From Occlusion to Transparency

An Occlusion-Based Explainability Approach for Legal Judgment Prediction in Switzerland

Bachelor Thesis

Nina Baumgartner

Faculty of Science, University of Bern

31.12.2022

Dr. Matthias Stürmer
Joel Niklaus
Research Center for Digital Sustainability
Institute of Computer Science
University of Bern, Switzerland

Declaration of Authorship

I, NINA BAUMGARTNER, declare that this thesis titled, "From Occlusion to Transparency" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date:

31.12.2022

Signed:

Acknowledgements

I want to express my sincere gratitude to Joel Niklaus for his supervision and guidance throughout the process of writing my thesis. I am also grateful to the legal experts Thomas Lüthi, Lynn Grau, and Angela Stefanelli for their annotations and helpful legal advice. I would like to extend my gratitude to Dr. Matthias Stürmer for giving me the opportunity to write a thesis in his research group and for broadening my horizon in the field of Explainability in Legal Judgment Prediction. I wish to express my appreciation and thankfulness to Ilias Chalkidis for the helpful advice he provided on conducting my experiments. I would further like to thank Alex Nyffenegger for his assistance with the first Prodigy setup and deployment and the NLP seminar group for their feedback on my work throughout the process. Furthermore, I thank my friends and family for their continuous support. I am also grateful to the Digital Humanities Bern team for their helpful advice and ideas for this thesis. Finally, I would like to express my deepest appreciation to Lukas Schacher for proofreading this thesis, for his helpful advice, and continuous support, and for enduring this process with me.

Abstract

Natural Language Processing (NLP) models have been used for more and more complex tasks such as Legal Judgment Prediction (LJP). A LJP model predicts the outcome of a legal case by utilizing its facts. This increasing deployment of Artificial Intelligence (AI) in high-stakes domains such as law and the involvement of sensitive data has increased the need for understanding such systems. We propose a multilingual occlusion-based explainability approach for LJP in Switzerland and conduct a study on the bias using Lower Court Insertion (LCI). We evaluate our results using different explainability metrics introduced in this thesis and by comparing them to high-quality Legal Expert Annotations using Inter Annotator Agreement. Our findings show that the model has a varying understanding of the semantic meaning and context of the facts section, and struggles to distinguish between legally relevant and irrelevant sentences. We also found that the insertion of a different lower court can have an effect on the prediction, but observed no distinct effects based on legal areas, cantons, or regions. However, we did identify a language disparity with Italian performing worse than the other languages due to representation inequality in the training data, which could lead to potential biases in the prediction in multilingual regions of Switzerland. Our results highlight the challenges and limitations of using NLP in the judicial field and the importance of addressing concerns about fairness, transparency, and potential bias in the development and use of NLP systems. The use of explainable artificial intelligence (XAI) techniques, such as occlusion and LCI, can help provide insight into the decision-making processes of NLP systems and identify areas for improvement. Finally, we identify areas for future research and development in this field in order to address the remaining limitations and challenges.

Contents

1 Introduction

Natural language processing (NLP) is a field at the intersection of Artificial Intelligence (AI) and linguistics, focused on the interaction between computers and human language (natural language). The ultimate goal of NLP is to create systems that understand natural language as humans do, taking into account all the contextual nuances. To achieve this goal, NLP systems can use a variety of techniques, including white box methods like decision trees and black box approaches like neural models. NLP has a wide range of applications across various fields, including science, criminal justice, humanities, and economics. It can be used to accurately extract, categorize, organize, and predict information contained in texts, making it particularly useful for legal documents such as court rulings and laws. Legal text processing, which includes the application of NLP to legal texts, is therefore a growing area of study.

One specific application of NLP in legal text processing is Legal Judgment Prediction (LJP), which uses the fact section of a court case to predict the final judgment. While this has the opens the possibility of speeding up the judicial process and potentially providing a more impartial judgment by removing the human element, it is important to note that data on crime, including court decisions, is known to be inherently biased and not reflective of reality due to the filtering that occurs in the criminal justice system (Neubacher, 2017). Therefore, it is crucial that any automated decision system in the criminal justice domain is able to explain its predictions in terms of how humans understand the legal process. This is where the field of Explainable Artificial Intelligence (XAI) comes in. XAI refers to the ability to understand and describe the inner workings of an AI system and how it arrived at a particular decision or prediction. There are various methods for explainability, including occlusion (Zeiler and Fergus 2013 & Li et al. 2017), LIME (Ribeiro et al., 2016) or integrated gradients (Sundararajan et al., 2017), as well as a focus on providing human-understandable explanations of decision-making processes by annotations.

In this thesis, we propose a multilingual occlusion-based explainability approach as an extension of the Swiss-Judgment-Prediction (SJP) (Niklaus et al., 2021) and explore the model's result using different metrics including Inter Annotator Agreement (IAA). Our aim is to address the concerns about fairness, transparency, and potential bias in training data in the high-stakes domain of criminal justice. Our contributions are as follows:

- We provide a multilingual set of gold-standard legal expert annotations which serve as the foundation for the occlusion and Lower Court Insertion (LCI) dataset and provide reliable ground truth for evaluating the gathered explanations.
- We publish comprehensive and well-structured Annotation Guidelines to ensure the quality and reproducibility of the legal expert annotation task.
- We present four different occlusion test sets and one LCI test set in German, French and Italian, which can be run in combination with the training and validation set provided by Niklaus et al. (2021).
- We perform a thorough analysis of the occlusion results using various explainability metrics including calculating the IAA between the model and human-produced explanation.
- We conduct a study on bias in SJP using the variation of the occlusion method LCI.
- We make our code publish to motivate further research on the topic of explainable LJP.

The rest of this thesis is organized as follows. Section 2 presents existing related work in the field of LJP by introducing the most important benchmark dataset for this thesis and giving an overview of XAI methods. Section 3 provides the theoretical background to the wide range of methods and metrics adopted in this thesis. In Section 4, we introduce the four different datasets being used for this work. Section 5 presents the steps for the implementation of the conducted experiments. Section 6 introduces the study results and in Section 7 we discuss our findings. Finally, in Section 8 we provide the conclusion of this thesis and discuss limitations and possible future work in the field of explainable LJP.

2 Related Work

In this section, we will review previous research on explainability in LJP, focusing on the most important benchmark dataset and the relevant literature on XAI.

2.1 Benchmark Datasets

Niklaus et al. (2021) released a multilingual corpus containing 85K cases from the Federal Supreme Court of Switzerland that covers the period from 2000 to 2020. In their work, they introduced the SJP task, which involves predicting the verdict from the fact section of a court ruling in a binarized LJP task. They released a multilingual corpus containing 85K cases from the Federal Supreme Court of Switerland covering the period from 2000 to 2020 and evaluated state-of-the-art BERT-based methods, including two variants that can handle longer input texts. They found that hierarchical BERT, similar to the one presented in Chalkidis et al. (2019) had the best performance with approximately 68-70% Macro-F1-Score in German and French. They also studied the impact of factors such as the canton of origin, year of publication, text length, and legal area on model performance.

Chalkidis et al. (2019) published an English LJP dataset, containing cases from the European Court of Human Rights. They presented three tasks on which they evaluated several neural models, establishing strong baselines that surpass previous feature-based models. In addition, Chalkidis et al. (2022) also explored fairness in legal text processing. Their multilingual benchmark suite, called FairLex, evaluates the fairness of pre-trained language models and the methods used to fine-tune them for downstream tasks in four jurisdictions (European Council, USA, Switzerland, and China), across five languages (English, German, French, Italian and Chinese), and for five attributes (gender, age, region, language, and legal area). In their experiments, they found that performance group disparities exist in many cases. None of the evaluated pre-trained language models with various group robust fine-tuning algorithms consistently mitigate group disparities or guarantee fairness. They also provided a quantitative and qualitative analysis of their results, highlighting open challenges in the development of robustness methods in legal NLP.

Malik et al. (2021) introduced the Indian Legal Documents Corpus. This corpus contains 35k Indian Supreme Court cases in English annotated with the court's original decision. In addition, they also created a separate test set annotated with gold-standard explanations by legal experts. The authors propose the Court Judgment Prediction and Explanation (CJPE) task with a hierarchical occlusion-based model for explainability. They chose occlusion (Zeiler and Fergus 2013 & Li et al. 2017) after experimenting with other explainability methods which were not suitable for their CJPE task. The proposed algorithm's analysis of explanations showed a significant difference between the algorithm's and legal experts' perspectives on explaining judgments, indicating the potential for further research in this area.

2.2 Explainable AI

Kakogeorgiou and Karantzalos (2021) evaluated ten XAI methods in multi-label classification tasks to improve transparency and produce human-interpretable explanations. They used and trained deep learning models with state-of-the-art performance, applied the XAI methods to understand and explain the models' predictions, and assessed and compared the performance of these methods using quantitative metrics. After conducting numerous experiments to evaluate the overall performance of the XAI methods, they found that occlusion (Zeiler and Fergus 2013 & Li et al. 2017) and LIME (Ribeiro et al., 2016) were the most interpretable and reliable, but at the cost of being computationally expensive.

In their survey, Danilevsky et al. (2020) provided a categorization of explainability methods and detailed the operations and techniques currently available for generating explanations for NLP model predictions. They discussed the distinction between inherently interpretable models and black box techniques, which require additional post-processing (post-hoc approaches) to be understandable. They also argued that a model's quality of explanation should be evaluated not only by its accuracy and performance but also by how well it provides explanations for its predictions. However, the authors emphasize that there is little agreement on how explanations should be evaluated, and while the majority of the reviewed works lacked standardized evaluation and only provided informal evaluation, a smaller number of papers examined more formal evaluation approaches, including leveraging ground truth data and human evaluation. The authors also highlighted the importance of providing high-quality ground truth when evaluating using human annotations.

In their work, Wiegreffe and Pinter (2019) focused on datasets containing human-annotated textual explanations and identified 65 datasets in this category. They classified these explanations into three main categories: highlights, free-text, and structured. They discussed how these annotations are used in different ways, including the creation of ground truth to evaluate model-generated explanations. They also summarized the

literature on collecting textual explanations, highlighted discrepancies in data collection that can have downstream effects on modeling, and provided recommendations for future dataset construction.

Bhambhoria et al. (2021) studied explainability in the context of LJP by using the most recent deep learning model, longformer, and achieving state-of-the-art performance with a limited amount of training data. However, their analysis also suggested that the improvement may have been due to the model's fitting to spurious correlations, in which the model made correct decisions based on information unrelated to the task. As a result, they cautioned that care should be taken when interpreting the obtained results. They also provided interpretability by applying post-hoc explanations to their task.

3 Methods

In this thesis, we aim to produce high-quality explanations for the decisions of the SJP task presented by Niklaus et al. (2021) and to evaluate their plausibility using legal expert annotations as a ground truth. To achieve this goal, we carefully selected the occlusion method as our explainability approach, as it has been shown to produce output similar to human annotations, as proposed by Malik et al. (2021) and inspired by the work of Zeiler and Fergus 2013 and Li et al. (2017). In this chapter, we justify our choice of the occlusion method and present its application in the context of our experiments. We also introduce the legal expert annotations, which serve as a basis for evaluating the plausibility of the model-generated explanations, and provide an overview of the IAA. Finally, we present the metrics we use to evaluate the results of the occlusion experiments, including the scaled confidence, explainability score, normalized explainability score, and explanation accuracy score. Note that since three out of four datasets for this thesis are derived from the legal expert annotations and are tailored specifically to the chosen explainability method we introduce all these datasets together in Section 4.

3.1 Explainability Methods

As discussed in the introduction of this thesis, explainability is becoming increasingly important as AI is being used in a wider range of applications, including high-stakes decision-making such as criminal justice. NLP can utilize both a white-box and a black-box approach when solving a problem. White-box techniques are inherently understandable and transparent to humans, they include “self-explaining" models such as decision trees and rule-based approaches. On the contrary, black-box techniques are only partially interpretable (if at all) and require additional steps after the prediction to be explainable (post-hoc explainability methods). Nevertheless, the popularity of various black-box techniques such as deep learning models has only increased in recent years (Danilevsky et al., 2020). The reason is that they often provide substantially advanced model quality. Unfortunately, this increase in performance comes at the expense of a model's explainability.

To categorize XAI methods Danilevsky et al. (2020) distinguish between **local** and **global explanations**. Local explanations refer to explanations that are specific to a particular input or subset of inputs, while global explanations refer to explanations that apply to the entire model or a large portion of the input space. **Local explanations** are typically used to provide insight into how a machine learning model is making decisions for a specific input or group of inputs. These kinds of explanations can be useful for understanding how the model is making decisions in specific cases, but they may not provide a complete understanding of the model's behavior. **Global explanations**, on the other hand, provide a more comprehensive view of how a machine-learning model is making decisions. These explanations may take into account the entire range of possible inputs and provide information about the model's overall behavior and decision-making process. Global explanations can be useful for understanding the strengths and limitations of a model, and for identifying any issues that may impact its performance. This thesis focuses on local explainability methods in order to examine the reasons behind the model's prediction of a specific judgment based on a specific fact section.

In addition to local and global explainability methods, we can further classify XAI methods as **model-agnostic** and **model-specific methods** (see Table 1) (Malik et al. (2021) & Kakogeorgiou and Karantzalos (2021)). **Model-agnostic** methods work independently of the model's characteristics, these methods include occlusion (Zeiler and Fergus 2013 & Li et al. 2017), LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017) and Anchors (Ribeiro et al., 2018). **Model-specific** explainability methods are techniques that are designed to provide explanations for the decisions made by specific types of machine learning models. These methods include integrated gradients (Sundararajan et al., 2017), gradient saliency, and attention-based method. There is also the category of attribution or feature-based methods. These methods involve ranking the input features that have the most significant impact on the model's output, included in this group are Layerwise Relevance Propagation (LRP) (Bach et al., 2015) and DeepLIFT (Shrikumar et al., 2019).

Local Explainable AI Methods		
Model-Agnostic	**Model-Specific**	**Attribute-Based**
Occlusion	Integrated Gradients	LRP
LIME	Gradient Saliency	DeepLIFT
SHAP	Attention-Based Methods	
Anchors		

Table 1: Overview of the different XAI methods

As one can see from the section above there is no one-size-fits-all method for explaining AI systems, the choice of method depends on the specific characteristics of the model and the needs of the user. Malik et al. (2021) experimented with different explainability methods for their CJPE task before deciding on occlusion. Starting with **gradient saliency** and **integrated gradients**: **Gradient saliency** is a method that is used to explain the decisions made by a single input (e.g. an image), while **integrated gradients** is a method that is used to provide more robust explanations that take into account the entire range of possible inputs. Both methods can be used to explain the decisions made by machine learning models by identifying the input features that had the most significant influence on the model's prediction. However, most of the time they are applied to computer vision tasks (e.g. saliency maps for image classification), if applied in NLP they only produce a few tokens as an explanation. This is a problem these methods have in common with LRP, DeepLIFT, LIME, SHAP, Anchors, and attention-based methods. All these approaches produce only a few tokens or short phrases as an explanation, even when considering large sample sizes. This makes them only partially usable for an explainability task with human annotations. For this thesis especially, we needed a method with an adaptable explanation size, since the sentence length the explanation would be compared to also varies greatly. We hope to achieve a better quality of explanation by producing longer strings since this provides more context and thus makes the explanation more human-understandable.

Beyond that Danilevsky et al. (2020) raises fidelity concerns when using surrogate models such as **LIME**. LIME works by perturbing the input data and observing the effect on the model's predictions. Then it uses this information to create a simplified linear model that approximates the behavior of the original model in the local region around the input data. In consequence, it may be the case that the learned surrogate models and the original models have completely different mechanisms to make predictions, reducing the fidelity of the produced explanation. The use of attention weights as an explanation in **attention-based** methods is still debated as a potential drawback. Jain and Wallace (2019) argue that it is unclear if there exists a relationship between attention weights and model outputs. In their work, they performed extensive experiments across a variety of NLP tasks to assess the degree to which attention weights provide meaningful "explanations" for predictions. They found that they largely do not. On the contrary, Wiegreffe and Pinter (2019) argue in their paper that attention weights can very much be used as an explanation when adapting one's definition of explanation. They propose four alternative tests to identify when and whether attention can be used as an explanation, which each allows for a meaningful interpretation of attention mechanisms. Even still, they confirm Jain and Wallace's opinion that attention distributions are not ideal when searching for the one true, faithful explanation of the link a model creates between a certain input and output. This leaves us with occlusion.

Occlusion was first suggested by Zeiler and Fergus (2013) for computer vision tasks. Later Li et al. (2016) introduced occlusion as a broad methodology to investigate and explain prediction from a neural model by examining the effect erasing various parts of the representation, such as input words, has on a model's decision. According to Kakogeorgiou and Karantzalos (2021) a big upside of this method, besides that it is an agnostic model and has variable input size, is that it is very easy to implement especially when applying the occlusion only to the input of the model. This makes occlusion analysis one of the most interpretable and reliable XAI methods, since it does not change the initial model setup and can therefore give a reality-based explanation. Still, occlusion is not perfect, as it can be computationally expensive due to the involved perturbation. For this thesis, the issue of scale is not a concern as we apply occlusion to a small subset of the SCRC, with an average experiment size of approximately 2'100 predictions.

3.2 Occlusion Method with hierarchical BERT

Concerning our implementation of occlusion, we produce explanations by removing elements from the input of the SJP task (fact section) and analyze the prediction confidence in comparison to a non-occluded baseline. In our setup, we do not change the conditions for the model's prediction, keeping the circumstances identical to the SJP task. We use the same training and validation set as Niklaus et al. (2021) and only change the test set.

For the occlusion test sets, we choose a subset from the SJP original test set. We then apply a sentence permutation, occluding each sentence of the fact section once. We also do further experiments with larger sentence groups of up to 4 sentences (for more details see Section 5.5). After applying the occlusion to the chosen cases we let the model make a prediction for each occlusion experiment, resulting in a confidence score and a binary prediction of 0 and 1. For the prediction, we use the hierarchical version of BERT since Niklaus et al. (2021) achieved the best results with it. Hierarchical BERT uses a shared BERT encoder to independently encode segments of up to 512 tokens, and then combines the encodings of all the segments using a BiLSTM encoder. The final output states of the LSTM are concatenated to create a single representation of the document, which is then used for classification. This method is similar to the one described in Chalkidis et al. (2019). In addition to keeping the setup identical to the SJP task our explanations have the advantage that they are easily

comparable to human annotations since we can choose the explanation size in the implementation. In fact, our implementation actually use the legal expert annotations itself to produce the occlusion experiment (see Section 5.5). Which makes the creation of the occlusion test set and the plausibility analysis much easier than using another method.

3.3 Lower Court Insertion – A study on Bias

As an addition to the “normal" occlusion we described above, we also experiment with a setup we call LCI (see Section 5.6) where we extract the lower court instances and insert each lower court in each case. This task was proposed by Ilias Chalkidis (Chalkidis et al., 2019) and Thomas Lüthi (legal expert) as a study on the bias from one lower court to another. These experiments keep the same setup described above, again only adding a new test set for the model's prediction. These experiments also result in a confidence score and a binary prediction.

3.4 Legal Expert Annotations

The Legal Expert annotations serve as the ground truth for evaluating the accuracy of the model-generated explanations. These explainability annotations are conducted in collaboration with three legal experts. According to Wiegreffe and Marasovic (2021) annotations are used downstream in three ways:

> [As] data augmentation to improve performance on a predictive task, as supervision to train models to produce explanations for their predictions, and as a ground truth to evaluate model-generated explanations.

The legal expert annotations conducted in this thesis focus on the last application: the evaluation of the plausibility of the model-generated explanations. The annotations are conducted over five months starting with the pilot annotations and ending with the gold-standard annotations. The team of legal experts consists of two law students, both studying for their master's degree and one lawyer. Each annotator was given access to their annotation interface on the annotation tool Prodigy (see Section 5.1), where the fact section of the ruling was displayed for labeling. In the end, the annotation only consists of the highlighted fact section with an occasional free-text explanation (see Section 3.4.1.2). The annotations are conducted in the three languages contained in the Swiss judicial system: German, French and Italian. All three legal experts speak German as their first language but have learned French at school. One legal expert is also fluent in Italian. Each fact section was at least annotated by one of the legal experts. The documents in German are labeled by all three experts.

The finished annotation only consists of the highlighted fact section. The question now pending is: why not annotate the entire court ruling for explainability? As mentioned in the previous section, the explainability annotations have the goal to provide ground truth for the occlusion method. Both the occlusion and the annotation try to explain the prediction made by the LJP Task presented by (Niklaus et al., 2021), which is a prediction based solely on the fact section. There is no need to annotate any other parts of the court ruling that will not be provided to the model during the occlusion process. The only difference between the human annotators and the model is, that even though they only highlighted sentences in the fact section, they had access to the entire ruling to make their annotation. The reasoning behind this is, that the annotation task focuses only on explainability and not whether or not humans could beat the model to predict the judgment.

3.4.1 Annotation Guidelines

Annotation guidelines are the instructions given to the annotator and are at the heart of a successful annotation task. Clear guidelines ensure that the task is understandable and reproducible for the annotators. Thus, the development of consistent and understandable annotation guidelines is an integral part of this thesis. The guidelines for these explainability annotations are influenced by the works of Reiter (2020), Leitner (2019) and Pustejovsky and Stubbs (2013). Note that the latest version of the complete guidelines can be found in the appendix of this thesis, they will be quoted with Baumgartner (2022).

3.4.1.1 Annotation Goal

Annotation guidelines should describe the annotation task as generically as possible, but simultaneously as precisely as necessary so that human annotators can annotate reliably. Pustejovsky and Stubbs (2013) describe that the annotation goal is one of the first pieces of information one should give to an annotator. In the introduction to the guidelines for this thesis the goal for the annotation is described as follows:

> To investigate explainability in the legal area of AI [this] annotation task has the goal to gather the human part of the explanation. With your annotation, you will give your insight as a legal expert and tag parts of the facts with specific labels. These guidelines should help you to identify the important parts of the facts and create consistent annotations.

In other words, the annotation goal is to explain court decisions from the viewpoint of a legal expert, by annotating the fact section of a ruling with explainability labels.

3.4.1.2 Annotation Entities and Categories

We have now defined the annotation goal and explained how using the fact section helps us reach this goal. As mentioned in the previous sections, we want to gather human explanations in a workable and interpretable structure, namely via text highlights in the fact section. Thus, it is of great importance for the guidelines to explain and narrow down exactly what entity is to be annotated. Reiter (2020) describes that when writing annotation guidelines we should describe exactly what should be annotated. Be it every paragraph, every sentence, every word, or only units that fulfill a certain condition. It is essential to choose units or entities as objectively and independently of interpretation as possible so that the real decision lies in the categorization of the units.

After consulting with legal experts, we decided to focus on sentences and sub-sentences in order to preserve as much context as possible in the highlighted units. This decision was made in the hope that it would enable us to better explain the meaning of different parts of sentences in the context of the judgment, and thus improve our understanding of the model's decisions. The guidelines describe these entities as follows:

> To add highlights you will label sentences or sub-sentences as supporting or opposing the judgment. For this task, we define a sentence as a self-contained linguistic unit consisting of multiple words, terminated with a period, semicolon, colon, question mark, or exclamation mark. An entire sentence is the largest entity to be annotated. A sentence can consist of multiple sub-sentences usually separated with a "and" or a comma. A sentence may contain two sub-sentences opposing each other, which should be consequently annotated with different labels. These sub-sentences are the smallest units that should be annotated. So single words or expressions should never be annotated. (Baumgartner, 2022)

In addition to sentences and sub-sentences, we also ask annotators to label the last lower court, which is indicated in the Rubrum of the ruling. We encourage annotators to consult the Rubrum to identify the correct lower court, as they always have access to the full ruling.

> The last lower court is composed of the name of the court e.g. "Verwaltungsgericht" and the location "Kanton Luzern". Please annotate all instances of the lower court where it appears as a complete constellation. So for example, if "Verwaltungsgericht des Kanton Luzern" appears multiple times in the facts please label it each time. Please Note that you should only annotate the lower court itself please do not label prepositions like "beim" or "zum" or verbs like "sprach" which are often found next to the lower court. (Baumgartner, 2022)

After defining the units to be annotated, we provide a description of the annotation categories to enable the legal experts to identify the appropriate sentences for each explainability label. To this end, we provide the annotators with the following annotation categories: `Supports Judgment`, `Opposes Judgment`, `Neutral`, and `Lower Court` The labels of `Supports Judgment` or `Opposes Judgment` are the most important explainability labels and require the most expert knowledge and time to identify. The annotators are asked to highlight each sentence or sub-sentence which supports or opposes the judgment with the matching label. For this purpose, they first have to read through the facts, the consideration, the ruling, and any other needed legal document to understand the court case and then annotate the correct sentences. The difficulty is that similar sentences may have different contexts if the verdict is approved or dismissed. This is one example of conflicts that needed resolving when creating the gold-standard set. Note that `Supports Judgment` and `Opposes Judgment` are not only opposite but also distinct categories. Therefore, instances of those two categories should not overlap.

These explainability labels were chosen because we consider them best suited for the comparison with the model's explanation. Since they are opposing each other they can be assigned a score (as explained in Section 3.7). Our version of occlusion removes them systematically from the text. Through this approach, we can evaluate whether the model's performance would improve if, for example, we removed a sentence opposing the judgment, as this would make the fact section more favorable to the judgment by its absence. With the `LowerCourt` label as explained in Section 3.3, we try to investigate a hypothesis of bias from one lower court to annotator as suggested by Ilias Chalkidis and Thomas Lüthi.

For `Neutral` parts of the facts, the annotators are instructed in the following way:

> Every not-labeled sub-sentence is considered neutral. This is not a label per se but merely how the system interprets words or sentences which are not assigned one of the labels above. It is important for the analysis that even the neutral sentences are annotated which in our case means to omit them. One example in German of a neutral expression that should not be tagged with a label is the word “Sachverhalt:". This word only indicates the beginning of the fact section and should be left out as a neutral part of the facts because it does not give us any further information on the explainability of the judgment. Another example of a neutral part of the facts are the section indicators labeled with capital letters (e.g. A., B., A.a., A.b and so on). Note that witnesses, accused persons, and other involved parties are also labeled with uppercase letters and should be annotated if part of a sentence [...]. (Baumgartner, 2022)

For the creation of the gold-standard annotation, the instruction for `Neutral` sentences is changed. The annotators are given a label on the interface to indicate them. This change is necessary for the implementation of the occlusion method because it split the fact section into more or less coherent sentences with minimal effort.

During the entire annotation process, the annotators are also given several options for dealing with problematic cases. We expect that this will further improve the dataset, as it did, for example, in the German subset, where additional cases were added dynamically during the annotation process. The annotators are also encouraged to provide free text explanations for cases that they found particularly challenging, which will provide valuable legal insight for the interpretation of the results.

3.5 Creation of the gold-standard Annotation Dataset

According to both Pustejovsky and Stubbs (2013) and Reiter (2020), the creation of annotation guidelines and the annotation process itself are iterative processes. Working in these cycles should ensure the quality of the guidelines, the annotation itself, and a consistent gold-standard annotation set.

> Using the annotation guidelines to identify the right parts of the text, multiple annotations by multiple individual annotators are done on the same input. Then these annotations are analyzed and the guidelines are adapted accordingly to provide consistency in the annotations. Therefore, it is important that for the first few cycles the annotations are done individually. Later the gold-standard annotations for this corpus emerge from this process. (Baumgartner, 2022)

In consequence, the guidelines stayed a work-in-progress during the entire time period the annotations were conducted and were only finalized after finishing the gold-standard annotation. Changes to the guidelines are continuously documented in a change log at the end of the document.

After finalizing the pilot annotations we started with a second cycle. Unfortunately, by then one of the annotators had to leave the team which left only one student annotator and the lawyer. Pustejovsky and Stubbs (2013) suggest calculating the IAA after each cycle. For this thesis, we chose a different approach. Because time was of the essence at this stage only a qualitative analysis was conducted. This first analysis revealed some flaws in the annotations and enabled us to improve the guidelines with more precise instructions. After the second cycle, we instructed the annotators on the gold-standard annotation process. Pustejovsky and Stubbs (2013) define this gold-standard as the final version of the annotations that uses the most up-to-date guidelines and contains only correctly annotated sentences. Since there is a lot of subjectivity involved in this annotation task, explanations may just differ from person to person regardless of how precisely one writes the guidelines. In consequence, conducting more cycles may not even have resulted in a greater agreement. This justifies our choice of stopping the annotations after 2 cycles and transferring over to the gold-standard annotations. To ensure the high quality of these last annotations we had an open discussion with the legal experts. We decided how to exactly resolve conflicts and how to merge the annotations in the best possible way. After this discussion one legal expert used the gold-standard annotation interface to compare all the annotations done for each case until this point and then chose the most fitting annotation, resolving all the pending conflicts. In addition, this legal expert also annotated all the neutral sentences.

3.6 Inter Annotator Agreement

Pustejovsky and Stubbs (2013), Malik et al. (2021) and Wiegreffe and Marasovic (2021) all suggest the IAA as a baseline for agreements among the annotators. In general, the IAA refers to the degree to which multiple annotators agree on the labels or categories assigned to a set of data. Measuring IAA is important because it can provide insight into the reliability and consistency of the labeling process, and can help identify areas where additional training or clarification may be needed. This score is always between 0 and 1, with 0 being no

agreement and 1 being perfect agreement. As mentioned before the IAA should be fairly high (> 0.4) to ensure that the annotation guidelines are comprehensive, making the annotation task reproducible.

