the class. Even with recent approaches, such as Retrieval Augmented Generation (Lewis et al. 2020), current models may not be up to the challenge, considering for instance the "lost in the middle" phenomenon (Liu et al. 2024).

Explanation Method

The aim with explanation methods is to provide a means for understanding model decisions and to facilitate trustworthiness assessment. Unfortunately, our findings question the role of feature attribution methods for these purposes. The current feature attribution methods seem to fall short on important dimensions in our setting. They are computationally expensive and do not overlap with the ground truth.

Runtime The average time to generate explanations for a single document is at least 5 minutes. This is impractical for an application because the latency exceeds a 10-second attention threshold (Nielsen 1993). If an explanation needs to be generated more quickly, it is possible to reduce the hyperparameters, such as the step size in the case of Integrated Gradients. However, this reduction leads to a larger approximation error, making the explanation less accurate, and is therefore not recommended. The consequence is that document processing, including explanation generation, needs to happen offline, requiring intelligent solutions for offloading. It is important to factor in the effort from the ML operations side from the start. As a takeaway, we recommend to consider alternative modeling approaches if response time is of essence.

Ground Truth Overlap For the overlap, the annotation style and the notion of explainability are discussed. A limitation, as result of the annotation style, is that the annotations are only provided in a sufficient manner. By design, post-hoc explanation methods provide scores for each input feature. Focusing on top-k token, in fact, makes an explanation more compact but less faithful. One extreme would be to highlight everything that the method returns, which could be overwhelming for the user. The other extreme, to highlight just a single token, is incomplete.

Another limitation is that plausibility evaluation with ground truth is not always ideal. First and foremost, an explanation method will reveal important features from a model perspective. A model can exhibit high performance but may learn spurious correlations which do not align with human ground truth explanations. In addition, it could be that the explanation method reveals relevant and reasonable information which an annotator would also consider important, but it does not match the ground truth due to the sufficient annotation style. Despite this, the expectation was that, on average, an explanation method should have some overlap with the ground truth. However, the results did not substantiate that. One way forward would be to perform user

studies for further investigation. In the current form, we see a risk of overinterpretation, as the tokens so far do not suggest this to be a sensible approach. As far as we know, it has not been discussed whether feature attribution methods should only be applied to "better" performing models. Therefore, we see future work in exploring how model performance and explanation utility are connected.

We see it as more fruitful to work on the modeling before performing user evaluation of these explanations. Additionally, it may be promising to reevaluate whether interpretable models can be used and to invest more research into concept-level methods, as both areas are not sufficiently studied in the NLP community (Calderon and Reichart 2024).

Precisely because this a real-world task, the findings highlight that explainability is lacking the maturity for operationalization. Until more suitable post-hoc methods and evaluation procedures are developed we recommend to rely on rigorous testing as proposed in recent work (Bilodeau et al. 2024).

Conclusion

With the surge and application of language models in various domains, ethical and legal matters gain significance. Regulatory policies are being implemented to manage AI systems. As part of this, high expectations are placed on explainability to provide insight into the decision process of a classifier. However, there seems to be a vast gap between expectations and current research reality. This work exposed limitations in the runtime and plausibility of post-hoc explanation methods in a complex setting with long text classification. Our findings call into question how useful common feature attribution methods are for similar settings. More research is needed to identify suitable methods for trustworthiness assessment, either by better utilizing current methods or by creating new ones to close the present gap.

Acknowledgments

The development of this publication was supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) as part of the project "KI-Tauglichkeit von Normen" (AI Readiness of Standards). The authors would like to thank the project partners and standardisation experts for the collaboration.

References

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020. A Diagnostic Study of Explainability Techniques for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3256–3274. Online: ACL.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 675–718. Nusa Dua, Bali: ACL.

Bastings, J.; and Filippova, K. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 149–155. Online: ACL.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150.

Bhambhoria, R.; Dahan, S.; and Zhu, X. 2021. Investigating the State-of-the-Art Performance and Explainability of Legal Judgment Prediction. In *Canadian AI 2021*. Canadian Artificial Intelligence Association (CAIAC).