Pustejovsky and Stubbs (2013) describes Cohen's Kappa (and its variation, Fleiss's Kappa) and Krippendorff's Alpha as the most commonly used metrics in computational and corpus linguistics. Unfortunately, these scores are not well suited to calculate IAA for the conducted annotations. Firstly, Cohen's Kappa only looks at two annotators annotating each instance with a category, the annotations for this thesis is in part conducted by three different annotators. Cohen's Kappa's extension Fleiss's Kappa and Krippendorff's Alpha would solve this problem, but both of these scores focus on the agreement between the number classifications and not on sentence vise agreement. In other words, when applying these scores one would see a fact section as a group of classified sentences with the classes: `LowerCourt`, `Supports Judgment`, `Opposes Judgment` and `Neutral`. Then, for each annotator, we would count the number of sentences in a class (e.g. Annotator 1 has five sentences with the label `Opposes Judgment` while annotator 2 has only three) and calculate the agreement from there. In consequence, when annotators would label different sentences in a fact section with the same label, but annotate the same amount of sentences, we would have an agreement of 1.

For this reason, we use the scores Malik et al. (2021) suggested in their work. They took inspiration from the machine translation community and used `ROUGE-L`, `ROUGE-1`, `ROUGE-2` (Lin, 2004), `BLEU` (Papineni et al., 2001), `METEOR` (Lavie and Agarwal, 2007), `Jaccard Similarity`, `OVERLAP Maximum`, and `OVERLAP Minimum` to measure IAA. In addition to these scores, we also calculate IAA with the `BERTScore` (Zhang et al., 2020). In the following, each of these scores is briefly introduced.

3.6.1 ROUGE

`ROUGE` was first introduced by Lin (2004) and stands for Recall-Oriented Understudy for Gisting Evaluation. Originally ROUGE was developed to automatically evaluate the quality of a model-produced summary by comparing it to other (ideal) summaries made by humans, however, it can also be used to calculate IAA between humans and humans and machines and humans. The ROUGE scores are calculated by counting the number of overlapping entities such as n-grams, word sequences, and word pairs between the two samples. There are four different ROUGE measures: `ROUGE-N`, `ROUGE-L`, `ROUGE-W`, and `ROUGE-S`. For this thesis, we will use `ROUGE-N` (`ROUGE-1`, `ROUGE-2`) and `ROUGE-L`.

`ROUGE-N` is an n-gram (n stands for the length of the n-gram) recall between a sample text and a set of reference texts. An n-gram is a contiguous sequence of n words in a given text. For example, in the annotation "Verwaltungsgericht des Kantons Zürich" the unigrams (n=1) are "Verwaltungsgericht", "des", "Kantons" and "Zürich". The bigrams (n=2) are "Verwaltungsgericht des", "des Kantons", and "Kantons Zürich". The `BLEU` score presented in Section 3.6.2 below is a closely related measure. `ROUGE-L` is a longest common subsequence (LCS) based F-measure. It measures agreement by looking at the LCS of two sample texts. Intuitively, the longer the LCS of two samples the more similar they are. With this property `ROUGE-L` is closely related to the `OVERLAP Maximum` and `OVERLAP Minimum` scores, though more sophisticated. `ROUGE-L` measures in-sequence co-occurrences. This score is hence able to capture sentence-level structure in a natural way. Lin (2004) thoroughly evaluated the different ROUGE measures concluding that `ROUGE-2`, `ROUGE-L` worked well in single document summarization tasks, while `ROUGE-1`, `ROUGE-L` performed great in evaluating very short summaries (or headline-like summaries).

3.6.2 BLEU

The `BLEU` (Bilingual Evaluation Understudy) score (Papineni et al., 2001) is a metric used to evaluate the quality of text generated by machine translation systems. It compares the generated text to a reference translation and calculates a score based on the number of matching n-grams (sequences of n words) between the two. If we look at a generated text as annotation A and the reference translation as annotation B we can adapt the `BLEU` score to an IAA metric. The score is calculated using the brevity penalty (BP), which penalizes the generated text if it is shorter than the reference, and the sum of the precision values of the generated text for each n-gram. The precision is defined as the number of matching n-grams in the generated text divided by the total number of n-grams in the generated text. Papineni et al. (2001) concluded in their evaluation of `BLEU`, that it provides a simple and effective way to compare the quality of different translation systems. The authors emphasize that resulting `BLEU` scores highly correlate with human judgment on the same translation.

3.6.3 METEOR

`METEOR` (Metric for Evaluation of Translation with Explicit Ordering (Lavie and Agarwal, 2007)) is a metric used to evaluate the quality of machine translation systems. It was developed to address some of the limitations

of traditional translation evaluation metrics, such as BLEU, which are based on n-gram overlap between the machine-translated text and a reference translation. METEOR demonstrates a high level of correlation with human judgment and significantly outperforms BLEU. By adapting the scheme to two annotations we can use METEOR as an IAA metric.

The METEOR score is calculated by considering both explicit word-to-word matches in two texts as well as the word order between these words. It also takes into account the stem of the words, allowing it to handle inflections and other variations and synonyms. To achieve this METEOR maps one word in string 1 to at least one word in string 2, creating an alignment. It then identifies and chooses the largest subset of these word mappings and calculates the score using a combination of precision and recall, with a higher weight given to recall to reflect the importance of translating all of the content in the source text. The final score is normalized to a value between 0 and 1.

3.6.4 Jaccard Similarity

Jaccard Similarity, is a measure of the similarity between two sets. It is defined as the size of the intersection of the sets divided by the size of the union of the sets. The Jaccard Similarity is calculated using the following Formula 1. A and B signify the two compared sets. The nominator $|A \bigcap B|$ is the set intersection between A and B and the denominator $|A \cup B|$ is the set union of A and B.

$$J = \frac{|A \bigcap B|}{|A \cup B|} = \frac{|A \bigcap B|}{|A| + |B| - |A \cup B|} \tag{1}$$

The Jaccard Similarity is a simple and efficient way to compare the similarity of two sets or documents, but it only considers the presence or absence of elements in the sets. It can range from 0 to 1, with values closer to 1 indicating greater similarity and values closer to 0 indicating lower similarity.

3.6.5 OVERLAP Maximum and OVERLAP Minimum

Malik et al. (2021) use OVERLAP Maximum and Min as other IAA metrics. The OVERLAP Maximum is calculated with the following Formula 2, where A and B signify the two sets to be compared. $max(|A \bigcap B|)$ is the maximal overlapping sequence which is divided by the maximum size of the two sample $max(A, B)$. The OVERLAP Minimum (Formula 3) is calculated equivalently with the difference that the denominator is the minimum sample size out of A and B.

$$o_{max} = \frac{max(|A \bigcap B|)}{max(A, B)} \tag{2}$$

$$o_{min} = \frac{max(|A \bigcap B|)}{min(A, B)} \tag{3}$$

OVERLAP Maximum and Min are very similar to the Jaccard Similarity, but in contrast, these scores take the order of the entries of a set into account. This property makes them well-suited to calculate IAA, especially when considering entire sentences. By looking at the LCS, these scores are also similar to ROUGE−L even if somewhat more rudimentary in their execution.

3.6.6 BERTScore

The pre-trained language model BERT can be fine-tuned for various NLP tasks. Zhang et al. (2020) propose BERTScore an automatic sentence-pair classification, in which the model receives a pair of sentences as input and is required to predict whether the sentences are semantically equivalent or not. With those functionalities, BERTScore is related to BLEU and METEOR and like all the other IAA metrics ranges from 0 to 1, with values closer to 1 indicating high semantic equivalence between the input sentences and values closer to 0 indicating lower similarity. The BERTScore is calculated by encoding the input sentences using the BERT model and comparing the resulting representations using the cosine similarity measure. This allows a soft measure of agreement instead of exact-string matching. The result is a recall and a precision value for each corresponding token between two texts, which are then combined to compute an F1 measure. This F1 measure is the value used for the IAA. Zhang et al. (2020) show in their evaluation of the BERTScore, that it surpasses other existing metrics in correlation with human judgments. Since the SJP task uses BERT, this IAA metric is well-suited for evaluating the agreement between the annotations and the occlusion process.

3.7 Explainability Metrics

In this section, we will introduce the metrics we used in addition to the IAA to evaluate the results from the occlusion and LCI experiments.

3.7.1 Scaled Confidence

The temperature-scaled confidence or scaled confidence c as we call it in this thesis was suggested by Joel Niklaus (Niklaus et al., 2021). He has drawn attention to the fact that for the reliable usage of the prediction confidence a temperature scaling should be applied to it. The reason is that modern neural networks are likely overconfident in their predictions compared to the observed accuracy (Küppers et al., 2020). This makes calibration essential for this thesis since much of the used metrics are derived from the confidence estimates of the classification. The framework used for the temperature scaling is Küppers et al. (2020)'s framework net:cal (see Section 5.4 for implementation details).

3.7.2 Explainability Score

To analyze the change of a prediction from the baseline to an occluded fact section we introduce a metric we call explainability score. The explainability score S_{exp} (Equation 4) is the difference between the temperature-scaled confidence of the baseline from a case and the temperature-scaled confidence of the specific occlusion experiment of a case. It ranges between -1 and 1, with scores close to 0 indicating small to no change in confidence.

$$S_{exp} = C_{baseline} - C_{occlusion} \tag{4}$$

As negative and positive scores have different implications for different predictions (e.g., a negative score for prediction $p = 0$ indicates less confidence in the model's decision), we also introduced the normalized explainability score $normS_{exp}$. This normalization (see Equation 5) simplifies interpretation because it ensures that every negative explainability score indicates a decrease in confidence and every positive explainability score indicates an increase in confidence. Note that the normalized explainability score still ranges between -1 an 1.

$$normS_{exp} = \begin{cases} p = 0: & C_{baseline} - C_{occlusion} \\ p = 1: & -1 * (C_{baseline} - C_{occlusion}) \end{cases} \tag{5}$$

3.7.3 T-Test

The t-test is a statistical test that is used to determine if there is a significant difference between the means of two groups. The t-test can be used to compare the means of two groups that are independent of each other. The formula for the t-test depends on whether you are performing a one-sample t-test, a two-sample t-test for independent samples, or a two-sample t-test for dependent samples. To determine if there is a significant difference between the baseline and experimental measures of scaled confidence (denoted as $C_{baseline}$ and $C_{occlusion}$, respectively) and explainability score (with the baseline score being $S_{exp} = 0$), we perform a one-sample t-test in our occlusion and LCI analysis. The one-sample t-test is calculated with the following Equation 6. Where $\bar{x}$ is the sample mean, μ is the hypothesized mean or population mean, s is the standard deviation and n is the sample size.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{6}$$

3.7.4 Confidence direction

Using the normalized explainability score we can assign a confidence direction $conf_{dir}$ to each occlusion experiment (Equation 7).

$$conf_{dir} = \begin{cases} normS_{exp} > 0: & 1 \\ normS_{exp} = 0: & 0 \\ normS_{exp} < 0: & -1 \end{cases} \tag{7}$$

The confidence direction allows us to assign an explainability label to each experiment, giving an indication of how the model interprets this sentence. Table 2 shows the explainability label with their numerical values x_{label}. It is worth noting that the numeric values of the explainability labels are not assigned in the potentially intuitive way (1 to `Supports Judgment` -1 to `Opposes Judgment`) because removing a sentence supporting the judgment should theoretically lead to a decrease in confidence according to the legal perspective. The same metrics are also applied to the lower court experiments, but there the confidence direction is used to indicate if a lower court influences the model positively or negatively toward the decision.

Explainability Label	x_{label}	Explanation
Opposes Judgment	1	Removal should make decision more confidence
Neutral	0	Removal should not change anything
Supports Judgment	-1	Removal should make decision less confident

Table 2: Explainability labels explanations and their numerical values.

3.7.5 Explanation Accuracy Score

This section discusses the process of how the actual textual explanations from the occlusion were gathered and evaluated. Note that this section concerns only the occlusion experiments and not the LCI. We introduce the explanation accuracy score aS_{exp} (Equation 8). This score ranges between 1 and 0, with numbers close to 1 indicating a high level of legal accuracy of an explanation and values close to 0 indicating a low level. This score uses the confidence direction $conf_{dir}$ to split the results from the occlusion experiments into a set of correct classifications ($conf_{dir} = x_{exp}$) and a set of incorrect classifications ($conf_{dir} \neq x_{exp}$). For correct classifications, the explanation accuracy score takes the number 1. For incorrect classifications, it takes the mean IAA value calculated from the `BERTScore` (see Section 3.6.6) between this sentence and all the sentences labeled with the same numerical value as the model. For example, if the model classifies a `Neutral` sentence with a $conf_{dir}$ of -1 (`Supports Judgment`), the mean IAA between this `Neutral` sentence and all actual `Supports Judgment` sentences from the same case gets calculated giving us a score between 0 and 1.

$$aS_{exp} = \begin{cases} conf_{dir} = x_{exp}: & 1 \\ conf_{dir} \neq x_{exp}: & mean(BERTScore) \end{cases} \tag{8}$$

4 Dataset Description

In this chapter, we describe the construction and composition of the four datasets which form the basis of this thesis: The legal expert annotation dataset, with the resulting gold-standard annotation dataset, and the occlusion and LCI dataset for the experimental part of this thesis.

4.1 Legal Expert Annotation Dataset

The Federal Supreme Court of Switzerland is the final level of appeal and only hears cases that the lower courts before were unable to resolve satisfactorily. In their decisions, the supreme court analyzes probable incorrect reasoning by the lower court and therefore only focuses on a highly condensed version of the case. As a result, these decisions are often particularly challenging and sensitive due to their finality (Niklaus et al., 2021).

The dataset for the legal expert annotations consists of a subset from the SJP dataset presented by Niklaus et al. (2021) containing 108 cases. The SJP dataset is a multilingual, diachronic LJP dataset of 85K (50K German, 31K French, and 4K Italian) Swiss Federal Supreme Court cases that span over 21 years (from 2000 to 2020). In addition to the court documents the publication years, legal areas, and cantons of origin were annotated and added to the dataset. The SJP dataset is split into a training, validation, and testing set. For this annotation task, only cases from the test and validation were selected[1]. This approach was taken because the goal of this thesis was not to train a new model, but rather to explain the prediction published in the work of Niklaus et al. (2021).

The 108 cases are equally distributed among the three languages. Each language set contains six cases over six years (2015 until 2020), with each year having two cases per legal area. One with the judgment "approved" and one with the judgment "dismissed". It is important to note that even though our annotation dataset is balanced concerning the judgments, the SJP dataset is not (contains $\frac{3}{4}$ dismissed cases). In consequence, the training and validation set we use are not as perfectly balanced concerning the judgment as the Annotation and gold-standard Annotation sets. The chosen legal areas are categorized as penal, social, and civil law, which are the legal areas with the best performances in the SJP task Niklaus et al. (2021). For each legal area, one case has the verdict approved and the other has the verdict dismissed. In addition, preference is given to cases where the model decided the correct judgment from the facts given to it, with some outliers in the French and Italian subsets. Further selection criteria implemented in the dataset creation script (`prodigy_dataset_creator.py`) were the selection of short fact sections[2]. The reasoning behind this was that the performance of the SJP model deteriorated with the cases getting more complex (longer facts) Niklaus et al. (2021), consequently, the explanation quality produced by the occlusion method might also be worse. In addition, a shorter fact section would also speed up the annotation task. After a discussion with the annotator on some example court cases and their corresponding fact sections, we made the decision to enforce a minimum character length of over 2'000 characters. Table 3 shows the mean distribution of the number of tokens in the fact section per legal area and year. Observe that except for some outliers in the Italian dataset, the token amount per fact section is rather equally distributed among the years and legal areas. The German dataset also has the shortest facts, with the total mean being around 324 tokens, while in French and Italian we had around 421 respectively 415 tokens. One reason for the shorter German cases could be that there were generally more cases to choose from in the German subset being that it is the largest of the SJP dataset, so there were more short facts available.

Note that the annotation process did start as stated above with a dataset of 108 with almost perfect balance. But since the annotation process involves humans and is a cyclic process it is very dynamic. To ensure the quality of this dataset and the fact section it contained, the legal experts were able to ignore cases and in consequence filter them out. The reason for ignoring a case mostly was that the underlying facts did not have anything to do with the court's decision. As annotator 1 wrote (translated from German to English):

> The Swiss Federal Supreme Court dismissed the case because the complainants were not legitimate in the first place (no sufficient argumentation in their notice of appeal). The decision did not deal with substantive complaints. [In consequence the] dismissal in this case ultimately does not refer to the case itself, but "only" to formal defects.

Since three cases were ignored, we added three new cases to the German dataset, tallying the German set to 39 cases and the total annotation set to 111 cases. As shown in the next section describing the gold-standard dataset, the balance of the annotations even further diminished after the gold session.

[1] These sets consist of cases from 2015 to 2020.
[2] Implemented by ordering shortest first

Years	Penal Law	Social Law	Civil Law
German Dataset			
2015	342.5	335.0	300.0
2016	310.0	303.0	321.5
2017	341.0	318.5	346.5
2018	331.5	316.5	330.0
2019	341.5	309.0	322.0
2020	327.0	316.34	315.5
French Dataset			
2015	414.0	406.5	410.5
2016	421.0	401.5	414.0
2017	409.0	416.0	442.0
2018	422.5	406.0	433.5
2019	437.0	432.0	430.5
2020	405.0	407.0	474.5
Italian Dataset			
2015	423.0	389.0	541.0
2016	355.5	529.0	391.0
2017	466.5	387.0	489.0
2018	374.0	392.0	391.5
2019	443.5	389.5	401.5
2020	385.5	403.5	373.0

Table 3: Distribution of the mean number of tokens in a facts section over years and legal area.

4.2 Gold-Standard Annotation Set

An important part of this was the creation of a gold-standard annotation set, which was then used to create the occlusion dataset. The gold-standard annotations are the last step of the annotation process and contain the consensus of the conducted annotation. In addition, neutral sentences are annotated to provide sentence splitting for the occlusion method.

The gold-standard annotation set consists of 108 cases. As previously explained, they are not evenly distributed. The German subset contains 38 cases and is the largest. The French set has 36 and the Italian subset contains 34 cases[3]. In each language, the cases are evenly distributed among the two judgments "dismissed" and "approved". With all language dataset having 50% approved decisions and 50% of dismissed decisions. It is very interesting to see that this equal distribution in original judgment does not lead to an equal distribution in prediction (see Sections 4.3 and 4.4). This gold-standard set also contains a total number of 19 comments in all three languages.

Considering the distribution of the explainability labels (Figure 1) we can observe that overall in Italian more tokens were annotated. German has the least amount of tokens annotated in the gold-standard set. This may be due to differences in the fact length, mean sentence length, and the number of sentences in each language. The German fact sections are the shortest roughly containing about 14 sentences each with a mean sentence length of around 22 tokens. Italian and French facts are in general longer even when containing a lower amount of sentences both about 12 each, but with their sentences both having a mean number of approximately 33 tokens.

[3]The Italian dataset is smaller because during the gold session two unforeseen cases were removed after the annotator reexamined the latest Guidelines.

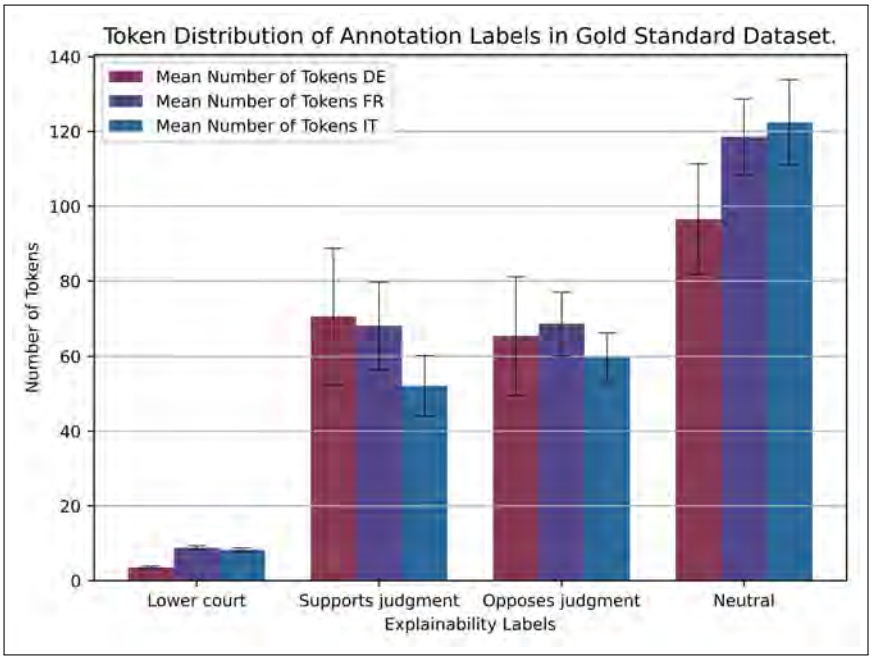


Figure 1: Distribution of the number of tokens per explainability label in the gold-standard dataset for each language.

Note that the sentence splitting is done manually by the annotators and according to the guidelines[4], so these numbers are only a rough estimation. We can also see that neutral sentences are the main component of facts according to the legal expert. This should be the case since the annotators should only annotate certain sentences with effective explainability labels. Another reason for this larger portion of Neutral sentences could be the case at hand itself, as one annotator put it (translated from German to English):

> The ruling regarding costs ensures that the request for free administration of justice is irrelevant. In addition, the Federal Supreme Court cannot decide on welfare. For this reason, 'many' sentences are neutrally annotated. [Since the judgment] refers to the local jurisdiction.

Further, it is interesting that the distribution of the explainability labels is quite similar in all languages. Consequently, we can assume that all three languages contain a similar amount of relevant sentences. Another reason for this equal distribution could also be that only one annotator labeled the Italian and French cases. This is a limitation of this thesis and further work could investigate if the distribution stays the same with more annotations from more legal experts.

4.3 Occlusion Dataset

The occlusion dataset is based on the gold-standard annotation set, with the difference that it only contains cases from the test set of the SJP dataset [5]. This decreases the number of distinct cases in each language set with German now having 27, French 24, and Italian 23 distinct cases. As mentioned in Section 3.2 we created four different datasets per language to conduct the occlusion analysis on the SJP task (occluding 1,2,3 and 4 sentences per experiment). From the previous section, we know that the number of tokens annotated in each explainability label is different. As a result, not all cases and languages contain the same amount of sentences per explainability label, making it so that the number of experiments is also quite different (see Table 4). The same reasoning applies to the occluded chunk length seen in Figure 2. In total, we have 28375 different occlusion experiments. German makes up the biggest part of the experiments since it also has the most distinct cases in the dataset. The French dataset is the smallest, which is to be expected since French facts have fewer sentences than German and a lower amount of tokens was annotated in French than in Italian. Looking at the mean chunk length[6] per experiment and language we see quite diverse distribution among the explainabil-

[4] In the guideline we defined a sentence as a self-contained linguistic unit consisting of multiple words, terminated with a period, semicolon, colon, question mark, or exclamation mark. We specifically allowed the splitting into sub-sentences.

[5] The annotation dataset and in consequence the gold-standard annotation set both contain cases from the validation and test set.

[6] For experiment 1 this is equal to the mean length of a sentence. For experiment 2 it is equal to the mean length of two sentences and so forth.

ity labels, particularly as more sentences are occluded. The difference is especially noticeable in the French experiments.

Language	Subset	Nr. of Experiments
German	de_1	427
German	de_2	1'366
German	de_3	3'567
German	de_4	7'235
French	fr_1	307
French	fr_2	854
French	fr_3	1'926
French	fr_4	3'279
Italian	it_1	299
Italian	it_2	919
Italian	it_3	2'493
Italian	it_4	5'733
	Total	28375

Table 4: Number of Occlusion Experiments per language and version

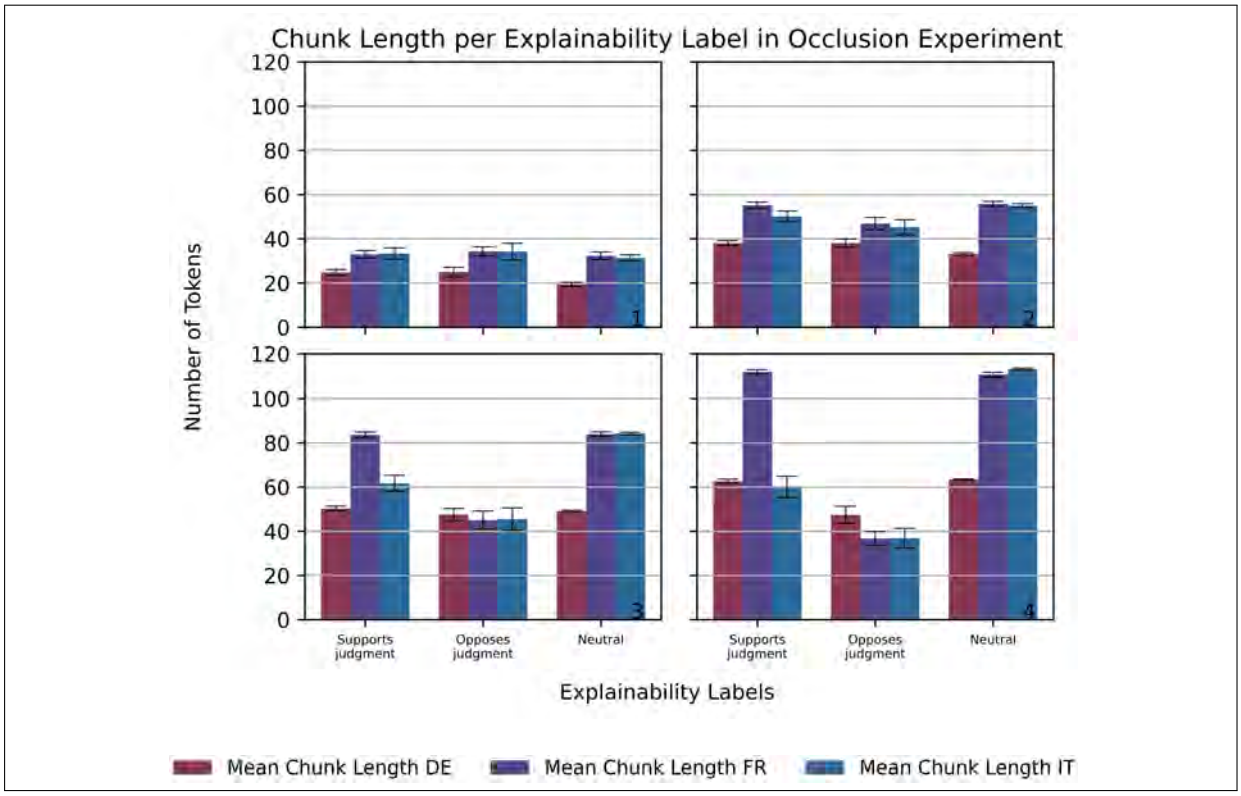


Figure 2: This plot shows the mean chunk length for each occlusion experiment. The numbers in the right corner indicate the experiment number (amount of sentences occluded).

4.4 Lower Court Insertion Dataset

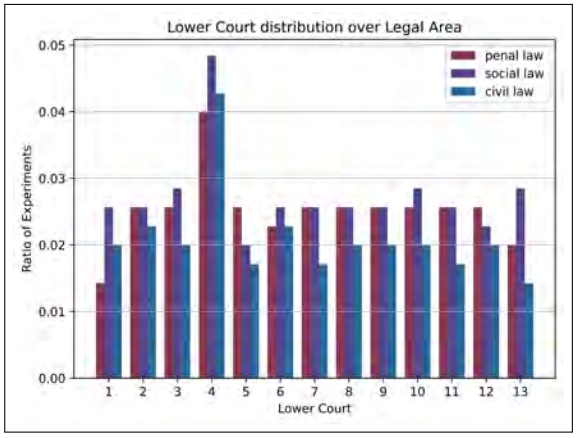
The LCI test sets are based on the test set part of the gold-standard annotation set, similar to the occlusion test sets. Thus, we have again 27, 24, and 23 distinct cases for German, French, and Italian respectively. This results in a total of 378 experiments in German, 414 in French, and 335 in Italian (Table 5).

Looking at the distribution of the lower courts in the dataset (Figure 3) we can see that except for some outliers they are all rather equally distributed in the dataset. Looking at the German distribution only the GL_VGer (Administrative Court of Glarus) appears more than others. In the Italian dataset, we have a higher occurrence of TI_TRAP (Appeal Court of the Canton of Ticino) and TI_CCivAP (Civil Chamber of the Appeal Court of the

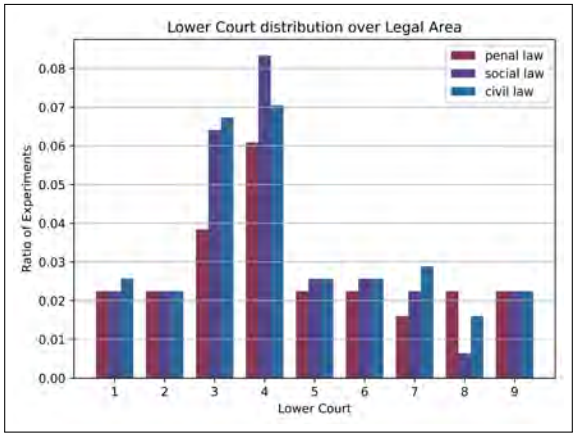
Language	Subset	Nr. of Experiments
German	de	378
French	fr	414
Italian	it	335
	Total	1'127

Table 5: Number of Lower Court Insertion experiments per language

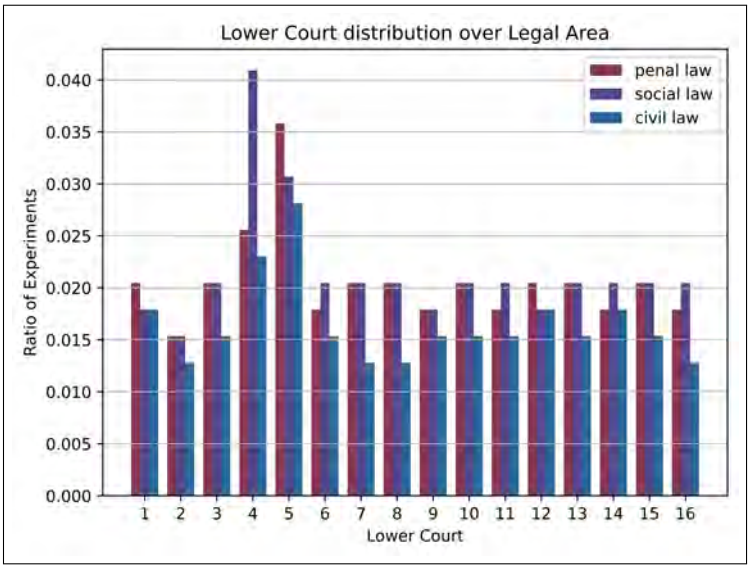
Canton of Ticino). In the French distribution of the courts, we have a higher occurrence of the VD_CASoTC (Social Insurance Court of the Cantonal Court of Vaud) and VD_ChRPeTC(Criminal Appeals Chamber of the Vaud Cantonal Court). We can also observe that social law is more prominent than other legal areas among the lower courts in all three language datasets.



(a) German



(b) French



(c) Italian

Figure 3: Distribution of the lower courts per legal area. Note that this plot shows the ratio of lower courts in the dataset split among the different legal areas.

Examining the represented cantons: In German nine cantons out of the 21 German-speaking cantons are represented (Berne is a multilingual Canton speaking both German and French). With Italian, two cantons are represented (Ticino and Grisons) these are also the only regions in Switzerland where Italian is one of the official languages. In the French dataset six cantons are represented with the canton Fribourg and Valais having both French and German as their official language. With these compositions, the Italian and French datasets have a complete amount of cantons in other words all French-speaking and Italian-speaking cantons are at

German Lower Courts		Italian Lower Courts		French Lower Courts	
1: AG_OGer	9: LU_KGer	1: CH_BVGer	9: TI_TRAP	1: CH_BGer	9: NE_CPe
2: AG_VGer	10: SZ_KGer	2: GR_VG		2: CH_BVGer	10: NE_CPuTC
3: AR_KGer	11: SZ_VGer	3: TI_CARP		3: FR_CAPCiv	11: VD_CAPPe
4: BE_Oger	12: ZU_KGer	4: TI_CCivAP		4: GE_CJ	12: VD_CASoTC
5: BE_VGer	13: ZU_OGer	5: TI_CEFTRAP		5: GE_CJRC	13: VD_ChCTA
6: BL_KGer		6: TI_CPTRAP		6: GE_ChRPeCJ	14: VD_ChRPeTC
7: BS_AppGer		7: TI_CRPTA		7: JU_CCiTC	15: VD_TC
8: GL_VGer		8: TI_TCAS		8: NE_CEA	16: VS_ChCivTC

Table 6: Legend of Figure 3

least represented once. The total amount of cantons represented in this dataset is 17 out of 26 which is about 2/3 of the cantons. Looking representation of each region in Table 7 we can see that each region is represented across the different language datasets. The R. lémanique and Ticino have the highest occurrence across the dataset, even when not appearing in the German dataset. This is due to the fact that these language regions are not as geographically dispersed across Switzerland and therefore make up the bulk of their respective language set. The German cantons are more spread across Switzerland hence we see a more diverse region distribution in this language set. The only two less-represented German regions are E. Switzerland and Zürich.

Regions according to \citet{swiss_regions}	**German LCI**	**French LCI**	**Italian LCI**	**Total**
Région lémanique (VD, VS, GE)	0,00%	66,30%	0,00%	23,67%
Espace Mittelland (BE, FR, SO, NE, JU)	21,37%	22,28%	0,00%	15,23%
Nordwest Switzerland (BS, BL, AG - 980)	28,49%	0,00%	0,00%	9,70%
Zürich (ZH)	13,68%	0,00%	0,00%	4,66%
East Switzerland (GL, SH, AR, AI, SG, GR, TG)	7,41%	0,00%	0,00%	4,75%
Central Switzerland (LU, UR, SZ, OW, NW, ZG)	29,06%	0,00%	7,37%	9,89%
Ticino (TI)	0,00%	0,00%	85,26%	25,80%
Federation (CH)	0,00%	11,41%	7,37%	6,30%

Table 7: Distribution of the regions in the different LCI datasets and in total

5 Experiments

In this section, we detail the process of implementing the legal expert annotation task using Prodigy, a proprietary Python library and describe the necessary post-processing steps required to prepare the annotated data the occlusion and LCI experiments with hierarchical BERT. In addition, we present the implementation of the IAA score and the temperature scaling and further discuss how we derived the occlusion and LCI test sets form the gold-standard Annotations and the steps we took to run the experiments so that they resemble the SJP task as closely as possible.

5.1 Explainability Annotations with Prodigy

In this section, we describe the implementation of the legal-expert annotation task. To implement the interface for the legal expert annotation we decided to use Prodigy. Prodigy is a proprietary python library with a range of pre-built workflows, command-line commands, and other components well suited for conducting annotations. Prodigy works of so-called recipes (python scripts), which customize the annotation recipe. For this thesis' annotation task two such custom recipes were created. To deploy the Prodigy application we used Docker and the server infrastructure provided by the Research Center for Digital Sustainability. The first version of the Prodigy Docker setup was provided by Nyffenegger [7].

Creating the custom Prodigy recipe was a straightforward process since they only require a few lines of code. A Prodigy recipe is a Python function returning a dictionary of its components. The arguments of the recipe function will become available from the command line so that the passing of parameters is possible. Using the following Plac syntax the recipe can be given a custom name and a variable number of arguments for the annotations. For this annotation task, we customized the built-in `spans.manual` recipe. Source Code 1 shows an excerpt of the recipe for the facts-annotation task. This recipe lets an annotator mark entity spans in a text by highlighting them and selecting the respective labels (e.g. "Lower court"). It first loads the correct language model for the tokenization of the input text (a fact section), then loads the input file (JSONL), and assigns the accurate output dataset and port. Next, it uses the built-in block Annotation Interface[8] to display the judgment, the link to the judgment, the fact section, and the free text explanation field. Note that each annotator had a custom link o conduct the annotations individually by adding the suffix `?session=annotator_name`.

After an annotation is completed Prodigy dumps the annotations as a nested string in a database file, which was previously defined in the configuration file (prodigy.json). The annotations from the database can be extracted as a JSONL file using Prodigy's built-in `db-out`. Thesis JSONL files are the basis for the conducted annotation analysis and the construction of the occlusion experiment test set.

```
@prodigy.recipe(
    "facts-annotation",
    language=("The language to use for the recipe.", 'positional', None, str),
)
# function called by the @prodigy-recipe definition
def facts_annotation(language: str):
    nlp = spacy.load(f"{language}_core_news_sm") # Load the spaCy model for tokenization.
    stream = JSONL(f"annotation_input_set_{language}.jsonl")# Input dataset
    dataset = f"annotations_{language}" # Output dataset
    port = PORTS[language]

    # Tokenize the incoming examples and add a "tokens" property to each example.
    stream = add_tokens(nlp, stream, use_chars=None)

    return {
        "dataset": dataset,
        "view_id": "blocks",
        "stream": stream,
        "config": {
            "port": port,
```

[7] The original work of Nyffenegger (2022) can be found in this https:github.comSkatingerSwissCourtRulingCorpustreeprodigyprodigy. He created the Dockerfile and bash scripts which were later adapted to fit the purpose of this annotation task. Note that in the process of this thesis his work was merged with the main branch of SCRC, the full code of the Prodigy implementation can be found in the https:github.comninabaumgartnerSwissCourtRulingCorpustreeprodigyscrcannotationjudgment_explainability.

[8] Please reference the annotation guidelines in the appendix 8 for screenshots of the interface.

```
            "blocks": [
            # define blocks here
            ]
        },

    }
```

Source Code 1: Excerpt facts_annotation.py

5.2 Post-processing of Annotations

To conduct the occlusion and apply the IAA scores to the legal expert annotation and subsequently to the occluded sentences, we first must do some extensive post-processing. The difficulty with the prodigy output format is that the resulting JSONL file is very nested. For example, the token with their entries text, start, end, id, and ws[9] are stored in a separate dictionary from the spans (the actual annotation). The separate span dictionary contains the start and end of the span, the label assigned by the annotator, and the ids of the tokens in the span, but not the text of the span. An additional problem that we encounter is that some fact sections (of the same case) are not tokenized the same as their counterpart in another annotation. The reason for this is that in the process of the annotation we add line breaks for better readability to the facts section. In consequence, annotations conducted after this change have a different tokenization (meaning different ids for the same word at the same position) than before. For IAA scores that require only a textual input, this is not a problem since the sentences stayed the same, but some scores are implemented with a numerical comparison (comparing the token id of a word), which means that a normalization of the token id is necessary. For those reasons, we had to do extensive merging, rearranging, and normalization of the data, which took a lot of time and postponed the analysis process drastically. After post-processing, the annotations for each document were stored with each row representing a document and its corresponding annotations (see Table 8). Once the annotations are transformed into this usable form, we are able to take the steps for the occlusion and LCI and apply the IAA scores.

id	tokens_text_1	normalized_tokens_dict	normalized_tokens_1
474906	Verwaltungsgericht des Kantons Zürich	{'Verwaltungsgericht': 9, 'des': 10, 'Kantons': 11, 'Zürich': 12, 'das': 8, 'Nan': 10000}	[9, 10, 11, 12]

Table 8: Excerpt of the preprocessed annotation table

5.3 Implementation of Inter Annotator Agreement Scores

The implementation of `ROUGE`, `BLEU`, `METEOR`, and `BERTscore` was straightforward since we could use python libraries. For `ROUGE` we could use the native python implementation (Google LLC, 2022), designed to replicate the results from the original `perl package`. For the `BERTscore`, there was also a python package (Zhang et al., 2022) published with the paper. `BLEU` and `METEOR` are both provided by the `nltk.translate` python package (Loper and Bird, 2002). Note that for the `Jaccard Similarity` as well as for the `OVERLAP Maximum` and `OVERLAP Minimum` the token id is used to calculate IAA. This is the reason the normalization is necessary so that the same word equals the same token id and the longest common sequence and set intersection can be correctly calculated. For the more sophisticated scores (`ROUGE`, `BLEU`, `METEOR`, and `BERTScore`) the annotated respectively occluded text could be used.

The `Jaccard Similarity` is implemented using Formula 1 and using Sidyakov (2021) article as a guide. The `OVERLAP Maximum` and Min following Formulas 2 and 3 are implemented by first getting the maximal and Minimal sample size (the nominator) and then calculating the LCS between each annotation by comparing their normalized token id lists.

5.4 Temperature Scaling

As mentioned in our methods section, we use Küppers et al. (2020)'s framework net:cal to temperature scale our confidence estimates. `Net:cal` is a Python 3 calibration framework library for measuring and mitigating the miscalibration of uncertainty estimations by a neural network.

[9] The value *ws* is a boolean indicating the presence of a whitespace character after the token.

For a binary classification task like the SJP one method of post-processing calibration, is temperature scaling. This method was proposed by Guo et al. (2017). It applies a confidence mapping g on top of a miscalibrated scoring classifier $\hat{p} = h(x)$ to provide a calibrated confidence score $\hat{q} = g(h(x))$. In other words, the confidence estimates are scaled to the calibrated confidence estimates. We implemented the net cal temperature scaling in the following way (see Source Code 2 for the function `temp_scaling()`).

```
def temp_scaling(df: pd.DataFrame) -> pd.DataFrame:
    """
    Replaces the judgment labels with int 0,1.
    Creates two NumPy 1-D and 2-D arrays.
    Applies temperature scaling to the values and returns calibrated DataFrame

    Uses TemperatureScaling() from Kueppers et al.
    via https://github.com/EFS-OpenSource/calibration-framework#calibration-framework
    """
    df["prediction"] = np.where(df["prediction"] == "dismissal", 0, 1)
    # ground truth digits between 0-1 - shape: (n_samples,)
    ground_truth = np.array(df["prediction"].values)
     # confidence estimates between 0-1 - shape: (n_samples, n_classes)
    confidences = np.array(
        df[["prediction", "confidence"]].values)

    temperature = TemperatureScaling()
    temperature.fit(confidences, ground_truth)
    df["confidence_scaled"] = temperature.transform(confidences)
    return df
```

Source Code 2: Implementation of temperature scaling using net:cal.

As one can see from the code above we called the Class `TemperatureScaling()` after transforming the prediction values from dismissal/approval to 0/1 and creating two `numpy.arrays` one for the ground truth (only binary prediction) and one with the `prediction` and `confidence` variables. Then as suggested by the `net:cal` documentation for the temperature scaling we used methods `fit()` to build a logistic calibration model and `transform()` to get calibrated outputs of uncalibrated confidence estimates. These scaled values are then scored under the column "scaled_confidence" and were used for the explainability score S_{exp} and the confidence direction $conf_{dir}$ presented in the next sections.

5.5 Occlusion with hierarchical BERT

Remember that the occlusion test set only contains cases from the original SJP test set. Since the gold-standard dataset contains both cases from the validation and test set, the first step after the post-possessing described in Section 5.2 is the filtering out of the validation cases. This leaves us with cases from the years 2017 to 2020, which decreases the total number of different cases in each language set with German now having 27, French 24, and Italian 23 distinct cases. We then split the annotations into sentences again since the applied post-post-processing joins all annotations for each label. This is done using the original prodigy output file (JSNOL), which gives us a span dictionary where each span or annotated sentence is a separate entry.

With the split sentences or sub-sentences for each case, we can do the permutations. For experiment 1 we occlude each sentence in a case once adding no marker or trace of the occlusion in the fact section to leave it as similar and natural as possible. This produces one row per experiment containing the fact section without the occluded part, a field for occluded text, and the corresponding explainability label the occluded text was assigned by the annotators. For experiment 2 we compute all combinations of two sentences for each explainability label and case. For example, if sentences A, B, and C are annotated with the label `Supports Judgment`, we once occlude combination AB, AC, and BC. We then apply the same steps for experiments 3 and 4 using sentence combinations of 3 and 4 sentences respectively. This workflow leaves us with four different test sets per language, each having different parts of the text occluded and annotated with the appropriate explainability label. Concerning the baseline of these experiments: As described in Section 3 we do not change the conditions for the model's prediction and use the same training and validation set as Niklaus et al. (2021). In consequence, our baseline for the occlusion and the LCI consists of a total of 74 distinct cases contained in the three-language dataset. We do not change anything in their fact section, meaning there is no occluded text. They are labeled with the explainability label “Baseline" to distinguish them from the experiments.

For the experimental setup, we again tried to keep it as similar to Niklaus et al. (2021) previous experiments as possible. During the training process, they focused on cases that represent the minority class (approval) by oversampling them. To train and evaluate the BERT-based methods, they used Early Stopping on the development data, a learning rate of 3e-5, and a batch size of 64. For hierarchical BERT, the maximum sequence length was set to 2048. We evaluate the models using the macro-F1 score on the test set, which takes into account both classes and gives more weight to examples from the minority class. Each experiment is run with a single random seed[10] on a single GeForce RTX 3090 GPU with mixed precision and gradient accumulation. In addition, we use the Hugging Face Transformers library (Wolf et al., 2020) and the BERT models available from the Hugging Face.

5.6 Implementation of Lower Court Insertion with hierarchical BERT

The LCI has a different kind of test set than the occlusion, but the rest of the above-described setup and framework also applies to the LCI. In this experiment, we try to investigate bias from one lower court to another. To create the lower court test set we apply the same post-processing to the gold-standard annotation as described in Section 5.3. Therefore, we also obtain the same number of distinct cases as in the occlusion dataset. We then extract all distinct entities of the lower court. This results in 13 lower courts for the German, 16 for the French, and 9 for the Italian dataset. For a better overview and normalization, we abbreviate each lower court (see Appendix, Table 22). After extracting each lower court and inserting a provisional marking at the location of extraction, we add a column containing the original lower court. This is important to distinguish the baseline from the experiments. Finally, we insert each lower court once except for the original lower court. Note that we do not apply any normalization to the lower courts before inserting them, the abbreviation seen in the LCI dataset description is only added in the analysis step of this thesis. In consequence, some lower courts appear twice in the test set[11], but are later summarized into one court.

[10] seed 2, which provided the best results according to Niklaus et al. (2021)

[11] One example would be the BE_OGer, which appears twice once as "Obergericht Bern" and once as "beim Obergericht Bern".

6 Analysis

In this section, we will present the key findings from our quantitative analysis of the annotations, the occlusion, and the LCI experiments. We will also discuss the methods we used to evaluate and visualize the results, including the metrics we employed and how we aggregated the data. We aim to provide a comprehensive and detailed overview of the outcomes of our experiments, so you can fully understand the significance of our findings.

6.1 Main Results – Legal Expert Annotations

In this section, we will discuss the main results of the legal expert annotations. First, we will have a look at the overall explainability label distribution, then briefly explain how we implemented IAA, and then presented the results from the IAA agreement scores.

6.1.1 Explainability Label Distribution

As mentioned in Section 3.5, a total of three cycles were conducted in this annotation task. Figure 4 below illustrates the distribution of the explainability labels in the German dataset[12]. In the German subset, all three annotators used all three labels. Annotator 1 annotated the least amount of tokens[13], while annotator 3 annotated the most, especially when using the `Supports Judgment` label. One annotator only annotated the French dataset. While the Italian set was annotated by two of the experts, expert 1 was only able to work on a handful of cases before resigning. This is the reason why in the following IAA analysis only the German dataset was considered. The equivalent figures of these languages can be found in the Appendix (Figures 21 and 22).

Note, that the used IAA scores and the analysis tools can be adapted to all the remaining languages. The extension of the annotations and, in consequence, the analysis could be a focus for future work.

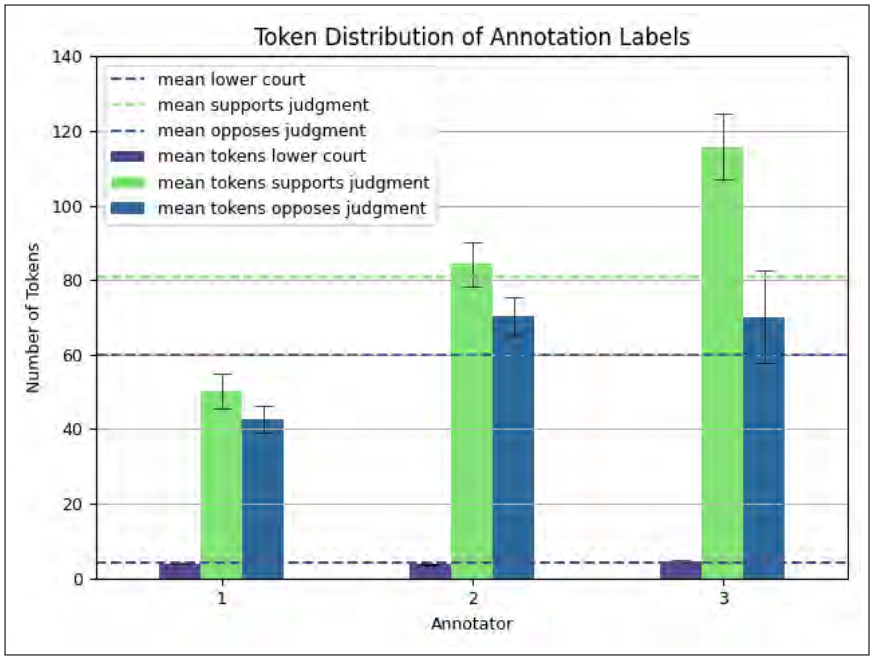


Figure 4: Distribution of the number of tokens per explainability label for the different annotators in German cases.

6.1.2 Inter Annotation Agreement Results

Table 9 shows the mean IAA for each score for the German dataset in the first cycle of the annotations. Note that in this table the agreement for the individual labels is combined. The full results for each label can be found in the appendix (Table 20). The shown IAA scores are calculated as explained in Sections 3.6 and 5.3. We show only the results for the first round since in the second cycle only two annotators were involved and only some new annotations were added. The agreement is roughly the same (see Table 21 in the appendix).

[12] Latest version of the annotation before the gold-standard

[13] In the context of this thesis the term "tokens" references the words in a string. However, since it cannot be guaranteed that the automatic tokenization has worked perfectly, we use this term.

IAA Score	A1 and A2	A1 and A3	A2 and A3	Mean Score
bert_score	0.91	0.868	0.938	0.905
bleu_score	0.757	0.695	0.855	0.769
jaccard_similarity	0.731	0.64	0.822	0.731
meteor_score	0.776	0.71	0.884	0.79
overlap_maximum	0.689	0.612	0.745	0.682
overlap_minimum	0.836	0.735	0.816	0.796
rouge1	0.788	0.696	0.877	0.787
rouge2	0.749	0.649	0.852	0.75
rougeL	0.779	0.687	0.873	0.779

Table 9: Mean Inter Annotator Agreement between the annotators in the first cycle

As one can see in Table 9 we achieved very high agreement for each of the scores, with values ranging between 0.7 and 0.9. Using BERTScore we achieve the highest agreement. The reason for this could be that this metric has the best ability to find similarities in non-exact matches. OVERLAP Minimum also achieves good results, which could indicate that the annotations were often a subsequence of each other e.g. different starting points of the same sentences. The OVERLAP Minimum illustrates this fact since the longest common subsequence is divided by the smaller sample size. Looking at the agreement between the different annotators: Legal experts 2 and 3 have the highest agreement. This could be because these two legal experts were involved until the end of the annotation process and received the best training and the most up-to-date guidelines.

We now want to look at some more detailed visualization of these results. The violin plot in Figure 5 shows the IAA calculated with BERTscore, METEOR, OVERLAP Minimum, and ROUGE – L for all three explainability labels per annotator. The agreement in the categories LowerCourt and Supports Judgment is very high. With the minimum for both scores being at around 0.4 and most of the values accumulating in the upper interquartile range. We can also observe that the results for these labels are very similar for all four scores, with the difference that the range of values with BERTScore and ROUGE – L is smaller and even more pronounced in the upper field than with the other two scores.

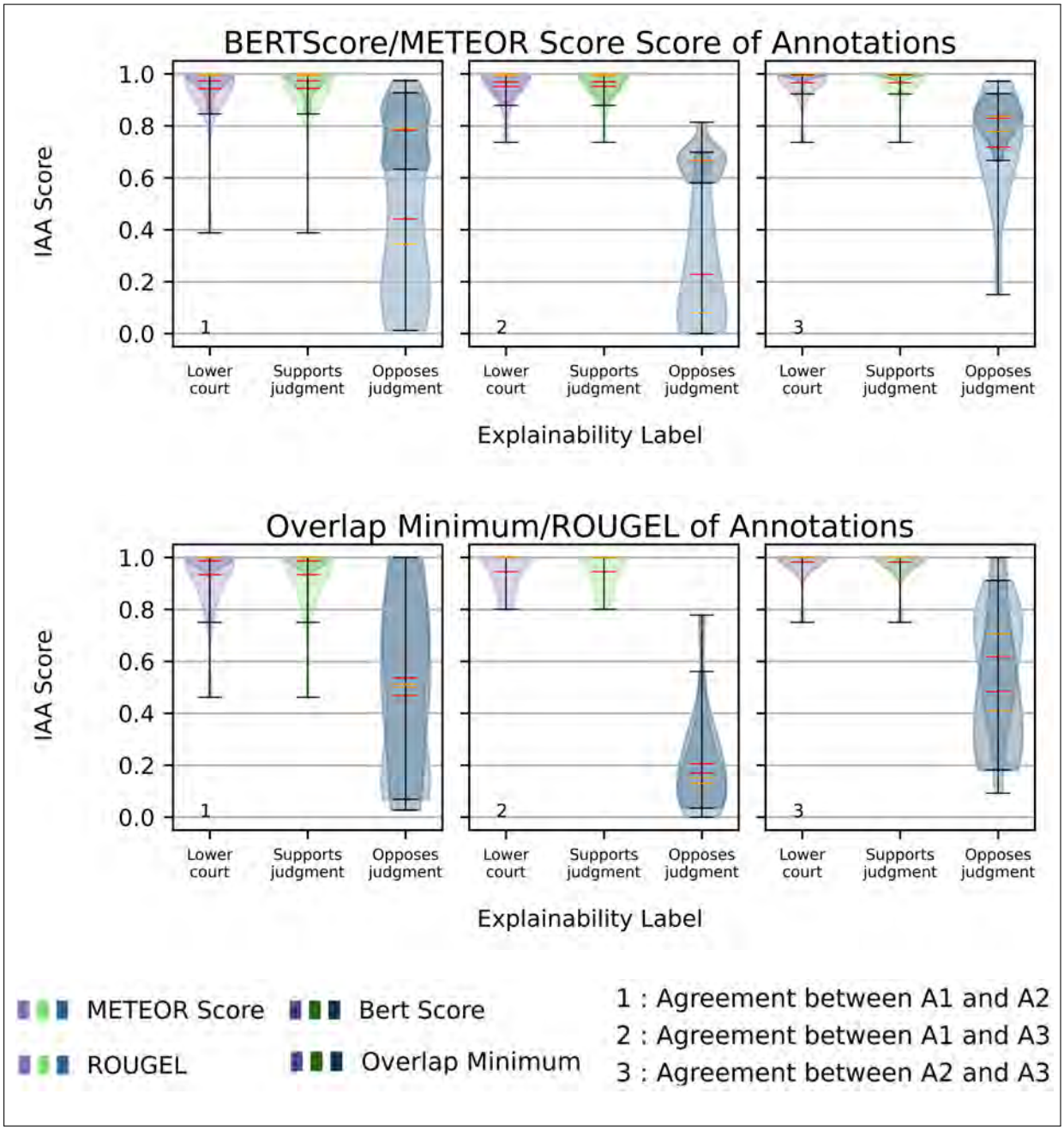


Figure 5: Results of BERTScore, METEOR, OVERLAP Minimum and ROUGE – L in the first cycle. The darker parts are the results from BERTScore and ROUGE – L, the lighter parts show the results from METEOR and OVERLAP Minimum. The red dash indicates the mean and the orange dash indicates the median. The numbers at the bottom indicate annotator combinations.

Looking at the Opposes Judgment label we notice a bigger span of values, especially with METEOR, ROUGE – L and OVERLAP Minimum. Here the values accumulate in the upper and lower quantile with the plot being skinnier in the middle. Except for the agreement between annotators 1 and 2, where the violin is more pronounced in the lower third. A reason for those worse results in theOpposes Judgment category could be that these sentences were harder to identify, thus leading to more conflicting annotations. The experts also confirmed this during the discussion regarding the gold-standard annotations. They explained that depending on the judgment, a very similar sentence, most of the time at the end of the facts, could be either opposing or supporting the judgment depending on the verdict of the case. If an expert did not remember the verdict or did not read the sentence word for word, it was possible that these sentences were annotated in opposite ways. This is also an example of a conflict that was resolved during the gold-standard annotations. The Prodigy interface clearly indicates opposing annotations, which makes it very easy to resolve conflicting annotations. Another possibil-

ity to increase the agreement in the `Opposes Judgment` category would be the implementation of more cycles in the annotation. Looking at the results of this second round (Table 21 in appendix), we see that this is no guarantee for higher agreement since the results in this category were very similar to the first round.

In conclusion, the results from the legal expert annotations are very promising. Even with a few rounds, we achieve an overall agreement of approximately 0.78. These results speak for the quality of the guidelines and the training the legal expert received. It also gives us a good indication of the quality of the gold-standard annotation and the resulting explainability dataset. The results of the IAA calculations showed that the annotations were generally consistent, with most scores falling above the 0.4 threshold for acceptable agreement. However, there were some areas where additional training or clarification was needed. These issues were addressed before conducting the gold-standard annotations to ensure the quality of the annotations. Overall, these IAA played a crucial role in ensuring the reliability and trustworthiness of the Legal Expert Annotations, which serve as the ground truth for evaluating the model-generated explanations in this thesis.

6.2 Main Results – Occlusion

In this section, we present the results from the occlusion method. Specifically, we aim to identify any patterns or trends that may shed light on the reason why the model chose a certain prediction. For this purpose, we look at the overall performance and examine correctly and incorrectly classified sentences. We also look at the general trends for the explainability label and present our results from the IAA study and the obtained textual explanations.

To examine the overall performance of the occlusion with hierarchical BERT (seen in Table 10) Niklaus et al. (2021) suggests that the macro-average measures Macro-F1 is a more appropriate measure. The reason is that the classification task is quite challenging when label skewness is present. In other words, it can be challenging to outperform dummy baselines (such as always predicting the majority class). Hence the Macro-F1 measures take into account the performance of both classes and are able to distinguish between more effective methods. These measures favor models that are able to learn the task and are able to accurately classify the two classes, rather than simply predicting the majority class every time. The performance of the model is similar between the German and French subsets, which have 35K and 21K training samples, respectively. However, their performance is much worse in the Italian subset, which has only 3K training samples. In German and French, occlusion performs better than the original SJP. In Italian, the SJP outperforms occlusion in every experiment. The good results in German and French may be due to the selection of mostly priorly correctly predicted cases in the original SJP task, leading to a higher probability that the model will predict the correct judgment again.

Model	de		fr		it	
	Macro-F1 (SJP)	Macro-F1 (OCC)	Macro-F1 (SJP)	Macro-F1 (OCC)	Macro-F1 (SJP)	Macro-F1 (OCC)
hierarchical (two-tier 4× 512 tokens)						
1 Native BERT	68.5 ± 1.6	**86.0**	70.2 ± 1.1	**88.9**	**57.1 ± 6.1**	53.4
2 Native BERT	68.5 ± 1.6	83.2	70.2 ± 1.1	85.9	57.1 ± 6.1	54.0
3 Native BERT	68.5 ± 1.6	82.8	70.2 ± 1.1	82.7	57.1 ± 6.1	50.3
4 Native BERT	68.5 ± 1.6	81.8	70.2 ± 1.1	881.1	57.1 ± 6.1	44.7

Table 10: Comparison between the results from the SJP and the Occlusion experiments using the Macro-F1. The numbers at the beginning of the row indicate the Occlusion experiment. The models were all trained and tested in the same language. "Native BERT" refers to the BERT model that was pre-trained in that language. The best scores for each language are highlighted in bold. In the SJP results, the standard deviation between different seeds is shown, but in the Occlusion results, the model was only run using one random seed, so this information is not necessary.