Bilodeau, B.; Jaques, N.; Koh, P. W.; and Kim, B. 2024. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2): e2304406120.

Burkart, N.; and Huber, M. F. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70: 245–317.

Calderon, N.; and Reichart, R. 2024. On Behalf of the Stakeholders: Trends in NLP Model Interpretability in the Era of LLMs. arXiv:2407.19200.

Danilevsky, M.; Dhanorkar, S.; Li, Y.; Popa, L.; Qian, K.; and Xu, A. 2021. Explainability for natural language processing. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 4033–4034.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: ACL.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458. Online: ACL.

DIN EN 12791:2018-01. 2018. Chemische Desinfektionsmittel und Antiseptika - Chirurgische Händedesinfektionsmittel - Prüfverfahren und Anforderungen. Standard.

Doshi-Velez, F.; and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608.

Duden. n.d. The longest words in the Duden dictionary. https://www.duden.de/sprachwissen/sprachratgeber/Die-langsten-Worter-im-Duden. Accessed: 2024-07-25.

Gurrapu, S.; Kulkarni, A.; Huang, L.; Lourentzou, I.; and Batarseh, F. A. 2023. Rationalization for explainable NLP: a survey. *Frontiers in Artificial Intelligence*, 6: 1225093.

Görge, R.; Haedecke, E.; Schmitz, A.; Borowski, M.; Seeliger, A.; and Poretschkin, M. 2024. Digital Governance: Confronting the Challenges Posed by Artificial Intelligence. Forthcoming.

Herrewijnen, E.; Nguyen, D.; Bex, F.; and van Deemter, K. 2024. Human-annotated rationales and explainable text classification: a survey. *Frontiers in Artificial Intelligence*, 7: 1260952.

Hou, W.-H.; Wang, X.-K.; Wang, Y.-N.; Wang, J.-Q.; and Xiao, F. 2024. Modelling long medical documents and code associations for explainable automatic ICD coding. *Expert Systems with Applications*, 249: 123519.

Jesus, S.; Belém, C.; Balayan, V.; Bento, J.; Saleiro, P.; Bizarro, P.; and Gama, J. 2021. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, 805–815. ACM.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Lu, X.; Li, J.; Wan, Z.; Lin, X.; Takeuchi, K.; and Kashima, H. 2024. Evaluating Saliency Explanations in NLP by Crowdsourcing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 6431–6443. Torino, Italia: ELRA & ICCL.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, 4765–4774.

Mamakas, D.; Tsotsi, P.; Androutsopoulos, I.; and Chalkidis, I. 2022. Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, 130–142. Abu Dhabi, United Arab Emirates (Hybrid): ACL.

Mendez Guzman, E. A.; Schlegel, V.; and Batista-Navarro, R. 2024. From Outputs to Insights: A Survey of Rationalisation Approaches for Explainable Text Classification. *Frontiers in Artificial Intelligence*, 7: 1363531.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; and Seifert, C. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.*, 55(13s).

Nielsen, J. 1993. Response Times: The 3 Important Limits. https://www.nngroup.com/articles/response-times-3-important-limits/. Accessed: 2024-07-25.

Pappagari, R.; Zelasko, P.; Villalba, J.; Carmiel, Y.; and Dehak, N. 2019. Hierarchical Transformers for Long Document Classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 838–844. IEEE.

Park, H.; Vyas, Y.; and Shah, K. 2022. Efficient Classification of Long Documents Using Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 702–709. Dublin, Ireland: ACL.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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*, 1135–1144.

Stremmel, J.; Hill, B. L.; Hertzberg, J.; Murillo, J.; Allotey, L.; and Halperin, E. 2022. Extend and Explain: Interpreting Very Long Language Models. In *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, 218–258. PMLR.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Tuteja, M.; and González Juclà, D. 2023. Long Text Classification using Transformers with Paragraph Selection Strategies. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, 17–24. Singapore: ACL.

Valvoda, J.; and Cotterell, R. 2024. Towards Explainability in Legal Outcome Prediction Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7269–7289. Mexico City, Mexico: ACL.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30: 5998–6008.

Wiegreffe, S.; and Marasović, A. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.