Next, we examine the prediction distribution of each experiment using Table 11. Notice that the baseline predictions except for Italian are relatively equally distributed. For all the experiments the predictions for German and Italian lean more towards dismissal while for French they stay rather equally distributed. The ratio of

dismissal in German becomes more extreme as more sentences are occluded, while in French, the ratio moves slightly towards approval, but to a much lesser degree. In Italian, the initial strong unequal distribution toward dismissal gets less pronounced with a higher experiment count.

Language	Approved Baseline	Dismissed Baseline	Approved Occlusion	Dismissed Occlusion
1 Sentence Occlusion				
German	44,44%	55,56%	38,11%	61,89%
French	50,00%	50,00%	51,96%	48,04%
Italian	21,74%	78,26%	21,72%	78,28%
2 Sentence Occlusion				
German	44,44%	55,56%	28,97%	71,03%
French	50,00%	50,00%	51,72%	48,28%
Italian	21,74%	78,26%	25,59%	74,41%
3 Sentence Occlusion				
German	44,44%	55,56%	20,03%	79,97%
French	50,00%	50,00%	54,36%	45,64%
Italian	21,74%	78,26%	34,65%	65,35%
4 Sentence Occlusion				
German	44,44%	55,56%	13,11%	86,89%
French	50,00%	50,00%	58,86%	41,14%
Italian	21,74%	78,26%	42,91%	57,09%

Table 11: Prediction Distribution in Occlusion Experiments

For the more detailed analysis of the occlusion, we used the explainability score S_{exp}, the scaled confidence $C_{occlusion}$, and the confidence direction $conf_{dir}$. Using these metrics we are able to analyze the changes in the predictions from the occlusion. In the next section, we explain how we aggregated these scores and visualized the results.

6.2.1 Impact of correctly Classified Sentences

To examine the impact an occluded sentence has on the model we first separate the occlusion results into a set of correct and incorrect classifications using the confidence direction. For example, if a sentence chunk was originally annotated with the explainability label `Supports Judgment` ($conf_{dir} = -1$, see Table 2) and occluding it also resulted in a negative $conf_{dir}$, the model classifies this sentence chunk correctly. We then apply the t-test to the explainability score and to the scaled confidence direction and visualize the results using the scatter plots below. Note that points plotted as significant must have significant differences[14] to the baseline in both their means. It is also important to note here that we remove all sentences the model classified as neutral (models $conf_{dir} = 0$) since they do not matter to us as they did not impact the prediction.

We observe the impact of correctly classified sentences in Figures 6,7, 23 and Table 12. The maximum amount of correctly classified sentences can be found in the French experiments with a ratio of around 10% for the `Supports Judgment` category. We can also observe that same as with the overall performance the model's sentence classification was the worst in Italian. In general, we see much more correctly classified sentences with `Supports Judgment` than with`Opposes Judgment`. This is probably due to the fact that the `Opposes Judgment` category is least represented in the occlusion dataset (the least amount of tokens annotated), but it could also show a weakness of the model with classifying sentences of this label.

[14] Calculated using the one-sample t-test with significance levels $\alpha = 0.05$ value

Language	Supports Judgement	Opposes Judgement
1 Sentence Occlusion		
German	9,43%	6,74%
French	9,25%	7,12%
Italian	7,49%	4,12%
2 Sentence Occlusion		
German	9,16%	3,51%
French	10,34%	3,09%
Italian	4,15%	1,53%
3 Sentence Occlusion		
German	8,41%	1,38%
French	13,61%	1,23%
Italian	1,45%	0,51%
4 Sentence Occlusion		
German	7,64%	0,46%
French	15,71%	0,59%
Italian	0,43%	0,18%

Table 12: Distribution of the correctly classified sentence chunks per explainability label, language and experiment

Investigating the correct classification further, first of all, we want to draw attention to the x-axis of all upcoming plots, which depicts the “scaled confidence direction". The “scaled confidence direction" is the scaled confidence of a particular occlusion experiment multiplied by the confidence direction. This results in negative values for `Supports Judgment` since this label has confidence direction -1, which makes understanding the plots much easier. In addition to this simplification we also only look at the German and French plots in detail, since in Italian there were only a few correct classifications (see Figure 23 in the Appendix). When examining Figures 6 and 7 we are looking for clustering at the limits of the x-axis and y-axis. In other words, we want a lot of orange points on the left bottom quadrant and a lot of dark green points on the right upper quadrant. The reasoning behind this is that with opposing confidence directions (-1 for `Supports Judgment` and 1 for OPPOSE JUDGMENT) the desired output of the model should not only be a correct classification (assigning the right sign) but a strong correct classification (high negative/positive values).

Looking at the German plots (Figure 6) we observe that they all look rather similar except that the amount of correctly classified sentences and therefore points increases. Overall, as with the distribution, there is a more significant impact with the `Supports Judgment` category (orange), especially when looking at experiments 2 and 3. With this explainability label, we see the desired effect quite nicely illustrated, with a mixed orange cluster of both predictions gathering at the lower left side of the plot. We also see some significant impacts for `Opposes Judgment` in experiment 2, with some clustering at the middle right of this plot. We can further observe that the predictions seem rather equally distributed for the German correct classification, showing both some approved and dismissed predictions in all the plots.

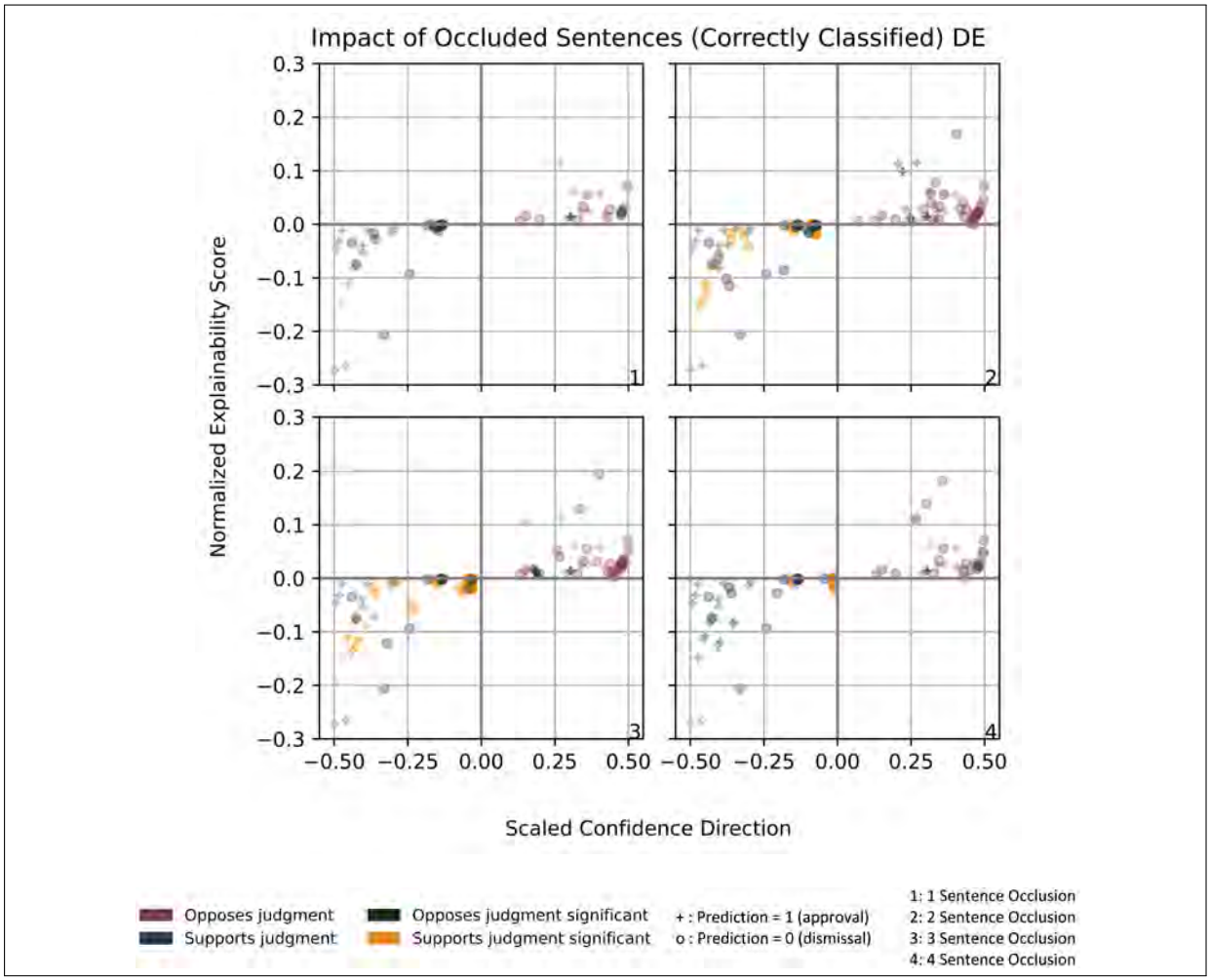


Figure 6: This plot shows the impact each correctly classified sentence has in German on the prediction. The further away a point is from the null axis the more impact it has on the model's prediction. The different markers indicate the prediction and the number on the bottom left is the experiment number.

Looking at the French results in Figure 7 we see much more pronounced plots, especially for `Supports Judgment`, where plus symbols for the approved predictions almost draw a line on the y-axis. When the model makes many similar predictions (same $C_{occlusion}$ for different cases), such a line is formed in the plot. The length of this line reflects the difference in the explainability score ($C_{baseline} - C_{occlusion}$) for these predictions. Here the impact of`Opposes Judgment` is much less significant than in the German dataset with only a few clusters in dark green mostly consisting of approved predictions. In addition, we see more approved decisions in this plot overall than in the German dataset, with the prediction equal 1 (plus symbols) being much more frequent. Interestingly enough, we also again have the case that the language plots are very similar to each other.

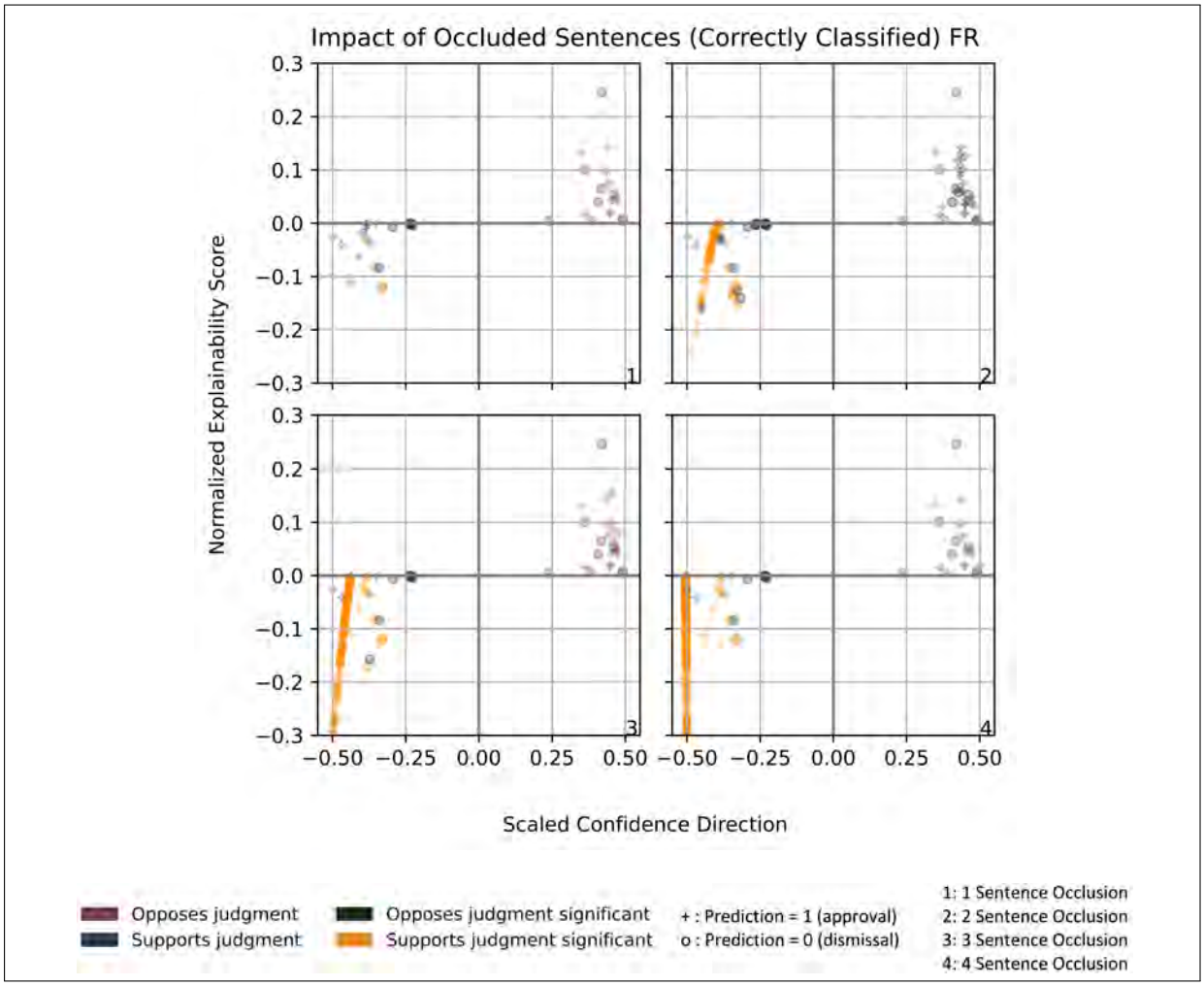


Figure 7: This plot shows the impact each correctly classified sentence has in French on the prediction.

Table 13 shows the distribution of the incorrectly classified sentences. Note that we have neutral sentences in this setup because these neutral sentences were incorrectly classified by the model as either supporting or opposing the judgment. We observe that there is an overwhelming amount of incorrectly classified neutral sentences. This high ratio in this category is probably due to two reasons: First, there were overall much more neutral sentences, to begin with, so these sentences made up the bulk of the occlusion experiments. Secondly, neutral sentences could be wrongly categorized as either `Supports Judgment` or `OpposeJudgment` while the other two categories could only be wrongly categorized as their opposite because we removed the model's classification of neutral sentences. Interestingly, the `OpposeJudgment` category stays the almost same as with the correct classification seen in Table 12. This confirms that the distribution of the classified sentences probably represents the different amount of occurrences these categories have in the dataset rather than a weakness of the model's classification ability in a specific category.

6.2.2 Impact of incorrectly Classified Sentences

To begin understanding the false classifications, we want to start by explaining what we are looking for in Figures 8, 9 and 10, because it is different than with the correctly categorized sentences above. With these visualizations, the desired result would be clustering around both the y and x axes. In German and Italian we see a resemblance of this desired clustering. However, we can also observe that especially `Neutral` sentences are distributed all over the plot in all three datasets. This suggests that these sentences, even though from a legal perspective unimportant, still have quite a significant positive and negative impact on the model's predictions. This effect is especially pronounced in the French dataset, where the clustering occurs at the axes limit like in the correctly classified sentences. As for the predictions we see a lot more dismissed predictions in all three of these plots suggesting that in those cases the model may have more difficulties identifying legally important sentences from unimportant ones.

Language	Supports Judgement	Opposes Judgement	Neutral
1 Sentence Occlusion			
German	14,02%	7,55%	62,26%
French	20,64%	4,98%	58,01%
Italian	10,11%	7,49%	70,79%
2 Sentence Occlusion			
German	7,38%	3,45%	76,50%
French	18,83%	1,86%	65,87%
Italian	5,59%	3,05%	85,68%
3 Sentence Occlusion			
German	2,97%	1,33%	85,91%
French	20,05%	0,64%	64,47%
Italian	2,29%	0,91%	94,85%
4 Sentence Occlusion			
German	1,24%	0,60%	90,06%
French	28,02%	0,39%	55,28%
Italian	0,82%	0,35%	98,21%

Table 13: Distribution of the incorrectly Classified Sentence chunks for each language and experiment

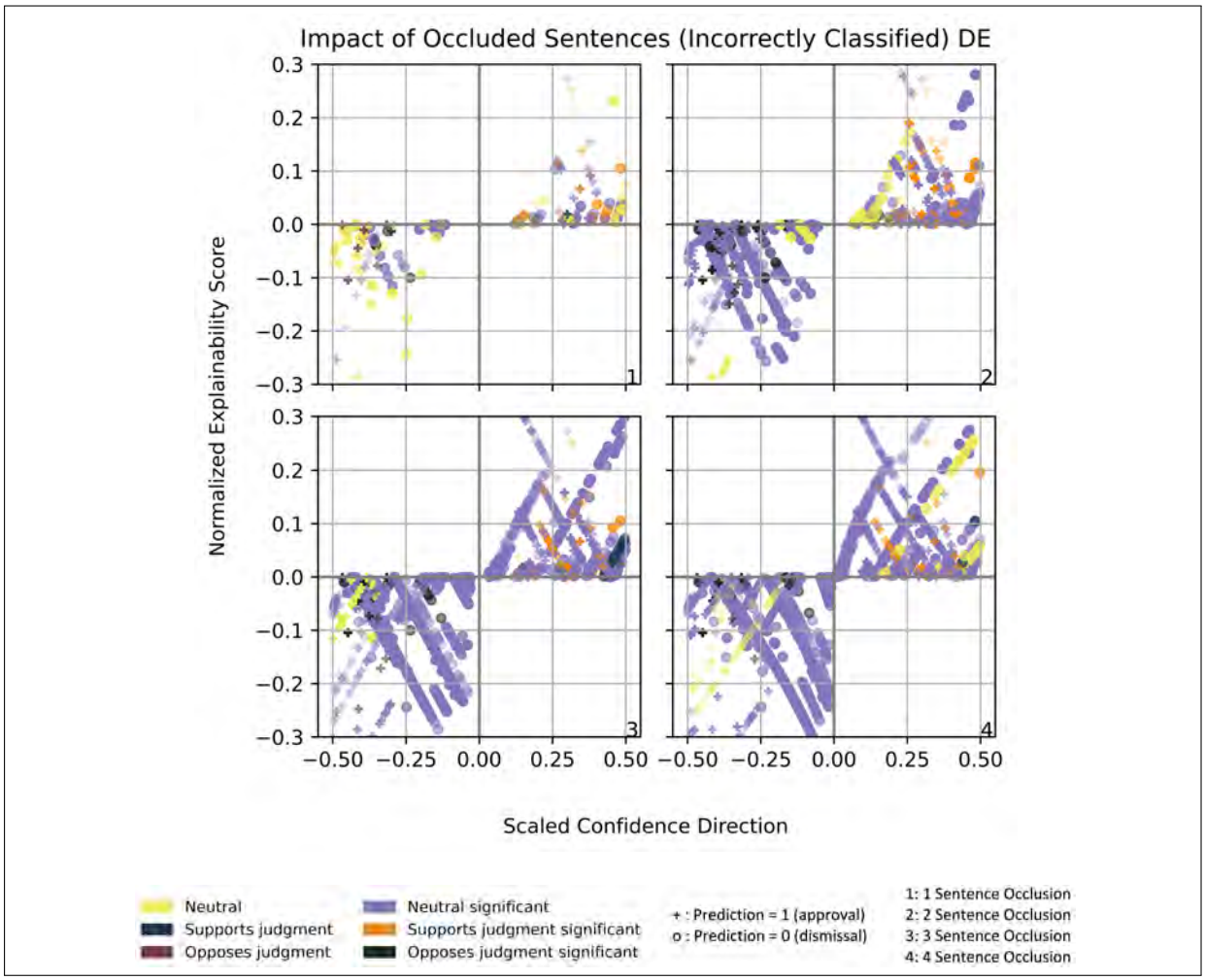


Figure 8: This plot shows the impact each incorrectly classified sentence in German has on the prediction. The further away a point is from the null axis the more impact it has on the model's prediction. The different markers indicate the prediction and the number on the bottom left is the experiment number. Note that with this false classification the clustering should occur on around both axes.

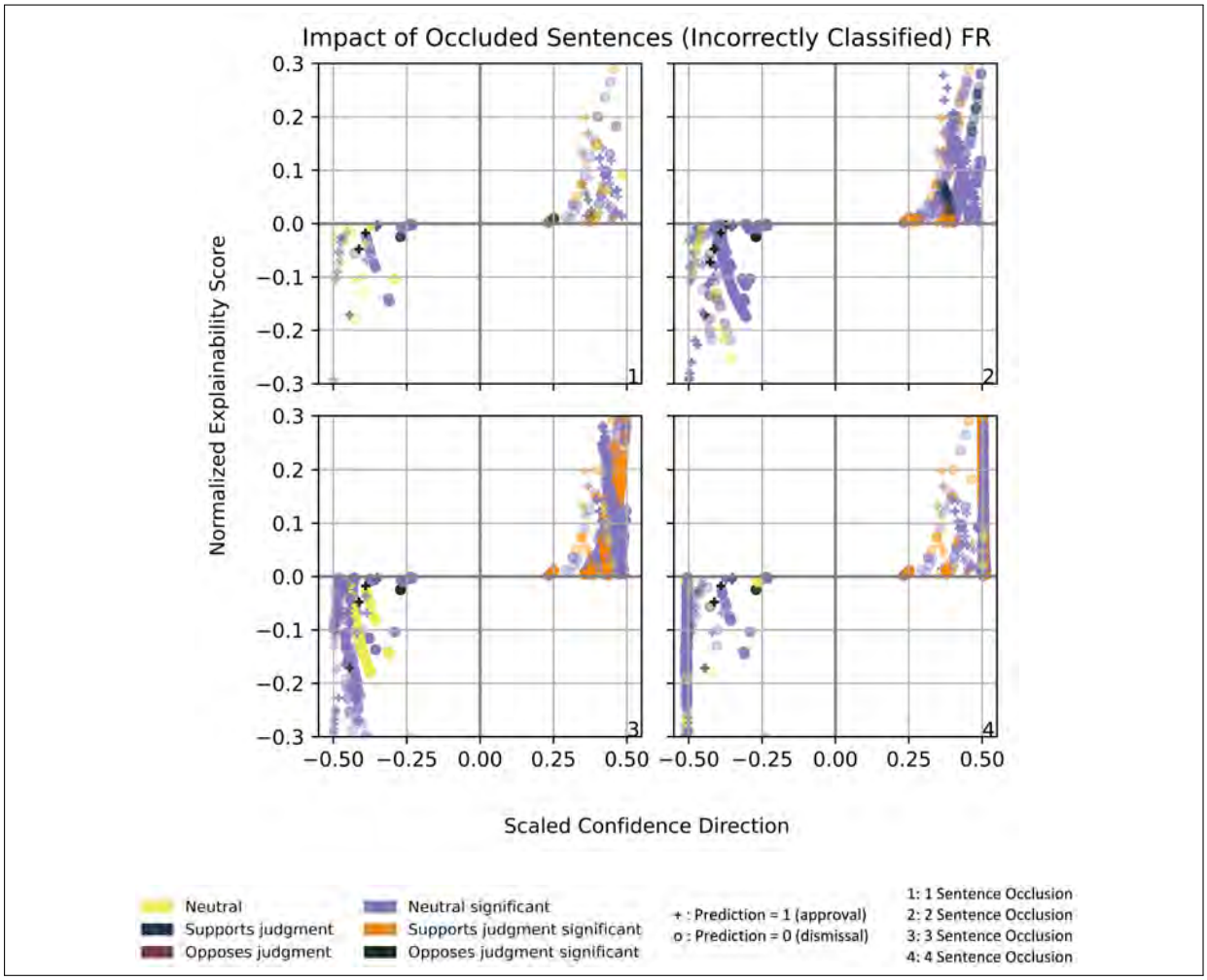


Figure 9: This plot shows the impact each correctly classified sentence in French has on the prediction.

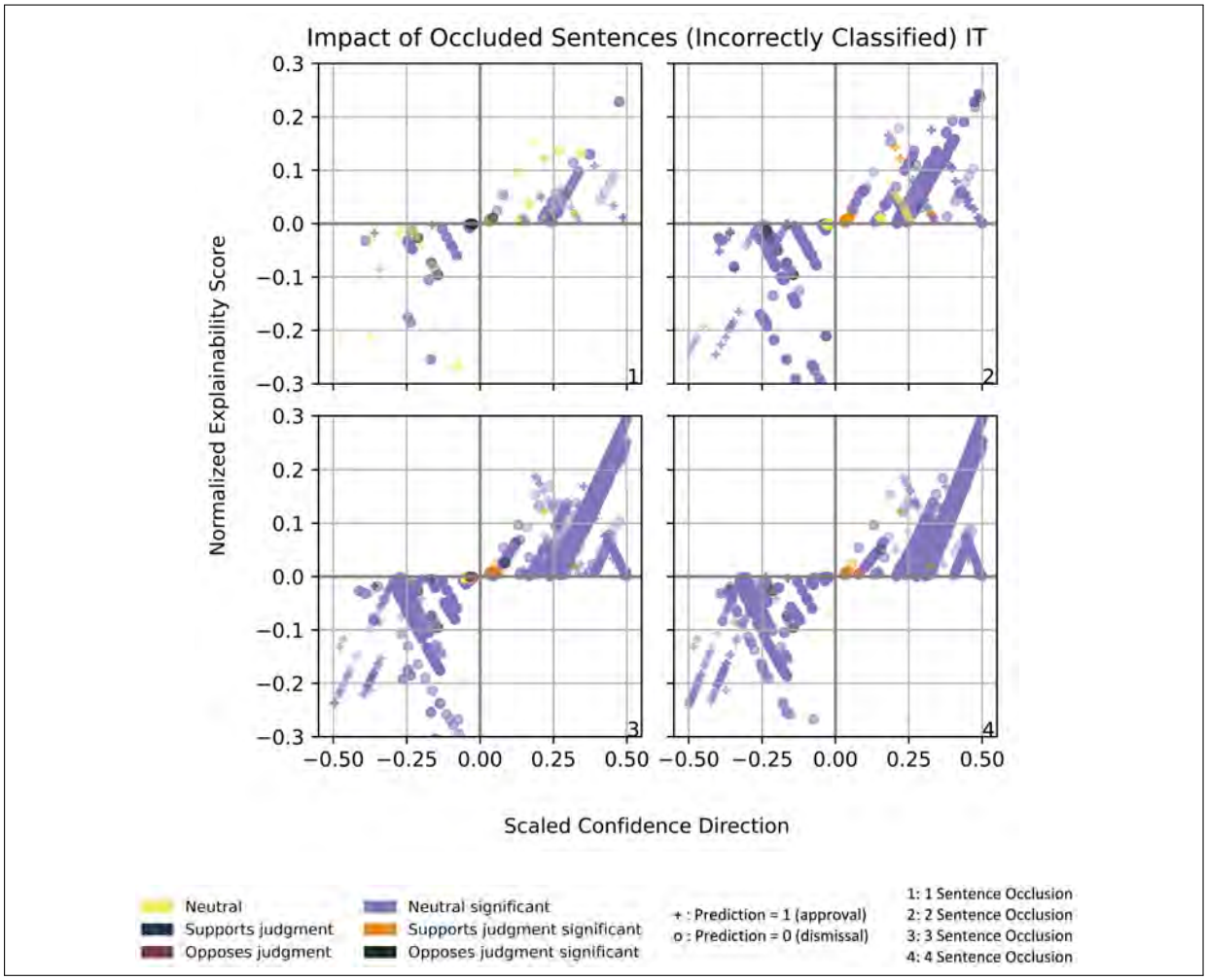


Figure 10: This plot shows the impact each correctly classified sentence in Italian has on the prediction.

This initial analysis reveals mixed results regarding the model's understanding of the semantics and context of the facts section. The correct classifications suggest that when the model correctly classifies a sentence, it is also quite influential in its decision-making process. This may indicate that BERT has a semantic understanding of the importance of these sentences and is able to make predictions based on legal reasoning in some cases. However, the wide range of classifications for neutral sentences indicates that these pose a significant challenge for the model. To further investigate the potential trends and impact of the explainability labels, we will present a further aggregation of these results in the next section

6.2.3 Trends Explainability Labels

We can examine the overall trends for each explainability label using Figures 11, 12,13. Note that with this visualization we again ignored sentences classified by the model as `Neutral`, only showing the trends for sentences the model labeled with the other two labels. For these plots, we have chosen a different scale and enlarged the relevant quadrant. For `Supports Judgment` this means an enlargement of the lower left part of the plot and of the upper right part of the plot for `Opposes Judgment`. First, we observe that for the three languages we have quite different plots. The German plots Figure 11 show the most clustering on the entire grid surface, French (Figure 12) again shows line patterns on the axes and in Italian we a mix of clusters and diagonal line patterns.

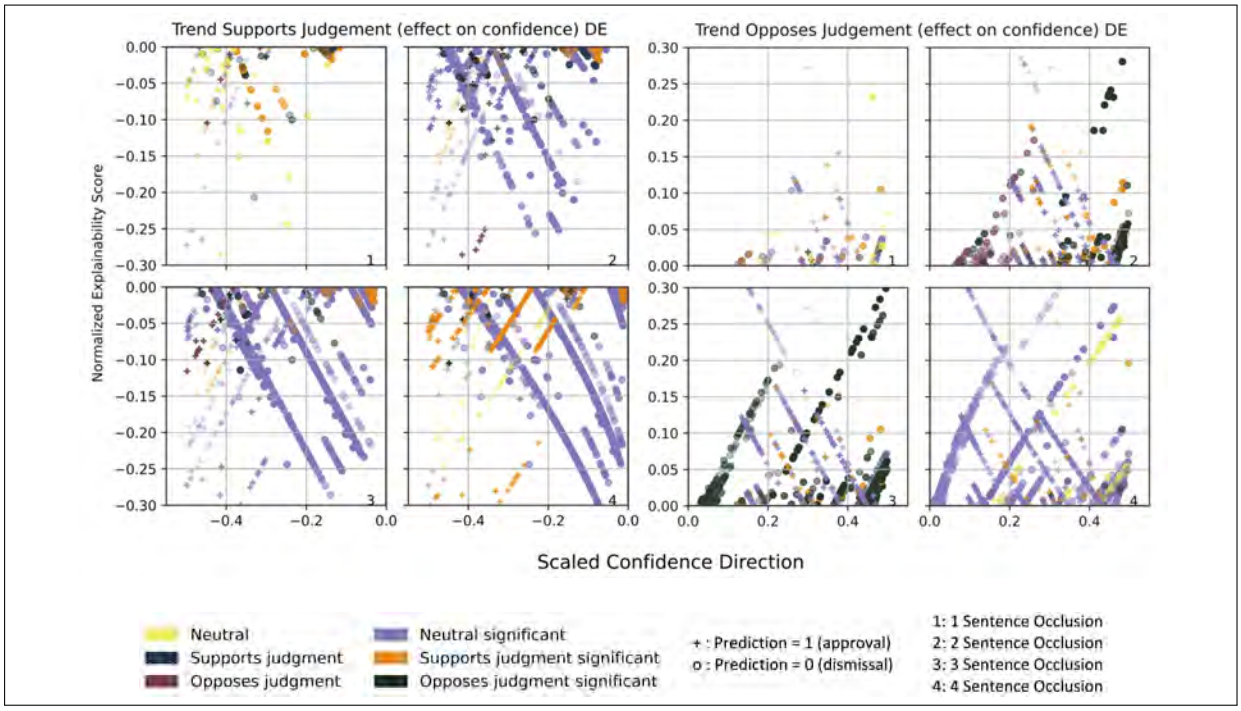


Figure 11: This plot shows the trend of the`Supports Judgment` and`Opposes Judgment` classification by the model for German cases. Note that this plot contains both correct and incorrect classifications. The actual human classification is indicated by the marker color.

These plots illustrate that the model appears to have a high level of confidence in correctly classifying `Opposes Judgment` in all three languages, as indicated by the dense clusters of dark green markers in the right subplots of Figures 11, 12, and 13. In contrast, the impact of false `Opposes Judgment` classifications is minimal, as shown by the smaller number of dark green markers in the left subplots. In terms of the green clusters, we see that in German, the majority of predictions are dismissals, in French the majority are approvals, and in Italian, both types of predictions are equally represented. Examining the occurrence of `Supports Judgment` we can see that in German both correctly and incorrectly classified sentences have only a small amount of impact (see orange clusters around the zero mark and at the edges of the of the Figure 11), with the exception being some approvals the fourth left subplot. French on the other hand shows a much greater effect concerning `Supports Judgment` with orange lines at the edges of almost all the subplots. In Italian, the models seem to be the strongest in clearly identifying the correct sentences as supporting judgment (especially for approvals) while having almost no incorrectly classified sentences in this category.

One weakness that is apparent in all languages is the difficulty in correctly classifying `Neutral` sentences. This is particularly evident in the `Supports Judgment` plots (left subplots), where the effects are particularly pronounced. This supports the finding from the previous section that the model tends to falsely view legally unimportant sentences as important, particularly as supporting the judgment.

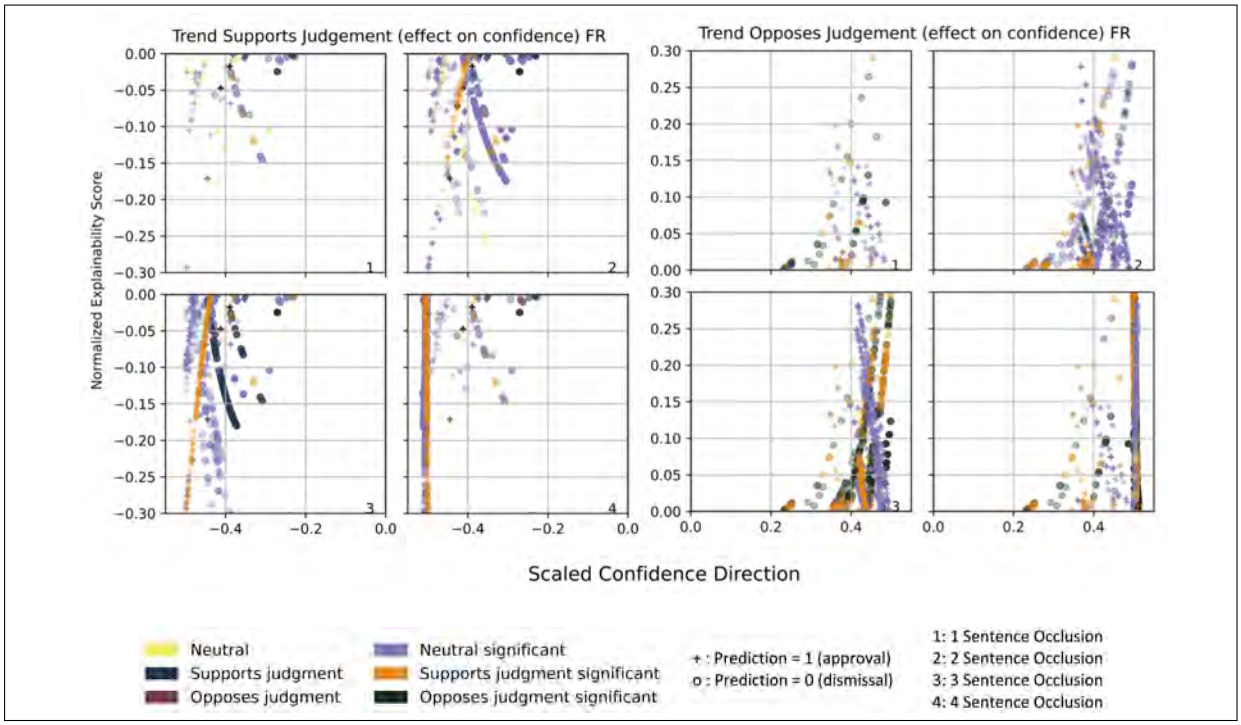


Figure 12: This plot shows the trend of the`Supports Judgment` and`Opposes Judgment` classification by the model for French cases.

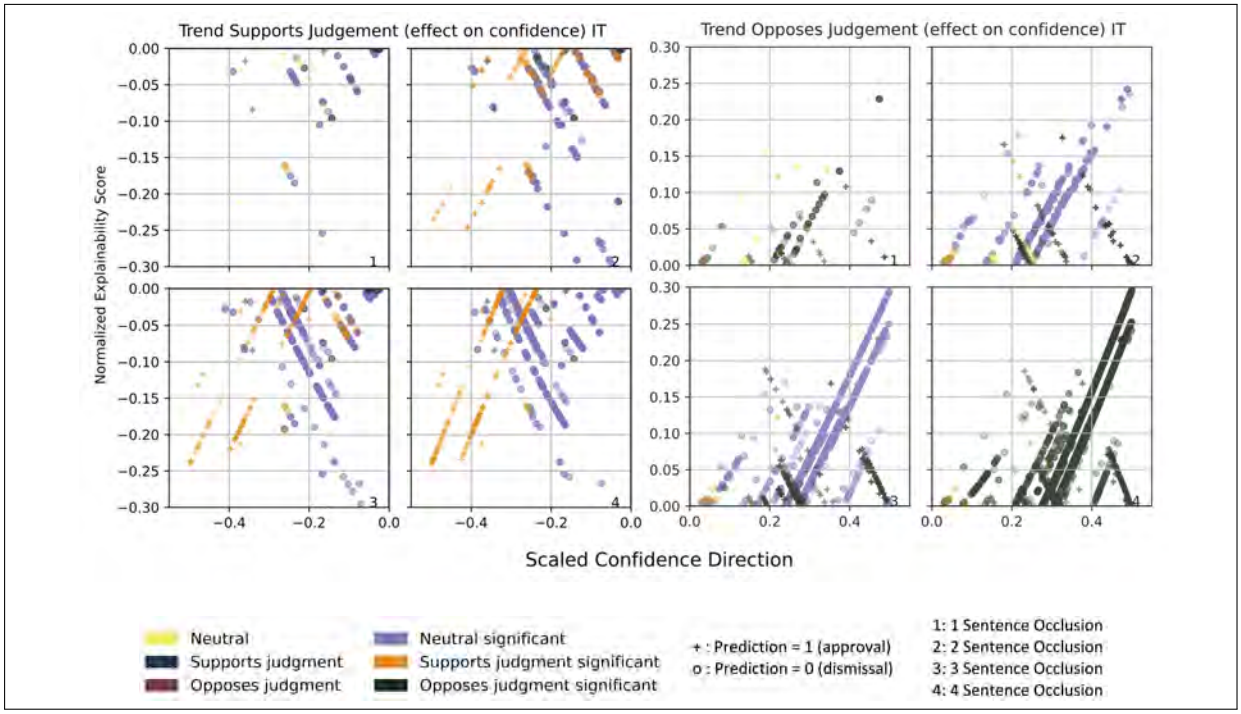


Figure 13: This plot shows the trend of the`Supports Judgment` and`Opposes Judgment` classification by the model for Italian cases.

6.2.4 Inter Annotator Agreement between Model and Legal Expert

The results above analyze the model's classification ability and show the impact the occluded sentences had on the prediction estimates. In this section, we further investigate the contents of the falsely classified sentences to determine if the model understands or has learned the legal finesses contained in a fact section. We present

the calculated the IAA of the incorrectly classified sentences with the gold-standard annotations. This enables us to investigate if incorrectly classified sentences may still be similar enough to the human annotations to suggest that the model understands the semantics of these sentences. To ensure comparability we apply the same IAA scores as with the annotations and calculate the IAA of each incorrectly classified sentence, with all the human-annotated sentences with this respective label in the same case. Table 14 shows the mean IAA for each language and score. Note that this mean consists of the average of both explainability labels and all occlusion experiments.

We see a rather similar agreement over all the languages with most of the scores ranging between 0.04 and 0.15. This indicates a very low agreement between the incorrectly classified sentences and the annotated sentences. Interestingly enough with `BERTScore`, we achieve an overall agreement of over 0.6, which indicates medium agreement. This is probably due to the fact `BERTScore` is one of the more sophisticated and newer IAA metrics and is able to determine if sentences are semantically equivalent to each other. The Violin Plot in Figure 14 visualizes the IAA using `BERTScore` as seen in the annotation analysis.

IAA Score	Mean German	Mean French	Mean Italian
bert_score	0.653	0.682	0.661
bleu_score	0.16	0.192	0.181
jaccard_similarity	0.084	0.128	0.086
meteor_score	0.094	0.135	0.087
overlap_maximum	0.041	0.041	0.037
overlap_minimum	0.127	0.111	0.118
rouge1	0.127	0.234	0.155
rouge2	0.033	0.061	0.024
rougeL	0.094	0.156	0.106

Table 14: Mean Inter Annotator Agreement between human and model for each language

Figure 14 shows that the agreement is similar across languages, experiments, and labels, with mean and median values around 0.65. In German and French, we see that `Opposes Judgment` sentences have a higher agreement than `Supports Judgment`, with values ranging from around 0.5 to 0.9. These values are mostly concentrated in the middle and upper parts of the violin plots. Italian exhibits slightly lower agreement in this category, with rounder plots due to a smaller range of values and a concentration in the middle (as shown by the blue violins in the Italian subplots in Figure 14). For `Supports Judgment`, we see a similar range of values in German and Italian across all languages and experiments, but with some more pronounced lower parts in the plots. Except for Italian which still has very round plots that are similar to their `Opposes Judgment` counterparts.

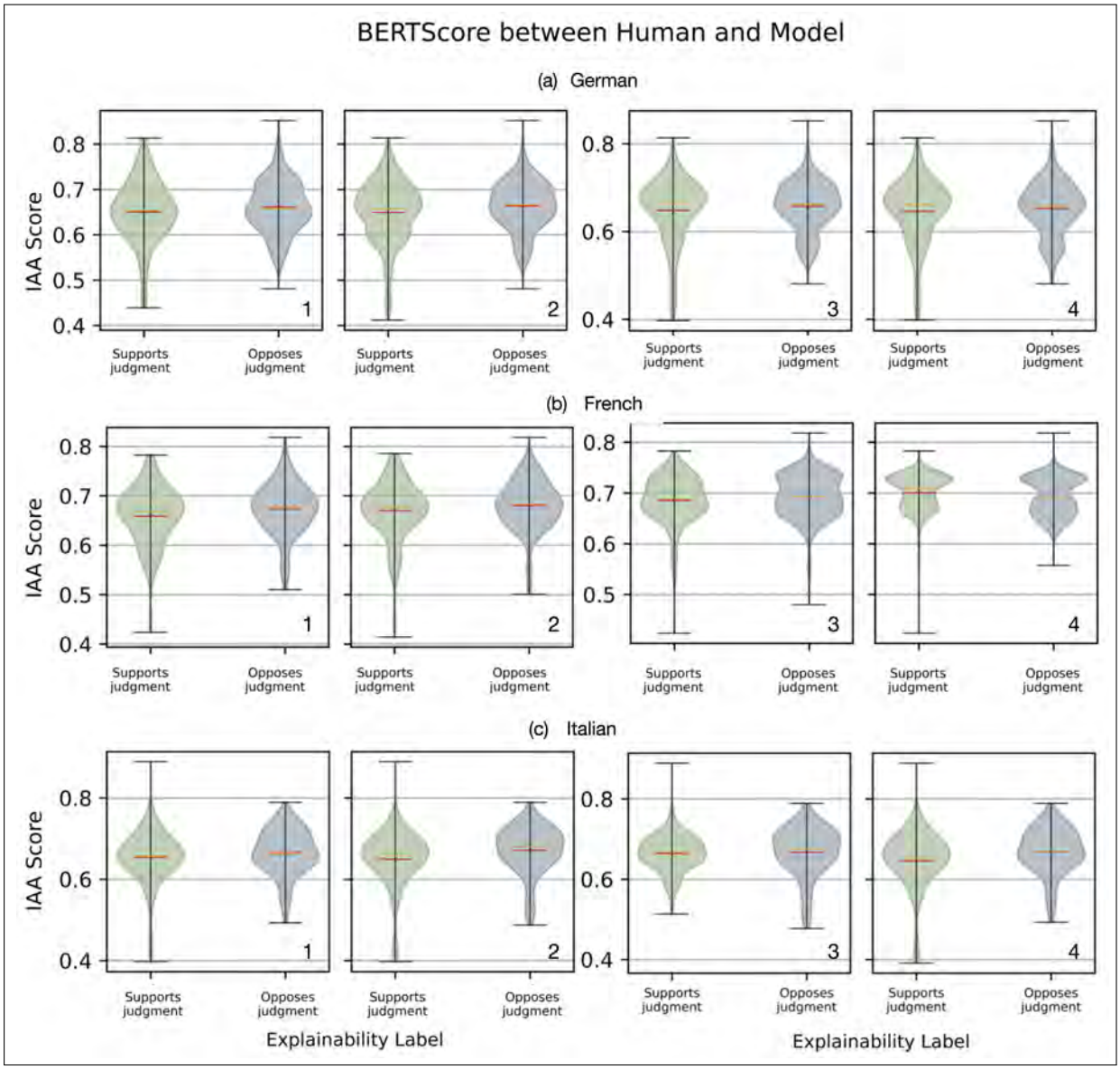


Figure 14: These violin plots show the results of the `BERTScore`. The red dash indicated the mean and the orange dash indicates the median. The number in the bottom corner indicates the occlusion experiment

The analysis of the IAA results shows mixed results. While most of the scores have a low agreement, the `BERTScore` metric exhibits higher agreement with values around 0.65, suggesting that the model may understand the semantics of the incorrectly classified sentences. This suggests that the model may not simply be making random classifications as legally important, but rather has some understanding of the content of these sentences. Overall, our results indicate that the model still has room for improvement in accurately identifying the importance of sentences and making accurate predictions. They also suggest that further investigation into different IAA metrics may be necessary, as they may not all be suitable for comparing human and model annotations. It is worth noting that the low agreement across most of the IAA scores may be due to the difficulty of the classification task, as well as the potential for label skewness in the dataset.

6.2.5 Explanation Accuracy

In Section 3.7.5 we introduced the explainability accuracy score aS_{exp}. This scoring system gives us the possibility of producing explanations with two flavors: a model-near explanation and a human-near explanation. For each case in the occlusion test sets, we chose 4 sentences as an explanation. Two model-near and two human-near explanations. Each of these two explanations consists of one sentence supporting the judgment and one opposing the judgment. The two flavors differentiate the filtering method applied to choosing these sentences. For the closest model explanation, we chose the sentences with the highest explainability score (for

Supports Judgment the highest negative score, and for Opposes Judgment the highest positive score). For the closest human explanation, we chose sentences with the highest IAA calculated using the BERTScore. To clarify using the explanation accuracy score aS_{exp} and the confidence direction $conf_{dir}$ we were able to filter explanations in two flavors and apply a score to them which classifies the legal accuracy of these explanations.

Note that the explanation length increases in the experiments with larger chunks (occlusion of 2,3,4 sentences), but the scoring system stayed the same. Each of these two sentences or two chunk explanations was assigned an explanation accuracy score giving them a value between 2 and 0.

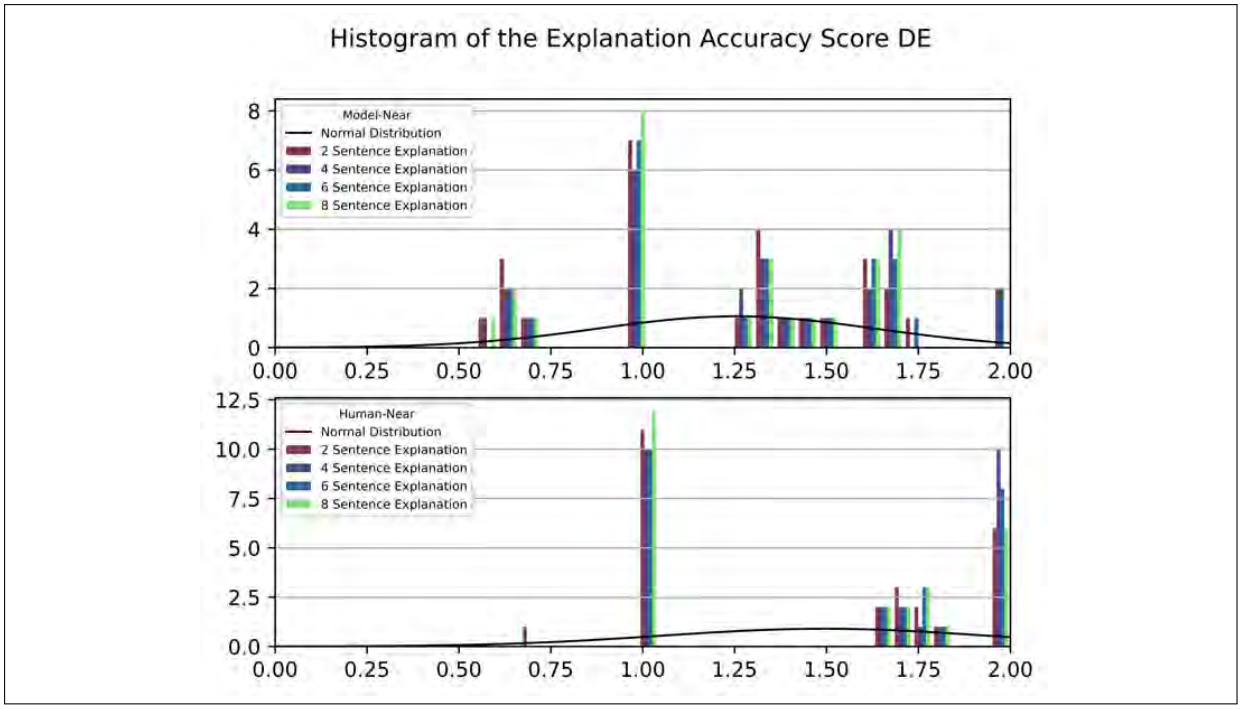


Figure 15: Distribution of the explanation accuracy score aS_{exp} for model-near and human-near explanations in German. The black plot indicates the normal distribution.

Figures 15,16, and 17 show the distribution of the explanation accuracy scores for the three languages. The human-near distributions (bottom histograms) are similar across all three languages, with peaks at 1[15], 1.75, and 2.0. One difference is that the German and French human-near explanations have a higher accumulation at the maximum value (2) than Italian. This is also true for the German and French model-near explanations, even when less pronounced. Looking at the model-near explanations the French subset has the worst explanation quality, with some values accumulating below the 1 threshold. Overall, these distributions suggest that both the model-near and human-near explanations are of mostly high quality (> 1). When analyzing the data more closely we also found that they often consist of the exact same sentences. Indicating that these two flavors are not as different as one would imagine. We also did not see any higher distribution when aggregating these results across prediction, legal areas, years or explainability labels, with the mean values always in a similar distribution than seen in Figures 15,16, and 17, with no distinct patterns apparent.

[15]Note that values accumulating at 1 often indicate that there was only one explainability label present in the case, allowing for an explanation containing only a sentence with that label (e.g. a sentence supporting the judgment).

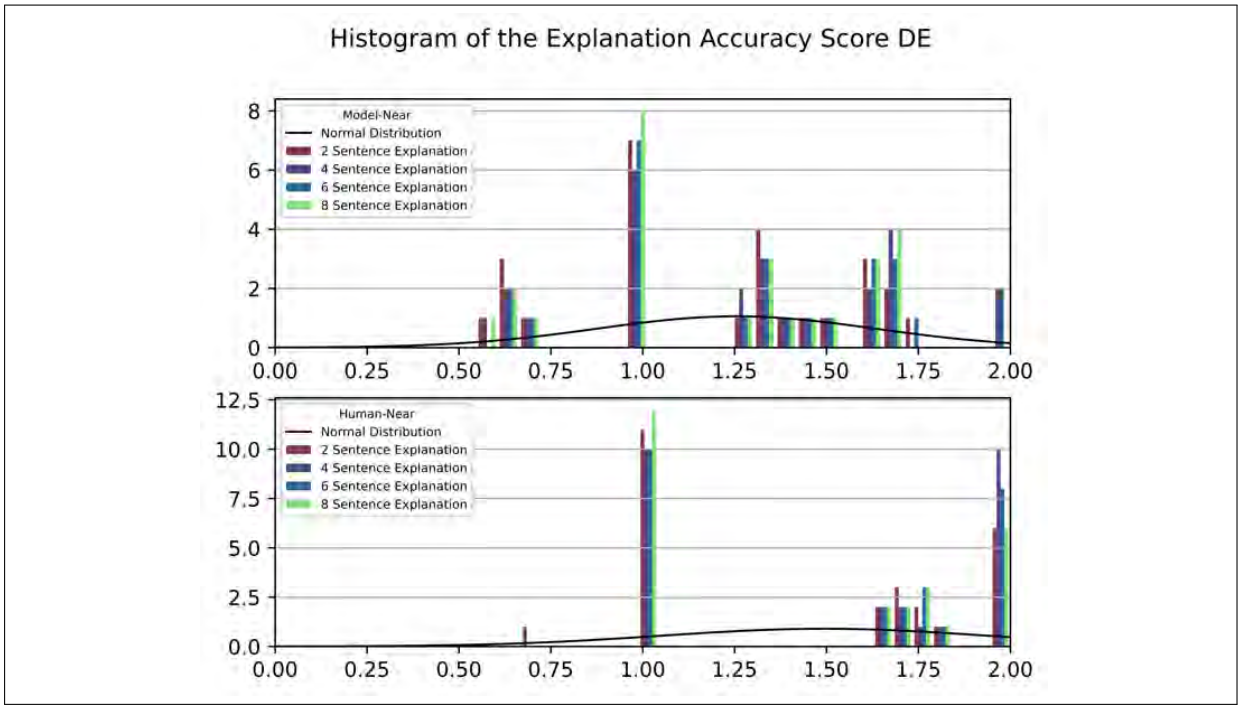


Figure 16: Distribution of the explanation accuracy score aS_{exp} for model-near and human-near explanations in German. The black plot indicates the normal distribution.

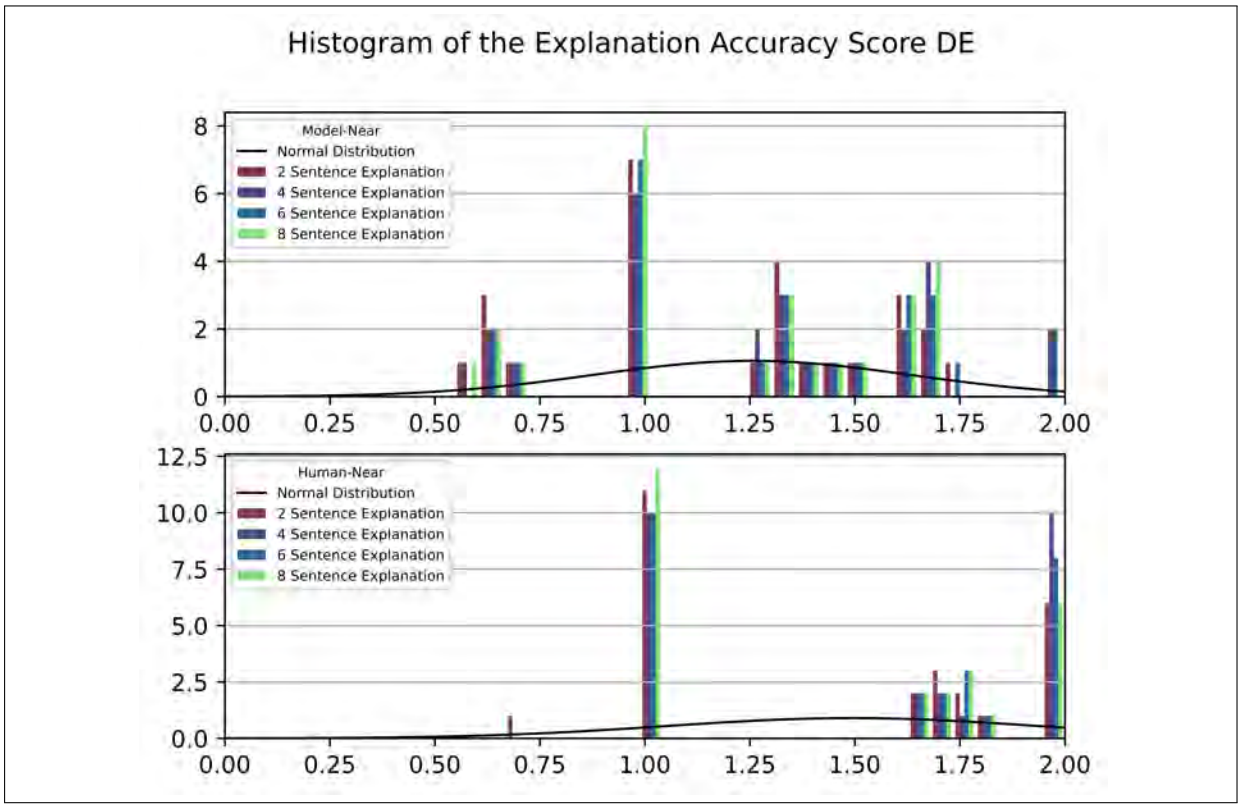


Figure 17: Distribution of the explanation accuracy score aS_{exp} for model-near and human-near explanations in German. The black plot indicates the normal distribution.

6.3 Main Results – Lower Court Insertion

In this section, we present the results of our investigation into potential bias within different lower courts using the LCI method. As previously discussed, we used the Macro-F1 measure to evaluate the performance of the LCI. In addition, we conducted a detailed study in each language to examine the impact of inserting a particular lower court on the model's prediction.

The results, as shown in Table 15, indicate that the performance of the model is similar between the German and French subsets, which have 35K and 21K training samples, respectively. However, the performance is weaker in the Italian subset, which has only 3K training samples. In German and French, the LCI performs better than the original SJP. In Italian, the SJP outperforms the LCI. These results suggest that the model is robust to small changes in the fact section and still is mostly able to make a correct prediction.

Model	de		fr		it	
	Macro-F1 (SJP)	Macro-F1 (LCI)	Macro-F1 (SJP)	Macro-F1 (LCI)	Macro-F1 (SJP)	Macro-F1 (LCI)
hierarchical (two-tier 4× 512 tokens)						
Native BERT	68.5 ± 1.6	**85.9**	70.2 ± 1.1	**88.9**	**57.1 ± 6.1**	53.7

Table 15: Comparison between the results from the SJP and the LCI experiments using the Macro-F1. The models were all trained and tested in the same language. "Native BERT" refers to the BERT model that was pre-trained in that language. The best scores for each language are highlighted in bold. In the SJP results, the standard deviation between different seeds is shown, but in the LCI results, the model was only run using one random seed, so this information is not necessary.

Regarding the prediction distribution for these experiments, we can see in Table 16 that especially the German and French datasets have quite an equal distribution. Concerning the effect a lower court has on the distribution, as explained in Section 3.6, we use the explainability score S_{exp} to analyze the changes in the predictions from the LCI. Figures 18, 19, and 20 show the influence each lower court has on the prediction. To better visualize this influence we separate negative S_{exp} from positive S_{exp} and calculate the mean for each direction. In other words, each result for one specific lower court is split into a set of positive influences and negative influences on the prediction. We also highlight the respective prediction with 0 (dismissal) having no hatch pattern and 1 (approval) having one. We also apply the t-test to each negative and positive mean comparing it to a hypothesized mean of 0, since theoretically if there is no bias from one lower court to another the explainability score should be 0 (no changes in prediction confidence). The results from the t-test are then aggregated with the mean explainability score giving us an indication if the mean is significantly different from the population mean (see darker parts in Figures 18, 19, 20). We also examine which court actually managed to flip a model's prediction from the baseline (Tables 17,18 and 19 and Figures 24, 25 and 26 in the appendix) and try to find a correlation between these two results. It is important to mention that we talk about very small changes in confidence below 0.1, but the LCI was also a very small disruption of the facts section, with a mean of around 7 tokens. We will now describe these results for each individual language starting with German.

Language	**Approved Baseline**	**Dismissed Baseline**	**Approved LCI**	**Dismissed LCI**
German	44.45%	55.55%	43.59%	56.41%
French	47.83%	52.17%	52.94%	47.08%
Italian	21.74%	78.26%	16.67%	83.34%

Table 16: Prediction distribution in LCI

6.3.1 German results

For the German lower courts we can observe that the administrative court Berne (Be_VGer), the high court of Aargau (AG_OGer), the appeals court of Basel-Stadt (BS_AppGer), and the administrative court from canton Schwyz (SZ_VGer) do not significantly influence the model's confidence. When we look at the plot (see Figure 18) in general we can see an all-around against approval and toward dismissal trend echoing the slightly skewed

prediction distribution seen in Table 16. The same is true for the flipped decision since there are more decision flips from approval to dismissal, indicating a pro-dismissal[16] trend.

Looking at courts with significant influence (darker parts in the plot): Inserting BE_Oger has almost the same influence in both directions, with a bit more positive influence for approved decisions. The high court of Zurich has the smallest influence in both directions. With both these courts this trend is also true regarding the flipped decisions with Berne having more than double predictions flipped from 0 to 1 (positive influence for approval) and Zurich having no flipped decisions at all (not influential).

The courts with the most consistent trend are courts from the inner part of Switzerland from the cantons Lucerne, Schwyz and Glarus (GL_VGer, Lu_KGer and SZ_KGer), from the northern part of Switzerland like the Administrative Court of Aargau and the cantonal courts of Basel-Landschaft (BL_KGer) and Zurich (ZU_KGer) and the Cantonal Court of Appenzell Ausserrhoden (AR_KGer). By a consistent trend, we mean that this court shows opposite directions for each prediction, resulting in a “Z" shape in the plot rather than a stacked pyramid. This is true for all these courts even if in different intensities. For the courts from Aargau, Appenzell Ausserrhoden, Basel-Landschaft, Glarus, and Zurich this trend is against approval or pro-dismissal. Consequently, we see as mentioned before an overall trend toward pro-dismissal (except for Lucerne). This trend is mostly reproducible in the flipped decisions where most of these courts have more flipped predictions from 1 to 0. An exception would be for example BL_KGer's only managed to flip predictions from 1 to 0 and AR_KGer's flipped more decisions from 1 to 0.

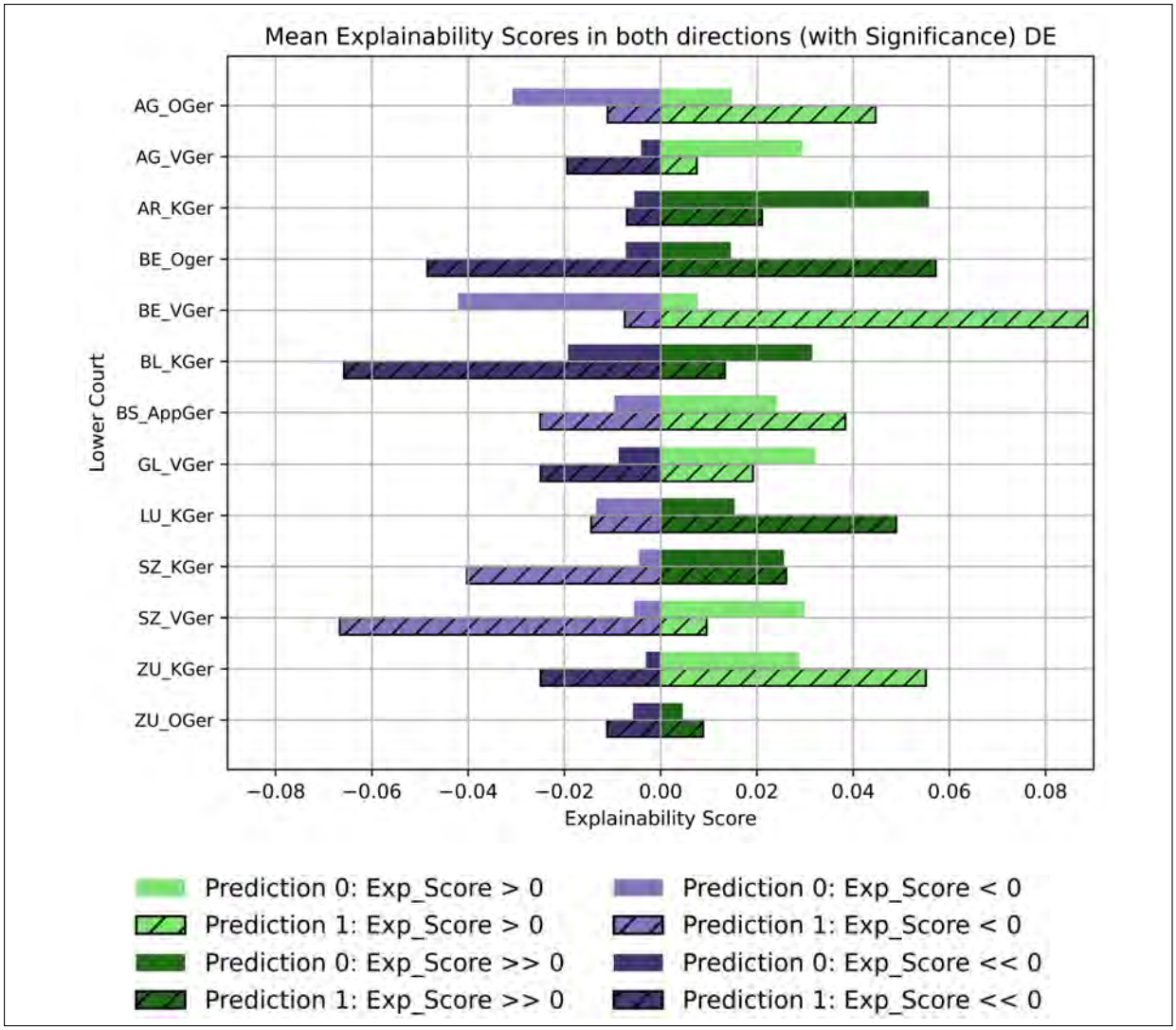


Figure 18: Effect on the confidence of each German lower court via two-sided explainability score. The darker parts indicate a significant difference from the baseline mean explainability score. The hatched bars show the approved predictions.

[16]The term 'pro-dismissal' refers to a situation where the outcome of a lower court case favors dismissal or is unfavorable for approval.

Looking now at the positive influence on the prediction we can see that the already mentioned courts SZ_KGer and LU_KGer have a significant positive influence on the model's decision while having little negative influence. This effect is especially pronounced for the prediction approved. When looking again at the flipped predictions this pro-approval[17] trend is not perfectly reflected. For example, the cantonal court of Lucerne only flipped decisions opposite to its trend (from 1 to 0). This suggests that these two metrics may not always be perfectly correlated. The most influential courts in the negative direction, are the AG_VGer, the BL_KGer, the GL_VGer (both administrative courts), and ZU_KGer. These courts show especially less confidence when looking at approvals while having no significant influence in the positive direction. This negative trend can again also be found in the flipped decision, where the trend for these courts except for AG_VGer (flips both decisions equally) is flipping predictions from approval to dismissal.

In conclusion, when looking at the German lower courts, there appear to be some trends for the different courts which correlate with the prediction distribution and the flipped predictions. Courts from cantons Schwyz, Lucerne, and Appenzell Ausserrhoden mostly have a significant positive influence on the model's decision, while the most influential courts in the negative direction are the AG_VGer, the BL_KGer, the GL_VGer, and ZU_KGer. In general, there is a strong trend against approval with a mixed positive trend (pro-approval as well as pro-dismissal).

Lower Courts	Total Nr. of Cases	Flipped Cases 1 → 0	Flipped Cases 0→ 1	Total Nr. Flipped Cases
AG_OGer	22	4,55%	0,00%	4,55%
AG_VGer	22	4,55%	4,55%	9,09%
AR_KGer	23	8,70%	4,35%	13,04%
BE_Oger	36	2,78%	5,56%	8,33%
BE_VGer	22	0,00%	0,00%	0,00%
BL_KGer	24	4,17%	0,00%	4,17%
BS_AppGer	21	0,00%	4,76%	4,76%
GL_VGer	22	4,55%	4,55%	9,09%
LU_KGer	23	4,35%	0,00%	4,35%
SZ_KGer	22	4,55%	4,55%	9,09%
SZ_VGer	22	4,55%	4,55%	9,09%
ZU_KGer	23	8,70%	4,35%	13,04%
ZU_OGer	20	0,00%	0,00%	0,00%

Table 17: Distribution of Flipped Cases in the German LCI. Note that most of the courts showing no flipped prediction also show a low influence on the model's prediction.

6.3.2 French results

In this section, we will focus on the results of the French lower courts illustrated in Figure 19. These courts are less geographically dispersed in Switzerland and are mostly located in the western or middle part of Switzerland. There are also fewer cantons as in the German dataset. First, we can ignore the social insurance court Vaud (VD_CHCTA), since it has no significant influence. Secondly, we can observe that there is an overwhelming amount of significant positive influence. In other words, the national administrative court (CH_BVGer), the civil appeal chamber of Fribourg, all the courts from Geneva, the civil court of Jura, and all but one court from the canton of Neuchatel and Vaud have a positive influence, with simultaneously little negative influence. This effect is especially pronounced for approved cases. This same pro-approved trend is also reflected in the prediction distribution, even when only very slightly leading more towards approval. Inspecting the flipped distribution using Table 18 we can first observe that the French cases flipped at a higher rate than cases from the other two datasets. With only one court having zero prediction flips. For example, for the three courts of Geneva (Ge_CJ, GE_CJR and GE_ChRPeCJ), the Court for the protection of children and adults and the Court of Public Law of Neuchatel (NE_CEA and NE_CPuTC), and for the Appeal and Insurance Court of Vaud (VD_CAPPe and VD_CASoTC), the positive influence on the approved decisions leads to a high amount of dismissed decisions being flipped (over 10%).

Looking at courts with consistent trends, most of the french lower courts show a more or less pronounced “Z" shape. With all but three of these “Z" shapes leaning pro-approval and against dismissal. The flipped prediction also mostly represents this trend even if not perfectly, particularly for less pronounced “Z" shape like in the VD_CASoTC (flips equal amount of decisions for both predictions). Examining the negative influence, we can observe that the Swiss National Supreme Court (CH_BGer), the Criminal Court of the Canton of Neuchâtel

[17] The term 'pro-approval' refers to a situation where the outcome of a lower court favors approval or is unfavorable for dismissal.

(NE_CPe), and the Civil Chamber of Valais (VS_ChCivTC) decrease the model's confidence, particularly for dismissed cases. This trend is also again somewhat similar for the flipped predictions, but the correlation is not perfect for example for the CH_BGer flips again an equal amount of both prediction 0 and 1.

Based on this first analysis it appears that the French lower courts have a mostly positive influence on the outcomes of cases, except for a few specific courts that have a negative influence. The positive influence of these courts is particularly pronounced in approved cases while the negative influence is pronounced in dismissed cases. Overall in French cases, the trends are very consistent with most courts showing a "Z" shape. The correlation with the flipped decisions is also quite high, but there are some courts where we rather see a kind of prediction swapping than flipping.

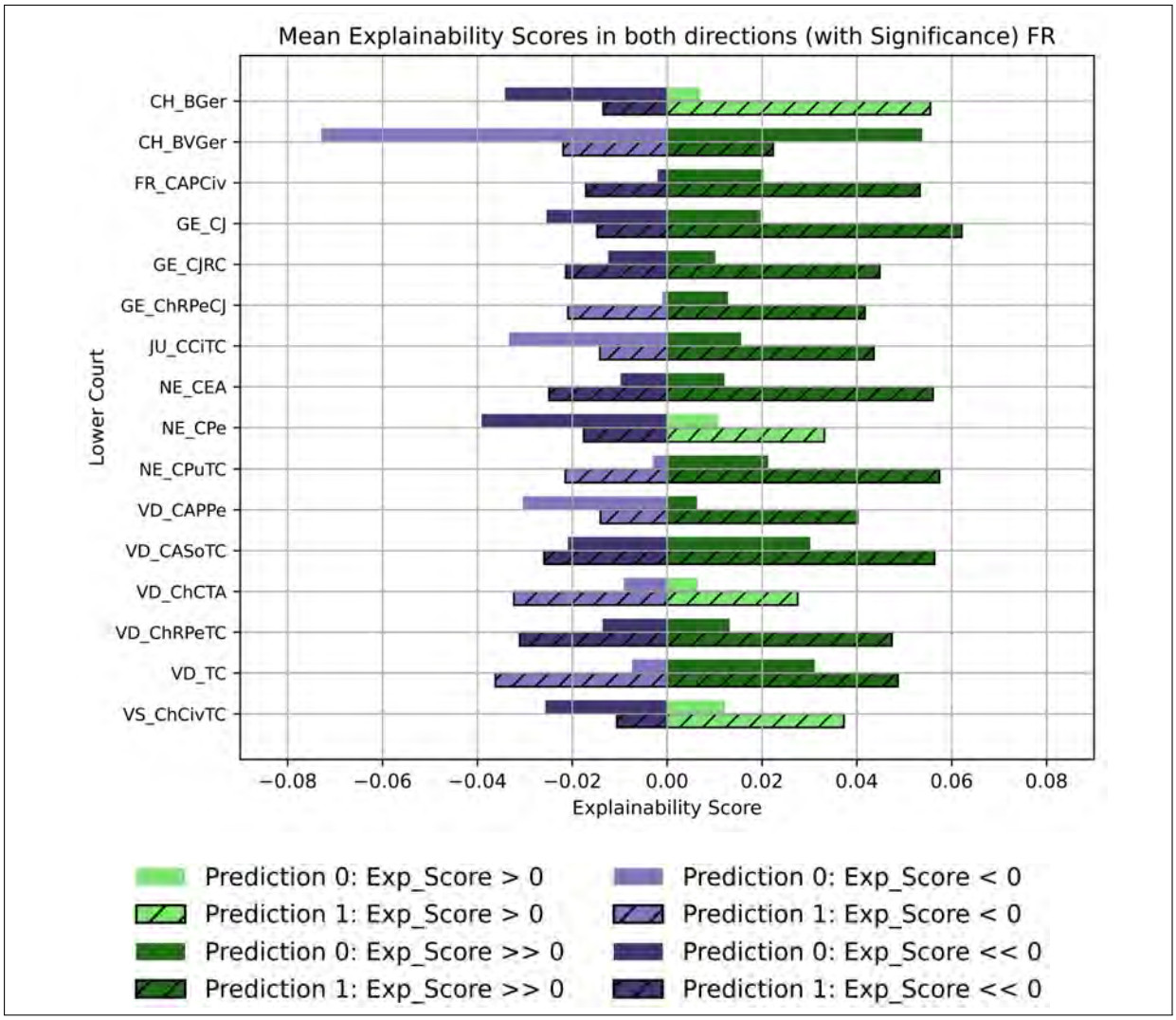


Figure 19: Effect on the confidence of each French lower court via two-sided explainability score. The darker parts indicate a significant difference from the baseline mean explainability score. The hatched bars show the approved predictions.

Lower Courts	Total Nr. of Cases	Flipped Cases 1 → 0	Flipped Cases 0→ 1	Total Nr. Flipped Cases
CH_BGer	19	5,26%	5,26%	10,53%
CH_BVGE	18	5,56%	11,11%	16,67%
FR_CAPCiv	18	0,00%	5,56%	5,56%
GE_CJ	29	0,00%	10,34%	10,34%
GE_CJRC	26	0,00%	15,38%	15,38%
GE_ChRPeCJ	19	0,00%	5,26%	5,26%
JU_CCiTC	20	0,00%	0,00%	0,00%
NE_CEA	17	0,00%	11,76%	11,76%
NE_CPe	18	5,56%	11,11%	16,67%
NE_CPuTC	18	5,56%	11,11%	16,67%
VD_CAPPe	17	0,00%	11,76%	11,76%
VD_CASoTC	19	10,53%	10,53%	21,05%
VD_ChCTA	19	0,00%	5,26%	5,26%
VD_ChRPeTC	19	0,00%	5,26%	5,26%
VD_TC	19	0,00%	0,00%	0,00%
VS_ChCivTC	17	0,00%	11,76%	11,76%

Table 18: Distribution of Flipped Cases in the French LCI

6.3.3 Italian results

Looking at the Italian Results (see Figure 20) firstly the Debt Enforcement and Bankruptcy Chamber of the Court of Appeal of the Canton of Ticino (TI_CEFTRAP) has no significant influence and hence can be ignored. Next, we can observe that almost all courts have a consistent trend. With the effect being more or less pronounced. For example, National Administrative Court (CH_VGer), the Civil Appeals Chamber (TI_CCivAP), the Protective Appeal Chamber (TI_CPTRAP), the Assurance Court (TI_TCAS) and the Appeals Court (TI_TRAP) exhibit a quite nice “Z" shape. The TI_CPTRAP and TI_CCivAP are more pro-approval and against dismissal respectively and the CH_BVGer, TI_TCAS, and TI_TRAP are more pro-dismissal. For the pro-dismissal courts, this trend is also echoed in the flipped decisions, with these courts having zero flips from dismissal to approval. For the pro-admissal courts, this is not the case since all the courts only managed to flip decisions from 1 to 0, meaning in the opposite direction of the trend.

Overall the Italian lower courts lean a bit more toward a negative trend against approval, which is represented gravely in the prediction distribution having only around 17% of approved predictions. It is also interesting to see that the effects are generally less pronounced for this dataset. The same is true for the flipped predictions, where most courts only managed to flip one decision and no decisions were flipped from dismissal to approval. The only two courts showing a largely positive influence are the aforementioned TI_TCAS (assurance court), the TI_TRAP (both strongly pro-dismissal), and again the TI_CPTRAP being the only court strongly pro-approval.

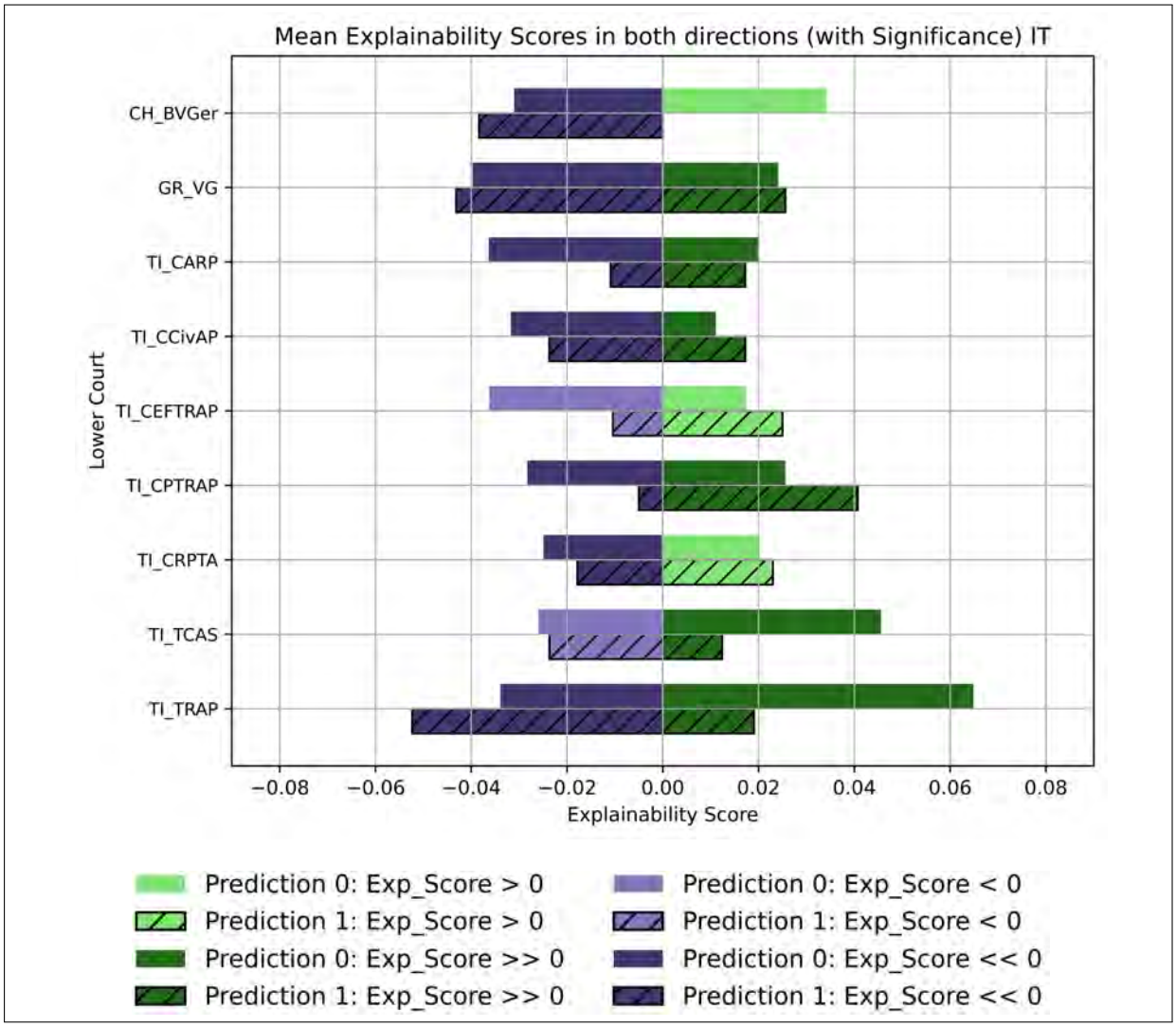


Figure 20: Effect on the confidence of each Italian lower court via two-sided explainability score. The darker parts indicate a significant difference from the baseline mean explainability score. The hatched bars show the approved predictions.

Lower Courts	Total Nr. of Cases	Flipped Cases 1 → 0	Flipped Cases 0→ 1	Total Nr. Flipped Cases
CH_BVGE	28	3,57%	0,00%	3,57%
GR_VG	27	3,70%	0,00%	3,70%
TI_CARP	63	4,76%	0,00%	4,76%
TI_CCivAP	83	4,82%	0,00%	4,82%
TI_CEFTRAP	27	3,70%	0,00%	3,70%
TI_CPTRAP	27	3,70%	0,00%	3,70%
TI_CRPTA	27	3,70%	0,00%	3,70%
TI_TCAS	23	0,00%	0,00%	0,00%
TI_TRAP	27	3,70%	0,00%	3,70%

Table 19: Distribution of Flipped Cases in the Italian LCI

7 Discussion

In this section, we present the findings of our occlusion analysis and provide examples of explanations generated by the model. We also examine a range of potential explanation approaches for interpreting the results of our LCI insertion study.

7.1 Explanaition from Occlusion

The IAA of the legal expert annotations is very high, especially when using `BERTScore` as a measure (see Section 6.1). This gives us confidence in using these annotations as our ground truth when discussing the results of the occlusion. Based on our impact analysis (see Section 6.2.3), it seems that the model has a somewhat mixed understanding of the semantics and context of the facts section. When the model correctly classifies a sentence, it tends to have a significant impact on its decision-making process, suggesting that BERT has some level of understanding of the relevance of these sentences and may be able to utilize legal reasoning in its predictions. Our trend analysis also showed that the model appears to be quite competent in correctly identifying sentences that oppose a judgment in all three languages. However, the model appears to struggle with classifying neutral sentences, which may be a challenge for it. This is particularly evident in the plots (Figures 11, 12, 13) for sentences that support a judgment, where the model tends to wrongly view legally unimportant sentences as important. This may be because neutral sentences in a legal context often mirror legally significant sections, using similar words, terms, and verbs but differing in their actual meaning. This notion of sentence similarity is also confirmed by the IAA results between model and human. Where we achieved medium to good agreement using `BERTScore`. We also illustrate this aspect in our explanation study below. In addition, we were able to quantify the quality of the gathered explanations using the explanation accuracy score aS_{exp}. The distribution of these scores across the experiments indicates that, in most cases, there is a good-quality explanation.

To deepen our understanding of the model even further we examine some example explanations. As a reminder the model-near explanations are chosen according to the explainability score, meaning that for `Opposes Judgment` the sentence with the maximum explainability score per case respectively the minimum for `Supports Judgment`. For a human-near explanation, we either chose sentences that were correctly classified by the model or showed a high agreement calculated using the best-performing `BERTScore`. The first few explanations concern the judgment of the Federal Supreme Court 6B 932/2019 of May 2020. This is a penal law case. For more context, we have translated the fact section and summarized it into the following key points.

- A._ (complainant) was convicted of defamation after failing to appear at the distinct court of Meilen in May 2013.
- In July 2013 A._ requested a new evaluation from the district court of Meilen, claiming that she had not been properly summoned for the main hearing.
- The district court rejected this request in September and the cantonal court of Zürich rejected the appeal in October 2013.
- A._ filed a revision request against the district court of Meilen sentence in November 2018, which was rejected by the cantonal court of Zürich in 2019.

A._ appealed the cantonal court of Zürich decision and requested that it be overturned and the Federal Supreme Court grant the revision request. The Federal Supreme Court approved the appeal and annulled the decision of the cantonal court of Zürich, referring the case back to the lower court for a new assessment. For the `OpposesJudgement` explanation we received the same sentence for the model-near and human-near explanation with a `BERTScore` of 0.77 (see excerpt below).

> A._ beantragt mit Beschwerde in Strafsachen sinngemäss, der Beschluss des Obergerichts des Kantons Zürich vom 2. Juli 2019 **sei aufzuheben** und zur Gutheissung des Revisionsgesuchs an die Vorinstanz zurückzuweisen. (Judgment of the Federal Supreme Court 6B 932/2019 of May 2020)

This sentence refers to the complainant's second appeal to the cantonal court of Zurich. According to the annotations, it is actually a`Supports Judgment` sentence. Looking at the only real`Opposes Judgment` sentence annotated in this case we can see that it is actually quite similar (see excerpt below).

> Das Obergericht des Kantons Zürich wies das Revisionsgesuch mit Beschluss vom 2. Juli 2019 ab. (Judgment of the Federal Supreme Court 6B 932/2019 of May 2020)

Both sentences talk about the cantonal court of Zurich, including the term “Beschluss" (decision), and the verb “zurückweisen" respectively “abweisen" (which both mean reject). The crucial difference between these is that the falsely classified sentence contains an imperative construction saying that the decision should be repealed (“sei aufzuheben"). Looking at the sentences chosen for `SupportsJudgement`, the human-near explanation presents the following correctly classified sentence:

> A._ stellte am 16. November 2018 ein Revisionsgesuch gegen das Urteil des Bezirksgerichts Meilen vom 29. Mai 2013. (Judgment of the Federal Supreme Court 6B 932/2019 of May 2020)

The model chose a neutral sentence that has a `BERTScore` of around 0.77. Which says that the involved court did not make statements concerning the appeal.

> D. Das Obergericht des Kantons Zürich verzichtet auf eine Stellungnahme. Das Bezirksgericht Meilen, B._ und C._ liessen sich innert Frist nicht vernehmen. (Judgment of the Federal Supreme Court 6B 932/2019 of May 2020)

Although the sentences for the model-near and human-near explanations for the `Opposes Judgement` label appear similar at first glance, they actually differ in their content and intended meaning. Both sentences discuss the cantonal court of Zurich and contain the term “Beschluss" (decision), but the falsely classified model-near sentence includes an imperative construction (“sei aufzuheben") that calls for the decision to be repealed, while the correctly classified human-near sentence does not. This suggests that BERT is able to classify sentences accurately as legally important or as having similar information to correctly classified sentences, as evidenced by the high `BERTScore` of 0.77.

Our analysis of other cases confirms this notion showing mixed results in the model-near explanation and mostly good examples of human-near explanations. These findings suggest that it may be useful to examine both types of explanations in order to identify areas for improvement, such as investigating possible explainability issues, analyzing bias, or improving performance. Especially the model-near explanations can reveal possible weaknesses of the model by disclosing which parts of the facts have the biggest impact on the model's prediction. Additionally, we recommend using a higher threshold for `BERTScore` to ensure that explanations with only medium agreement (e.g. 0.6 or higher) are reliable and meaningful. In cases where the explainability score is very small across all occlusion experiments, it may be necessary to consider using a different metric or introducing a threshold for the deviation between the baseline and occluded sentences. While we try using a t-test as a potential solution, our qualitative analysis shows that it is not strict enough.

The classification of `Neutral` sentences, especially suffered in this scenario, since one could argue that with only a small deviation from the baseline a sentence could be still considered technically neutral. A different evaluation method could potentially result in more `Neutral` sentences being classified correctly, but it may also risk affecting the classification of other explainability labels. It may be necessary to carefully consider and evaluate the trade-offs between improving the classification of `Neutral` sentences and potentially impacting the classification of other labels.

7.2 Lower Court Insertion – A study on Bias

The results of our lower court study show that the LCI has the greatest effect on cases in the German dataset and the least impact on cases in the Italian dataset. Interestingly, the French lower courts tend to have a pro-approval tendency, while the German and Italian courts tend to be more pro-dismissal. However, it's important to note that the changes in confidence resulting from the LCI are relatively small. Despite this, our findings suggest that inserting a lower court may have a significant impact on case outcomes and that trends and potential biases can be identified. In this section, we will try to give some explanation for the observed trends and talk about the implication of these possible biases.

7.2.1 Legal Areas

In our experiments, we have excluded the worse-performing legal area of public law Niklaus et al. (2021) and only focused on the well-performing legal areas[18]. As shown in the distribution of the lower courts (Figure 3) we do not have big disparities in the distribution among the legal areas suggesting that the observed effects on the prediction were not caused by them.

[18] These legal areas include penal law, social law, and civil law

7.2.2 Regional Bias

For the investigation of a geographical bias, we analyze the behavior of cantons and regions in the LCI experiments. We do not find a distinct bias between cantons. Most lower courts follow the overall trend of the language in which the LCI is performed, regardless of their canton. For example, even if a specific canton appeares strongly pro-approval, it simply reflects the trend of the entire language dataset. The same is true for regions. For instance, when we see a strong pro-approval trend for the R. Lémanique region, we also see the same trend for the entire language dataset. Concerning national courts, we can examine three examples from our experiments. The CH_VGer appears in French and Italian cases, while the CH_BGer appears only in French cases. The model reacts differently to these courts depending on the language. The CH_BGer is more pro-approval, while the CH_BVger strongly leans pro-dismissal (opposite each other). In Italian cases, inserting the CH_BVger has the opposite effect and results in a strong negative trend of dismissal. These results suggest that the canton or region where a court is located may not significantly impact the model's performance. These findings are supported by the results of Niklaus et al. (2021), which found that performance differences across cantons could not be correlated with any known factors, leading to the hypothesis that these differences may be due to the difficulty of particular cases or social and economic reasons.

7.2.3 Language Bias

It is important to consider the potential implications of language disparities in lower court performance, as they may lead to biases in the judicial system. The differences in performance on the SJP task and LCI between Italian and the other two languages are probably due to representation inequality in the training data[19]. In multilingual regions of Switzerland, such as the national courts and the German-French bilingual court of Berne, Fribourg, and Valais and Grison (Italian and German), these language disparities could lead to potential biases in the prediction. For example, if the same national court case is once predicted in Italian and once in German, the outputs could vary, since the German model has access to far more training samples than the Italian and is, therefore, able to accurately predict the outcome of the case or to extract more relevant information from the text. This could result in unfair treatment for parties involved in these cases, particularly if they speak Italian as their primary language.

Our study can show a specific example of these disparities. The national administrative court in French (CH_BVger) strongly leans towards dismissal. In Italian, the CH_BVger has the opposite effect and results in a strong negative trend of dismissal. This could indicate that in French the same case from the national court is more likely to be predicted as dismissed while in Italian the chance of an approved decision is greater. It is important to address these language disparities and strive for greater representation and fairness in the training data for LJP task. This could involve gathering more training data in languages that are underrepresented, or using techniques such as data augmentation or transfer learning to improve performance on these tasks in underrepresented languages. We can also suggest further investigating the trends of multilingual courts by extending the LCI.

[19] German and French have 35K and 21K training samples respectively, while Italian has only 3K training samples

8 Conclusion and Future Work

In this thesis, we presented a multilingual occlusion based explainability approach for LJP in Switzerland of 74 cases from the Federal Supreme Court of Switzerland, including cases in German, French, and Italian. In addition, we conducted a study on bias employing the variation of occlusion called LCI. We performed a detailed analysis utilizing different explainability metrics and provided ground truth for the occlusion using high-quality Legal Expert Annotations and calculating the Inter IAA.

Our results of the occlusion experiments show that the model has a varying understanding of the semantic meaning and context of the facts section. In some cases, the model is able to correctly identify a sentence with its respective explainability label. In these cases, these sentences also show a significant impact on the model's decision. This suggests that BERT has some level of understanding of the relevance of these sentences and may be able to utilize legal reasoning in its predictions. However, the model appears to struggle greatly with classifying `Neutral` sentences as such. This is particularly evident for sentences labeled with `Supports Judgment`. In this category, the model has difficulties distinguishing between legally relevant and irrelevant sentences. This may be due to the fact that legal sentences make up the majority of the facts section often mirroring legally significant sections, by using similar words, terms, and verbs but differing in their actual meaning. This notion of sentence similarity is also confirmed by the IAA results conducted between the model and the gold-standard Annotations. Using the `BERTScore` we achieved medium to high agreement. To illustrate this aspect we also conducted a qualitative analysis of the classified sentences where we compared correctly classified sentences with incorrect ones, showing that they are often quite similar. In addition, we quantify the quality of the gathered explanations using the introduced metric of the explanation accuracy score aS_{exp} and introduce two different flavors of explanations: model-near and human-near. The distribution of these scores across the experiments and flavors shows that in most cases there is at least one high-quality human-near explanation. The model-based explanations, in particular, can highlight weaknesses in the model by showing which aspects of the facts have the greatest impact on the model's prediction.

Our study on the bias with lower courts shows that the insertion of a different lower court has an effect on the prediction that may be caused by this insertion. However, our analysis shows no distinct effects concerning legal areas, cantons, and regions, suggesting that the observed effects on the prediction were not caused by group membership. We did recognize a language disparity with Italian performing much worse than the other language due to the representation inequality in the training data. In our results, we could observe possible implications of a language with this imbalance with the same courts showing the opposite effect in different languages. In multilingual regions of Switzerland, these language disparities could lead to potential biases in the prediction. With varying output for similar cases just based on the language.

Future work could address some of the limitations we faced in the process of this thesis. One approach would be to increase the sample size of the experiments, as this would allow for a more robust and reliable explanation of the model's behavior. Additionally, collecting more multilingual legal expert annotations, either by native speakers or through automation, could help to address language bias and improve the reliability of the ground truth data. Another potential avenue for improvement would be to transform the occlusion process into a classification task, using the gathered results as a training set. This could potentially provide more interpretability and insight into the model's behavior. To ensure that explanations are reliable and meaningful, it may also be useful to use a higher threshold for the `BERTScore` metric. In cases where the explainability score is very low across all occlusion experiments, it may be necessary to consider using a different metric or introducing a threshold for the deviation between the baseline and occluded sentences. This could help to improve the classification of `Neutral` sentences, but it is important to carefully evaluate the trade-offs between improving the classification of neutral sentences and potentially affecting the classification of other labels. Finally, we have provided detailed information about the steps taken to implement the annotation process, the occlusion process, and the LCI. We also make the annotation guidelines available to the public to ensure the reproducibility of the annotation task, and have made the datasets and code available to facilitate further development.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140.

Baumgartner, N. (2022). Annotation guidelines for explainability annotations for legal judgment prediction in switzerland.

Bhambhoria, R., Dahan, S., and Zhu, X. (2021). Investigating the State-of-the-Art Performance and Explainability of Legal Judgment Prediction. *Proceedings of the Canadian Conference on Artificial Intelligence*.

Chalkidis, I., Androutsopoulos, I., and Aletras, N. (2019). Neural Legal Judgment Prediction in English. Number: arXiv:1906.02059 arXiv:1906.02059 [cs].

Chalkidis, I., Pasini, T., Zhang, S., Tomada, L., Schwemer, S., and Søgaard, A. (2022). FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). A Survey of the State of Explainable AI for Natural Language Processing. Number: arXiv:2010.00711 arXiv:2010.00711 [cs].

Google LLC (2022).

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. Number: arXiv:1902.10186 arXiv:1902.10186 [cs].

Kakogeorgiou, I. and Karantzalos, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103:102520.

Küppers, F., Kronenberger, J., Shantia, A., and Haselhoff, A. (2020). Multivariate confidence calibration for object detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Lavie, A. and Agarwal, A. (2007). Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Leitner, E. (2019). Annotationsrichtlinien zum juristischen Korpus LER.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. (2016). Visualizing and Understanding Neural Models in NLP. Number: arXiv:1506.01066 arXiv:1506.01066 [cs].

Li, J., Monroe, W., and Jurafsky, D. (2017). Understanding Neural Networks through Representation Erasure. Number: arXiv:1612.08220 arXiv:1612.08220 [cs].

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. *CoRR*, cs.CL/0205028.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., Bhattacharya, A., and Modi, A. (2021). ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. Number: arXiv:2105.13562 arXiv:2105.13562 [cs].

Neubacher, F. (2017). *Kriminologie*. NomosLehrbuch Series. Nomos.

Niklaus, J., Chalkidis, I., and Stürmer, M. (2021). Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. Number: arXiv:2110.00806 arXiv:2110.00806 [cs].

Nyffenegger, A. (2022). Swisscourtrulingcorpus. https://github.com/Skatinger/SwissCourtRulingCorpus/tree/prodigy/prodigy.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Pustejovsky, J. and Stubbs, A. (2013). *Natural language annotation for machine learning*. O'Reilly Media, Sebastopol, CA. OCLC: ocn794362649.

Reiter, N. (2020). Anleitung zur Erstellung von Annotationsrichtlinien. In Reiter, N., Pichler, A., and Kuhn, J., editors, *Reflektierte algorithmische Textanalyse*, pages 193–202. De Gruyter.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Number: arXiv:1602.04938 arXiv:1602.04938 [cs, stat].

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):9.

Shrikumar, A., Greenside, P., and Kundaje, A. (2019). Learning Important Features Through Propagating Acti-

vation Differences. Number: arXiv:1704.02685 arXiv:1704.02685 [cs].

Sidyakov, M. M. (2021). Jaccard similarity and jaccard distance in python. https://pyshark.com/jaccard-similarity-and-jaccard-distance-in-python/.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. Number: arXiv:1703.01365 arXiv:1703.01365 [cs].

Wiegreffe, S. and Marasovic, A. (2021). Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. *CoRR*, abs/2102.12060:23.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

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

Zeiler, M. D. and Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. Number: arXiv:1311.2901 arXiv:1311.2901 [cs].

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. Number: arXiv:1904.09675 arXiv:1904.09675 [cs].

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2022).

List of Figures

Appendix

Additional Figures

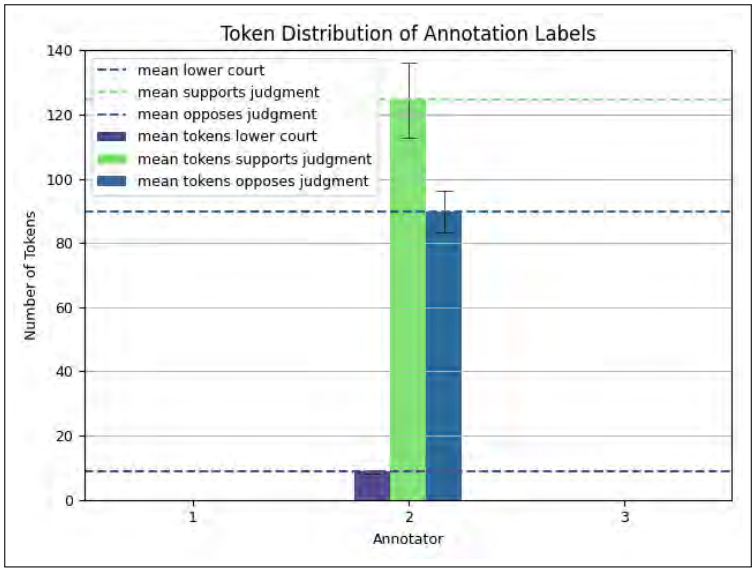


Figure 21: Distribution of the number of tokens per explainability label for the different annotators in French cases.

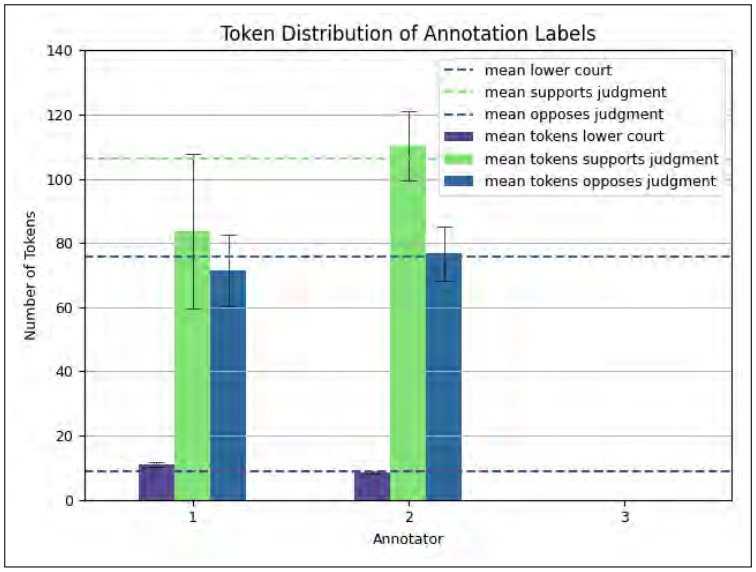


Figure 22: Distribution of the number of tokens per explainability label for the different annotators in Italian cases.

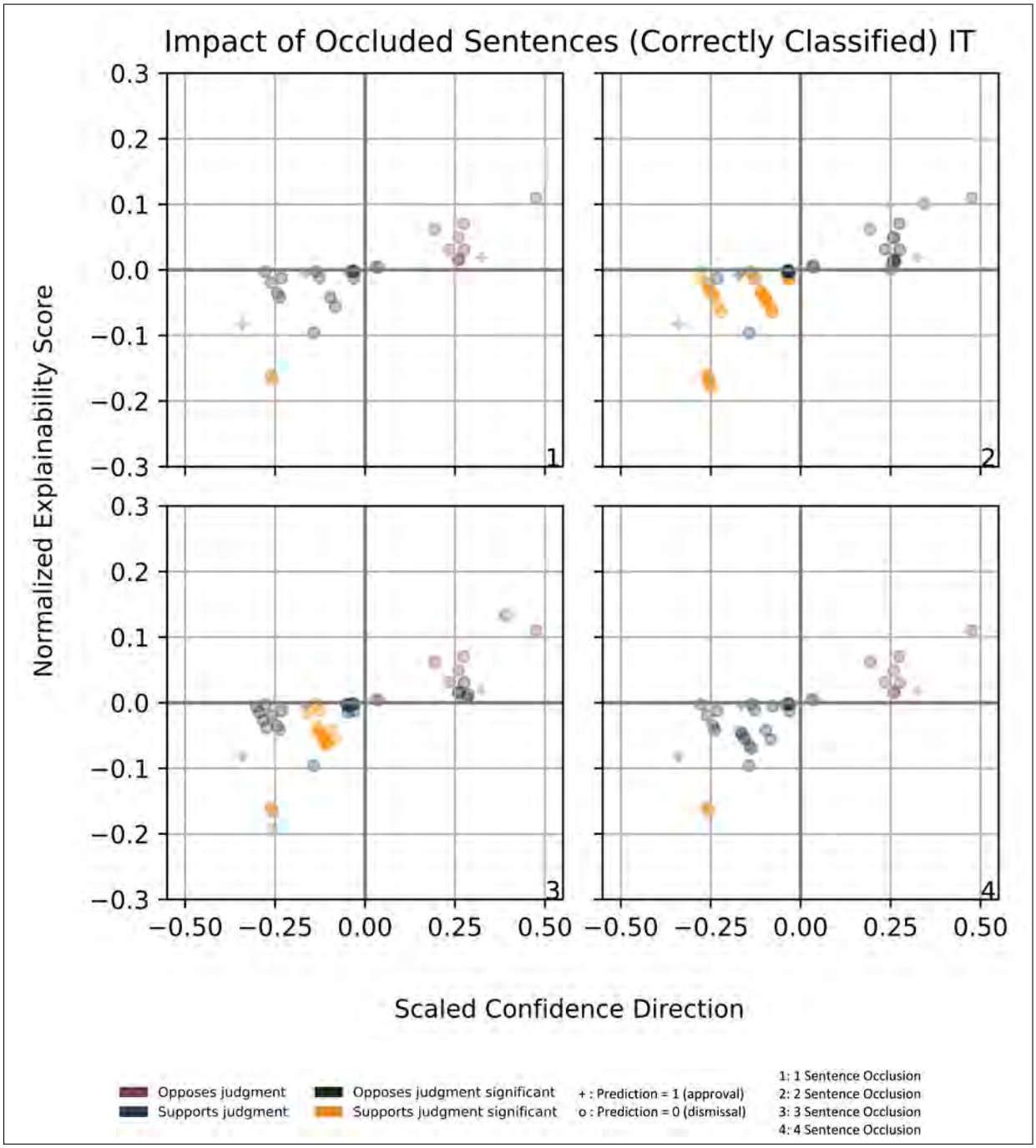


Figure 23: This plot shows the impact each correctly classified sentence in Italian has on the prediction. The further away a point is from the null axis the more impact it has on the model's prediction. The different markers indicate the prediction and the number on the bottom left the experiment number.

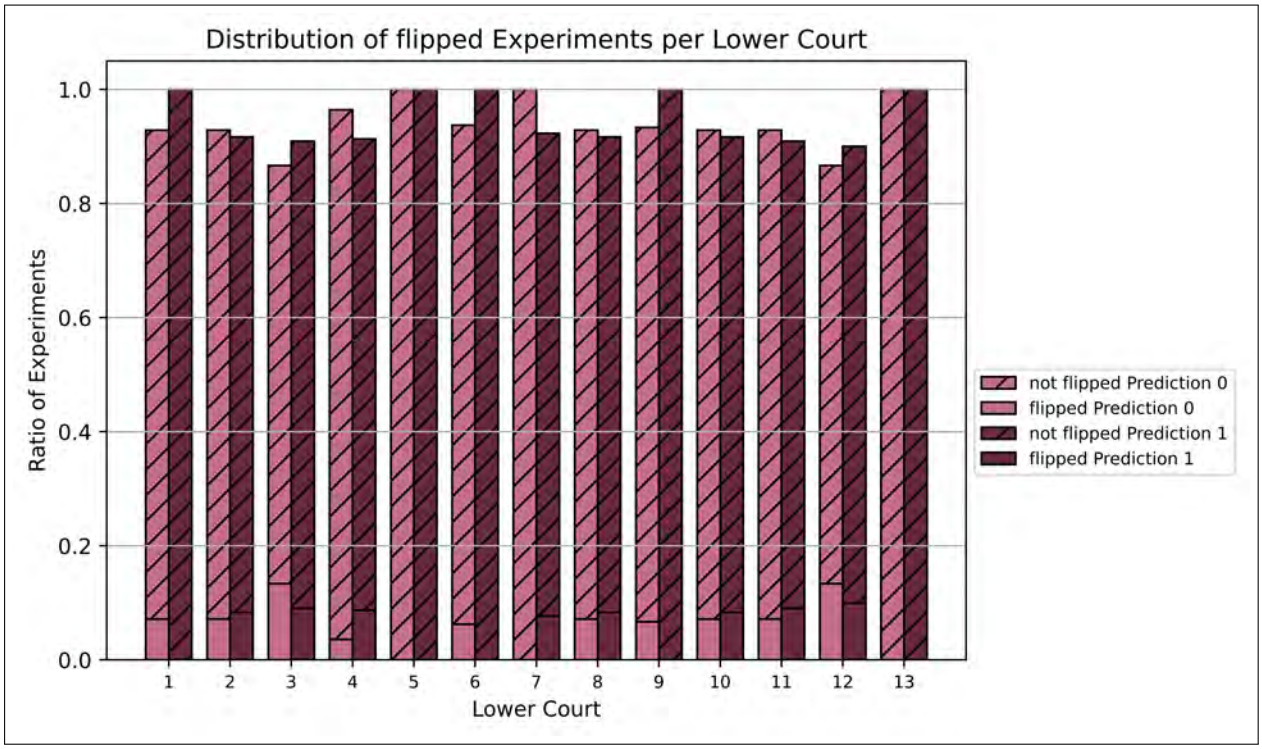


Figure 24: Distribution of flipped cases in German Lower Courts

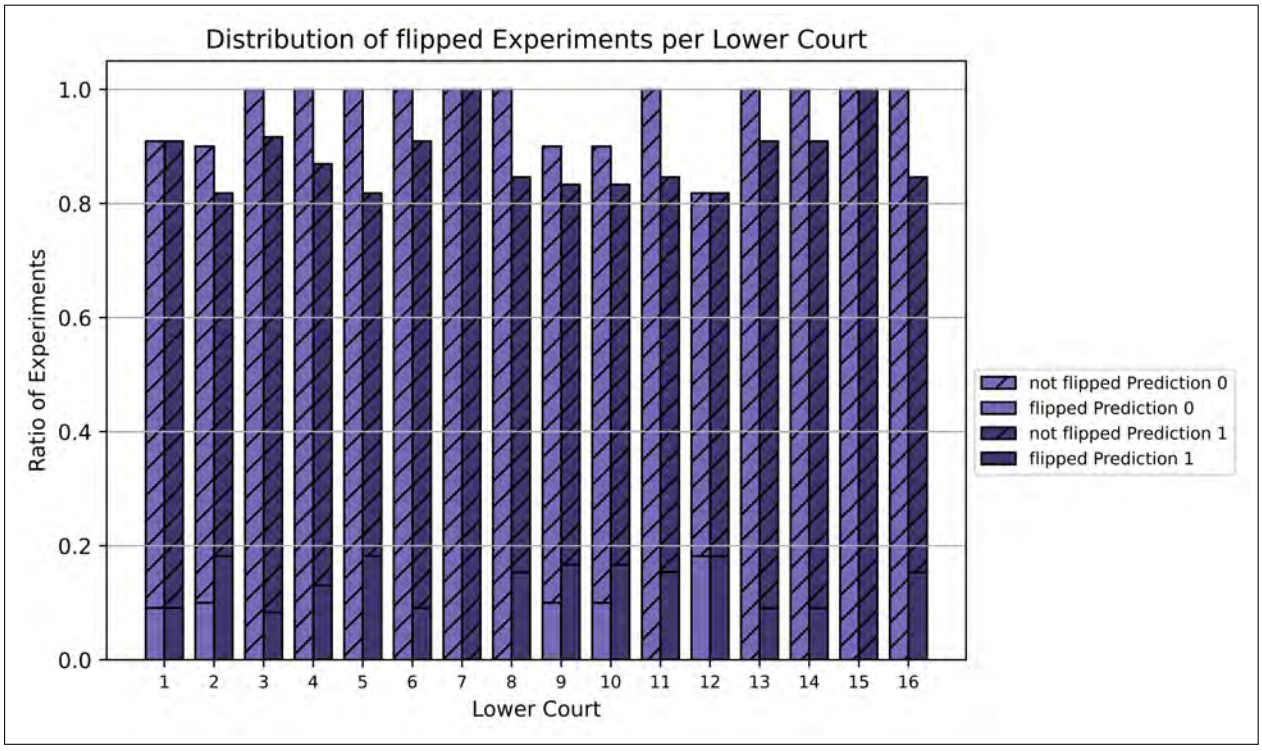


Figure 25: Distribution of Flipped Cases in French Lower Courts

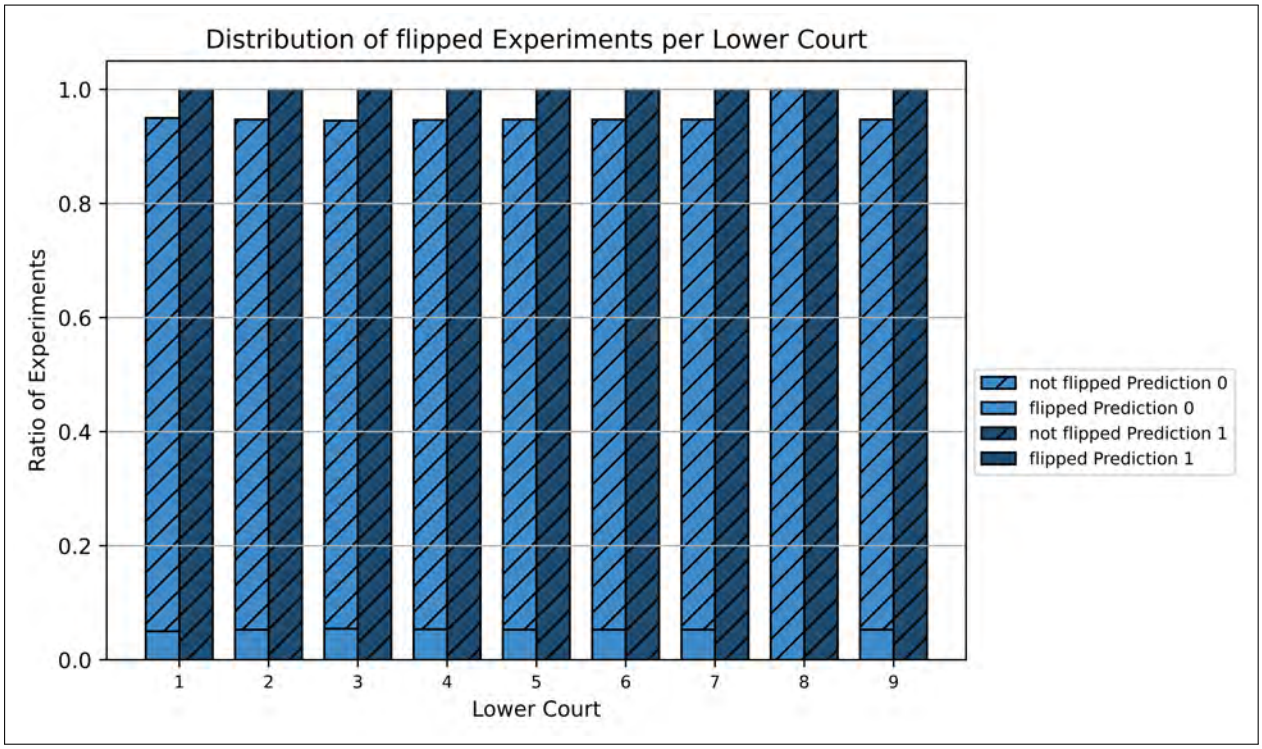


Figure 26: Distribution of Flipped Cases in Italian Lower Courts

Additional Tables

IAA Score	A1 and A2	A1 and A3	A2 and A3
Lower Court			
overlap_maximum	0.8909356725146198	0.8871794871794871	0.9678571428571429
overlap_minimum	0.9868421052631579	1.0	0.9821428571428571
jaccard_similarity	0.8830409356725146	0.8871794871794871	0.9571428571428572
meteor_score	0.9432028737333278	0.9516231778045803	0.9663561507936508
bleu_score	0.9219030288305766	0.9423475489598979	0.989927160494273
rouge1	0.9341266920214287	0.9451437451437451	1.0
rouge2	0.9105339105339105	0.9181929181929183	1.0
rougeL	0.9341266920214287	0.9451437451437451	1.0
bert_score	0.973115789473684	0.9689923076923077	0.9923071428571427
Supports Judgment			
overlap_maximum	0.8909356725146198	0.8871794871794871	0.9678571428571429
overlap_minimum	0.9868421052631579	1.0	0.9821428571428571
jaccard_similarity	0.8830409356725146	0.8871794871794871	0.9571428571428572
meteor_score	0.9432028737333278	0.9516231778045803	0.9663561507936508
bleu_score	0.9219030288305766	0.9423475489598979	0.989927160494273
rouge1	0.9341266920214287	0.9451437451437451	1.0
rouge2	0.9105339105339105	0.9181929181929183	1.0
rougeL	0.9341266920214287	0.9451437451437451	1.0
bert_score	0.973115789473684	0.9689923076923077	0.9923071428571427
Opposes Judgment			
overlap_maximum	0.28429452041093245	0.06135520350394367	0.29788434493905097
overlap_minimum	0.5354330285720722	0.20567215287769647	0.4831272197624061
jaccard_similarity	0.4259967593561048	0.14478335498324899	0.5522353348188802
meteor_score	0.4409498729410782	0.2276844547372545	0.7178908929650117
bleu_score	0.42789059173552735	0.1993542450223601	0.5866014013251174
rouge1	0.4971162801509688	0.19694415018237837	0.6302081240666247
rouge2	0.4263405748354715	0.11161035983847935	0.5565279809035123
rougeL	0.46765617121666775	0.1697221074294875	0.618128273875508
bert_score	0.7826684210526316	0.666876923076923	0.8280142857142857

Table 20: Mean IAA for the Labels (cycle 1)

IAA Score	A2 and A3
Lower Court	
overlap_maximum	0.9166666666666666
overlap_minimum	1.0
jaccard_similarity	1.0
meteor_score	0.9680397727272728
bleu_score	0.9152658292591042
rouge1	0.9444444444444443
rouge2	0.9358974358974358
rougeL	0.9444444444444443
bert_score	0.98405
Supports Judgment	
overlap_maximum	0.9166666666666666
overlap_minimum	1.0
jaccard_similarity	1.0
meteor_score	0.9680397727272728
bleu_score	0.9152658292591042
rouge1	0.9444444444444443
rouge2	0.9358974358974358
rougeL	0.9444444444444443
bert_score	0.98405
Opposes Judgment	
overlap_maximum	0.21711245389611764
overlap_minimum	0.42364373452946813
jaccard_similarity	0.3469953115119789
meteor_score	0.4898819506760849
bleu_score	0.38868404733188205
rouge1	0.46432964484183503
rouge2	0.3632447684247048
rougeL	0.3961754916553158
bert_score	0.7589999999999999

Table 21: Mean IAA for the Labels (cycle 2)

Lower Court	Abbrevation
German	
Obergericht des Kantons Aargau	AG_OGer
Versicherungsgericht des Kantons Aargau	AG_VGer
Obergericht des Kantons Appenzell Ausserrhoden	AR_KGer
Obergericht des Kantons Bern	BE_Oger
Verwaltungsgericht des Kantons Bern	BE_VGer
Kantonsgericht Basel-Landschaft	BL_KGer
Appellationsgericht Basel-Stadt	BS_AppGer
Verwaltungsgericht des Kantons Glarus	GL_VGer
Kantonsgericht Luzern	LU_KGer
Kantonsgericht Schwyz	SZ_KGer
Verwaltungsgericht des Kantons Schwyz	SZ_VGer
Sozialversicherungsgericht des Kantons Zürich	ZU_KGer
Obergericht des Kantons Zürich	ZU_OGer
French	
Tribunal fédéral	**CH_BGer**
Tribunal administratif fédéral	**CH_BVGer**
Cour d'appel civil du Tribunal cantonal fribourgeois	FR_CAPCiv
Chambre des recours pénale de la Cour de justice genevoise	GE_ChRPeCJ
Cour de justice	GE_CJ
Cour de justice de la République et canton de Genève	GE_CJRC
Cour civile du Tribunal cantonal du canton du Jura	JU_CCiTC
Cour des mesures de protection de l'enfant et de l'adulte du canton de Neuchâtel	NE_CEA
Cour pénale du Tribunal cantonal du canton de Neuchâtel	NE_CPe
Cour de droit public du Tribunal cantonal du canton de Neuchâtel	NE_CPuTC
Cour d'appel pénale du Tribunal cantonal du canton de Vaud	VD_CAPPe
Cour des assurances sociales du Tribunal cantonal du canton de Vaud	VD_CASoTC
Chambre des curatelles du Tribunal cantonal du canton de Vaud	VD_ChCTA
Chambre des recours pénale du Tribunal cantonal vaudois	VD_ChRPeTC
Tribunal cantonal du canton de Vaud	VD_TC
Chambre civile du Tribunal cantonal du canton du Valais	VS_ChCivTC
Italian	
Tribunale amministrativo federale	**CH_BVGer**
Tribunale amministrativo del Cantone dei Grigioni	GR_VG
Corte di appello e di revisione penale del Cantone Ticino	TI_CARP
Camera civile del Tribunale di appello del Canto ne Ticino	TI_CCivAP
Camera di esecuzione e fallimenti del Tribunale di appello del Cantone Ticino	TI_CEFTRAP
Camera di protezione del Tribunale d'appello del Cantone Ticino	TI_CPTRAP
Corte dei reclami penali del Tribunale d'appello	TI_CRPTA
Tribunale delle assicurazioni del Cantone Ticino	TI_TCAS
Tribunale di appello del Cantone Ticino	TI_TRAP

Table 22: Lower Court Abbreviation Table

Annotation Guidelines for Explainability Annotations for Legal Judgment Prediction in Switzerland

Nina Baumgartner

1 Introduction

1.1 Annotation Goal

Recently Niklaus, Chalkidis, and Stürmer (2021) presented a diachronic multilingual (German, French, Italian) dataset for Legal Judgment Prediction LJP including 85k Swiss Federal Supreme Court decisions. Using Hierarchical BERT, they achieved a Macro-F1 Score of up to 70%, considering penal law exclusively, they even achieved a score of up to 80%. To use Artificial Intelligence (AI) safely in high-stakes domains such as law we need explanations on how these decisions are made. To investigate explainabilty in the legal area of AI we want to gather some human and model-generated explanations of decisions from the SwissJudgmentPrediction (SJP) corpus.

This annotation task has the goal to gather the human part of the explanation. With your annotation, you will give your insight as a legal expert and tag parts of the facts with specific labels. These guidelines should help you to identify the important parts of the facts and create consistent annotations. They are based on the work of Reiter (2020), Leitner, Rehm, and Moreno-Schneider (2019) and Pustejovsky and Stubbs (2012). They are a work-in-progress in collaboration with Lynn Grau, Angela Stefanelli, and Thomas Lüthi.

1.2 Dataset

The SJP dataset is split into training, validation, and testing set. For this annotation task, a balanced subset of the SJP containing 108 cases taken from the test and validation set was created. The dataset is deemed balanced because the 108 cases are equally distributed among the three languages contained in the Swiss judicial system German, French and Italian. Each language set contains six cases over six years (2015 until 2020). With each year having two cases per legal area[1]: One with the verdict approved and one with the verdict dismissed. In addition, preference was given to cases where the model decided the correct judgment from the facts given to it, with some outliers in the French and Italian subsets.

1.3 Disclaimer

This document is a work-in-progress. If you have questions or find any errors in these instructions while doing the annotations please feel free to contact the maintainer. Please help with collecting examples to complete these guidelines.

2 The Annotation Cycle

To produce quality annotations and guidelines, which make the annotation task scalable and reproducible the annotations have to be done in cycles. Pustejovsky and Stubbs (2012) call this process the MAMA (Model-Annotate-Model-Annotate) cycle (see Figure 1 for details).

Using the annotation guidelines to identify the right parts of the text, multiple annotations by multiple individual annotators are done on the same input. Then these annotations are analyzed and the guidelines are adapted accordingly to provide consistency in the annotations. Therefore, it is important that for the first few cycles the annotations are done individually. Later the gold standard annotations for this corpus emerge from this process. Pustejovsky and Stubbs (2012) describe gold standard annotations as the final version of the annotations, which uses the most up-to-date guidelines and has everything labeled

[1]The chosen legal areas are categorized as penal law, social law and civil law

correctly. For this work, these gold standard annotations will be done as a team. For the practical aspect of this process please reference section 5.1.1, 5.1.2, 5.1.3.

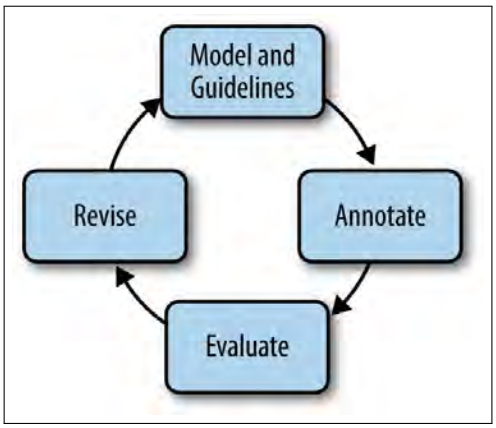


Figure 1: The inner workings of the MAMA cycle (Pustejovsky & Stubbs, 2012).

3 Annotation Entities

Although you will only be annotating the fact section of a ruling, you will have access to the full document (via a link on Prodigy) and the judgment will be indicated on the prodigy interface. You can and should use these other resources as an indicator to decide which parts of the facts are of greatest importance.

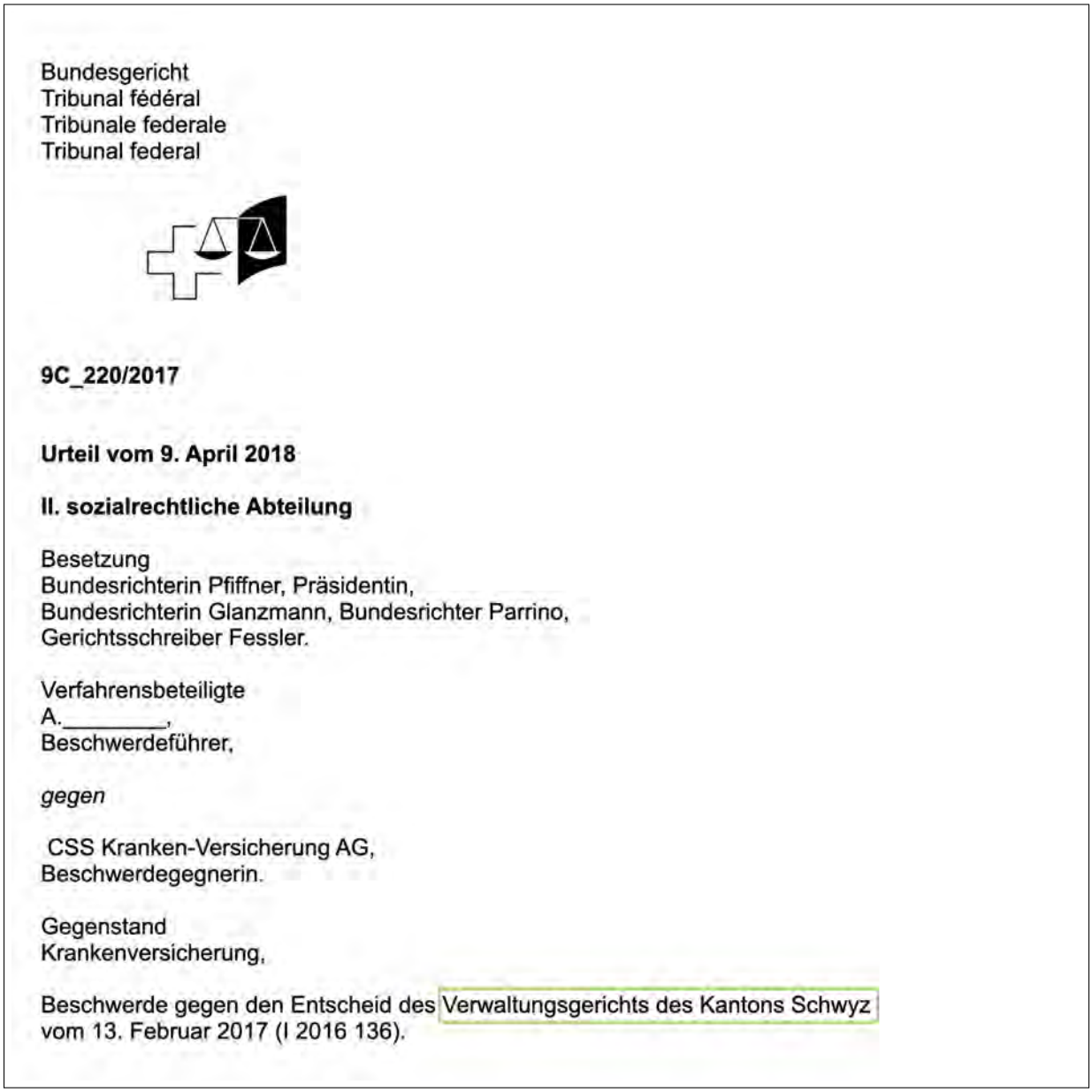
3.1 Sentences and Sub-Sentences

Wiegreffe and Marasovic (2021) identify three types of explanations in the Explainable Natural Language Processing (ExNLP) literature: highlights, free-text, and structured explanations. The explainability annotations for this task focus mainly on highlights with some addition of free-text explanations.

To add highlights you will label sentences or sub-sentences as supporting or opposing the judgment. For this task, we define a sentence as a self-contained linguistic unit consisting of multiple words, terminated with a period, semicolon, colon, question mark, or exclamation mark. An entire sentence is the largest entity to be annotated. A sentence can consist of multiple sub-sentences usually separated with a "and" or a comma. A sentence may contain two sub-sentences opposing each other, which should be consequently annotated with different labels. These sub-sentences are the smallest units that should be annotated. So single words or expressions should never be annotated. We hope that by choosing those units it is possible to indicate what the different parts of the sentences denote in the context of the judgment and to subsequently better explain the decisions of the model.

3.2 Lower Court

In addition to sentences, you will also have to annotate the last lower court of each case. As seen in Figure 2 the Rubrum of the ruling indicates the last lower court. The last lower court is composed of the name of the court e.g. "Verwaltungsgericht" and the location "Kanton Luzern". Please annotate all instances of the lower court where it appears as a complete constellation. So for example, if "Verwaltungsgericht des Kanton Luzern" appears multiple times in the facts please label it each time. Please Note that you should only annotate the lower court itself please do not label prepositions like "beim" or "zum" or verbs like "sprach" which are often found next to the lower court.


Bundesgericht
Tribunal fédéral
Tribunale federale
Tribunal federal

9C_220/2017

Urteil vom 9. April 2018

II. sozialrechtliche Abteilung

Besetzung
Bundesrichterin Pfiffner, Präsidentin,
Bundesrichterin Glanzmann, Bundesrichter Parrino,
Gerichtsschreiber Fessler.

Verfahrensbeteiligte
A._____,
Beschwerdeführer,

gegen

CSS Kranken-Versicherung AG,
Beschwerdegegnerin.

Gegenstand
Krankenversicherung,

Beschwerde gegen den Entscheid des Verwaltungsgerichts des Kantons Schwyz
vom 13. Februar 2017 (I 2016 136).

Figure 2: Screenshot of a Rubrum with the lower court highlighted Judgment (of the Federal Court) from September 8th 2017.

4 Annotation Categories

To annotate the sentences of each fact section you will be using two labels, *Supports judgment* and *Opposes verdict*. You should also highlight the lower court for each judgement. In addition, you will be given several options for dealing with problematic cases, which should help to improve the dataset, these guidelines, and the annotations themselves.

4.1 Supports Judgment

This label is used when a sentence or sub-sentence supports the judgment. Every sub-sentence that supports the judgment should be annotated.

4.2 Opposes Judgment

This label is used when a sentence or sub-sentence opposes the judgment. Every sub-sentence that opposes the judgment should be annotated.

4.3 Lower Court

This label is used to highlight the last lower court of the case. To label the last lower court highlight the name and the location of the court as one instance (see Figure 3).

Figure 3: Example of a highlighted lower court in Prodigy.

4.4 Neutral

Every not-labeled sub-sentence is considered neutral. This is not a label per se but merely how the system interprets words or sentences which are not assigned one of the labels above. It is important for the analysis that even the neutral sentences are annotated which in our case means to omit them.

One example in German of a neutral expression that should not be tagged with a label is the word *"Sachverhalt:"*. This word only indicates the beginning of the fact section and should be left out as a neutral part of the facts because it does not give us any further information on the explainability of the judgment.

Another example of a neutral part of the facts are the section indicators labeled with capital letters (e.g. *A., B., A.a., A.b and so on*). Note that witnesses, accused persons, and other involved parties are also labeled with uppercase letters and should be annotated if part of a sentence (see 4 below as illustration).

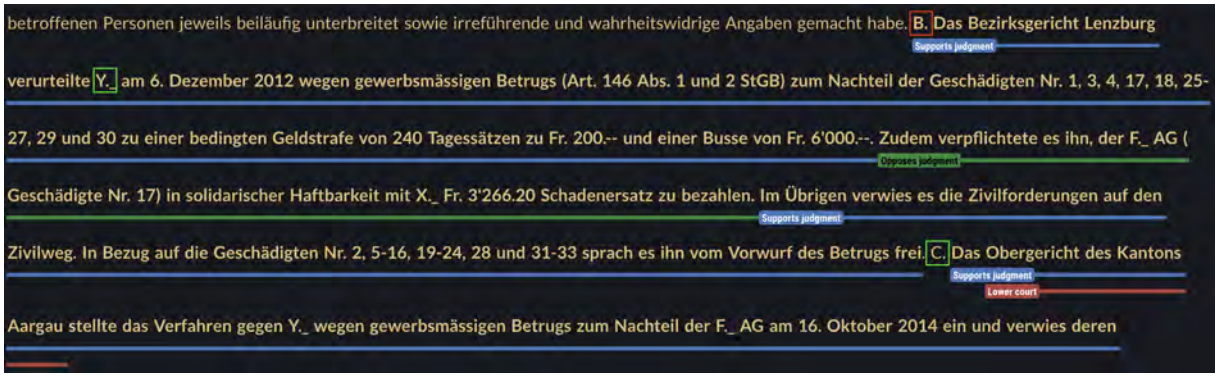


Figure 4: Example an annotation where the uppercase letters are first wrongly (marked red) and then correctly annotated (marked green).

Note that when annotating neutral sentences in the gold standard iteration the defined rules concerning sentences and sub-sentences should be followed. Please annotate each neutral sentence individually, with entire sentences being the biggest neutral instance.

4.5 Problematic Cases

Problematic cases can occur. For now, we differentiate between three possible types of such cases.

4.5.1 Rejected Cases

If a case is badly tokenized[2] or there is another formal error it should be rejected. Please state your reasoning in the comment window using the comment pattern below and reference the Reject or Ignore a Case section of this document for the details on how to properly reject a case. Figure 5 is an example of a case with formal errors.

[2]Tokenized means that the system did not properly separate the words.

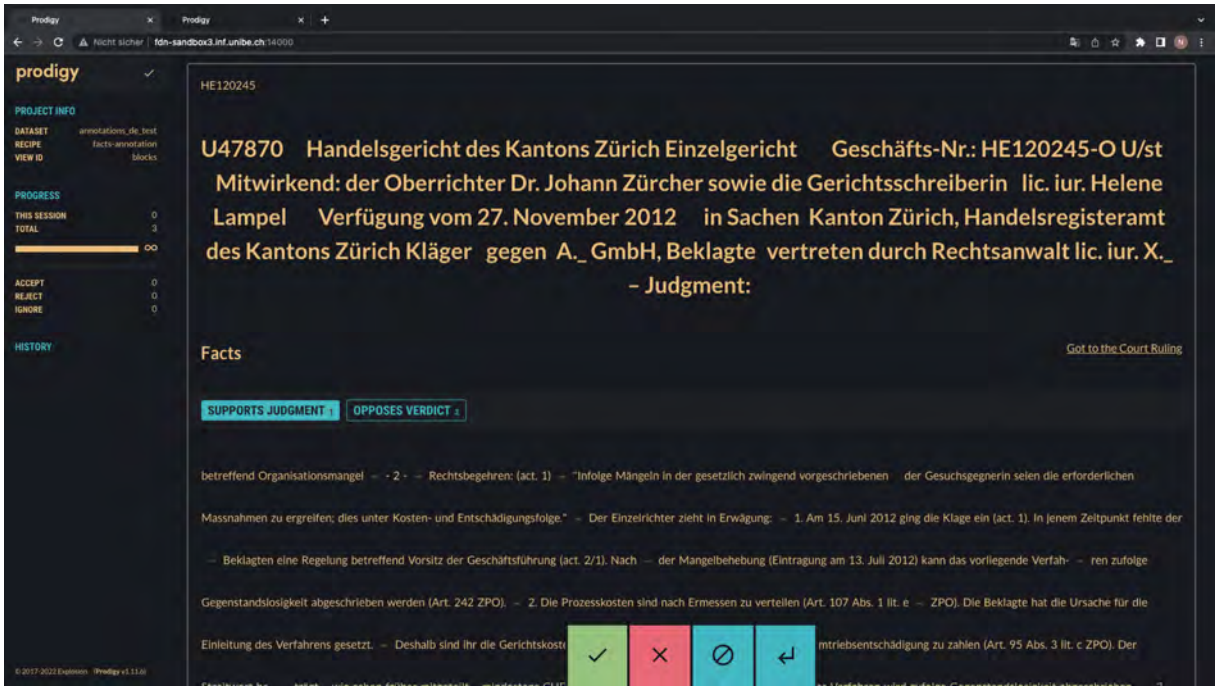


Figure 5: Example of a case containing a formal error that should be rejected. Here the title was parsed incorrectly, the judgment is missing and the facts are tokenized wrongly and incomplete.

4.5.2 Ignored Cases

If a case is too short or otherwise unfit for the annotation it should be ignored. To ignore it please state your reasoning in the comment section and follow the steps explained in the Reject or Ignore a Case section of this document below.

An example of a case that was ignored by an annotator is the Judgment (of the Federal Court) from April 8th 2020 (please reference the whole case online via link). The annotator who ignored this case explained his reasoning as follows in the free text explanation:

> "Before the Federal Court, only the question of party compensation was in dispute. The underlying facts, however, actually have nothing to do with the court's decision."

This argumentation can be supported by the following parts of the facts section from from April 8th 2020:

> "B. [...] Allerdings verpflichtete [das Verwaltungsgericht des Kantons Bern] die Suva-MV, A._____ eine Parteientschädigung in der Höhe von Fr. 3610.15 [...] zu bezahlen.
> C. Die Suva-MV erhebt Beschwerde in öffentlich-rechtlichen Angelegenheiten und beantragt sinngemäss, der angefochtene Entscheid sei bezüglich der Zusprache der Parteientschädigung aufzuheben. A._____ beantragt, auf die Beschwerde sei nicht einzutreten, eventuell sei sie abzuweisen."

4.5.3 Other Problematic Cases

There might be cases without formal errors where you have difficulties annotating (neither reject nor ignore). In such cases, please annotate to the best of your ability and explain your reasoning in the comment section.

4.5.4 Comment Structure

Comment for rejecting and ignoring case

Number of the case – Annotators name

- Why did you ignore/reject this case?

Comment for generally problematic case

Number of the case – Annotators name

- Why is this case problematic and difficult to annotate?
- How did you decide on your annotation?

5 Implementation: How to Annotate the Dataset using Prodigy

This section explains how to use the annotation tool Prodigy[3]. We built a custom recipe for this task which lets you annotate the facts section of a given court decision.

5.1 Access

The Prodigy instance can only be accessed via the University of Bern network. If you want to annotate from home you must use the VPN of the University of Bern[4].

If you are connected to the university network you can access Prodigy via one of the URLs in the following three sections. Before you can start you will be asked to provide a *username* and a *password*, which will be given to you by the maintainer of the annotation process. After the login procedure, you should now see an overview of the case and you can start with your annotation.

5.1.1 First cycle

The following links will be used for your pilot annotations (first iteration). If you completed the annotation on this dataset ignored and rejected cases will be replaced with other cases having the same legal area, year, and judgment. This process is ongoing until we reach 36 accepted cases.

- German case annotations:
  - Angela: http://fdn-sandbox3.inf.unibe.ch:11000/?session=angela
  - Lynn: http://fdn-sandbox3.inf.unibe.ch:11000/?session=lynn
  - Thomas: http://fdn-sandbox3.inf.unibe.ch:11000/?session=thomas
- French case annotations: http://fdn-sandbox3.inf.unibe.ch:12000/
- Italian case annotations: http://fdn-sandbox3.inf.unibe.ch:13000/

Note that sessions can be added dynamically by adding the suffix /?session=SessionName to the url.

5.1.2 Further Cycles and Corrections

If you have completed all the pending annotations on the above URLs Prodigy will display a message saying no task is available. This is your indicator to continue to this part of the annotations. Reference the Guideline for recent changes and adapt your annotations accordingly. You can repeat this process with a new session as often as you want (see session management example below).

- German case annotations:
  - Angela: http://fdn-sandbox3.inf.unibe.ch:11001/?session=angela
  - Lynn: http://fdn-sandbox3.inf.unibe.ch:11002/?session=lynn

[3] https://prodi.gy/

[4] https://serviceportal.unibe.ch/sp?id=kb$_article_viewsysparm_article = KB$0010032

  - Thomas: http://fdn-sandbox3.inf.unibe.ch:11003/?session=thomas

- French case annotations: http://fdn-sandbox3.inf.unibe.ch:12000/?session=lynn
- Italian case annotations: http://fdn-sandbox3.inf.unibe.ch:13000/?session=angela

If you need to do multiple corrections on the same case please add a number behind your link as seen below to distinguish between the sessions:

- Session 1: http://fdn-sandbox3.inf.unibe.ch:11001/?session=angela1
- Session 2: http://fdn-sandbox3.inf.unibe.ch:11001/?session=angela2

5.1.3 Final Gold Standard Annotations

After some iterations, you and the other annotator will get together and decide on the final annotation using the below link.

- German gold standard annotations:
  - http://fdn-sandbox3.inf.unibe.ch:8080/?session=gold

The gold standard annotation uses a different prodigy layout where the best annotation can be chosen and adapted according to the latest version of these guidelines. Figure 6 and figure 7 display the appearance of this prodigy setup.

Figure 6: Screenshot of review setup on prodigy (1).

Figure 7: Screenshot of review setup on prodigy (2). The different annotations can be selected by clicking on the name of the annotation (highlighted in yellow). If an annotation is highlighted blue it means that it was ignored by by this annotator.

Note that this review setup contains a new label called *Neutral*. Even though neutral sentences should not be annotated in the previous iteration of the annotation it is important to label them when doing the gold standard annotation. Please reference section 4.4 to learn what should be highlighted with the neutral label. The example in figure 8 below shows how to correctly annotate the facts when doing the gold standard annotation.

Figure 8: Screenshot of an annotated fact section with the neutral label.

5.2 Annotate a Sub-Sentence

To label a phrase with a tag, highlight it with your cursor and choose the corresponding label. To delete a tag simply click on the tagged words again. As seen in Figure 9 the two labels appear in two different colors. By hovering over an annotated section the delete toggle appears.

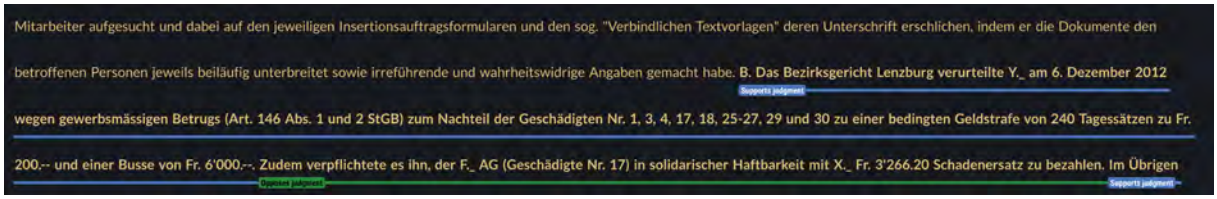


Figure 9: Screenshot of sentence labeling in prodigy.

If you are happy with your annotation you can accept it by clicking on the green check labeled with [1] in Figure 10 and save it by pressing the save button in the left corner referenced by the number [2]. To see your progress you can look at the information displayed on the left (see the number [3] on Figure 10). If you want to access the original document you can click on the link in the right corner (see the number [4]). Please do not forget to save your progress using the save button [2].

If you want to skip a case because you already annotated it. Please use the accept button [1] to get to the next case.

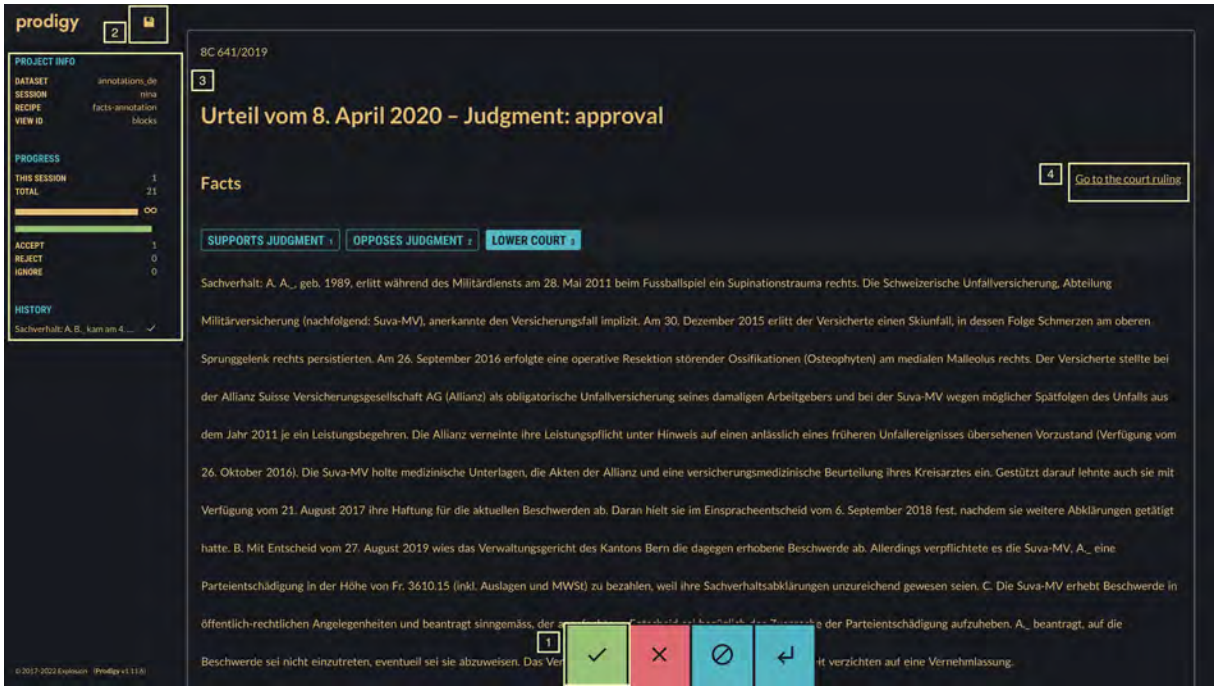


Figure 10: Screenshot of the case overview on prodigy

5.3 Reject or Ignore a Case

To reject a case state your reasoning in the comment section and press the red cross to reject it. To ignore it, press the blue button with the stop signal after commenting. Do not forget to save your progress. Figure 11 shows the interface of the comment section and the ignore and reject buttons.

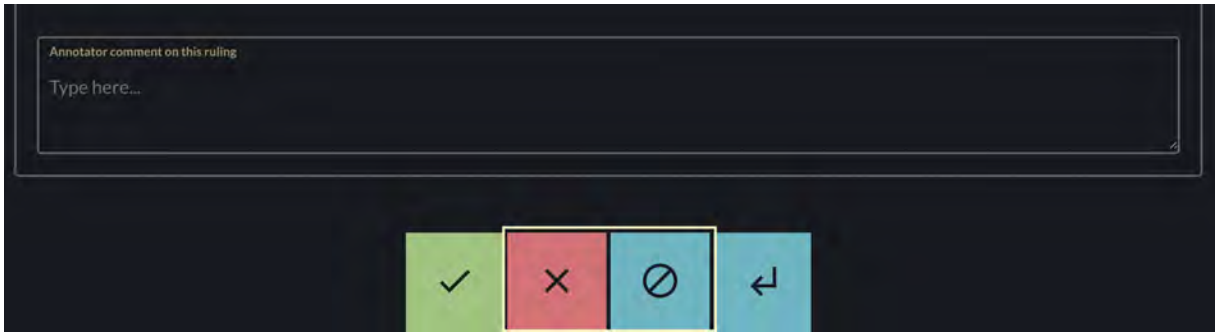


Figure 11: Reject and Ignore buttons

6 Change Log

This change log documents the progress of these guidelines. When adapting these guidelines please also add a new entry to the changelog using the following structure

Template

Date – Title of changes

- Which parts were changed in this iteration?
- Why was this part changed?

10.04.2022 – Formal changes after first feedback

- Which parts were changed in this iteration?
  - Changed E. Leitner reference to the published article.
  - Corrected some spelling errors.
  - Integrated the figures into the text of the guidelines.
  - Changed label "opposes verdict" to "opposes judgment"
- Why was this part changed?

  With this first adaption of the guidelines we mainly worked on some formal errors to standardize the format and clarify the instruction (especially with integrating the figures into the text). The label was changed to make the annotation and their interpretation more consistent.

23.04.2022 – Changes to Prodigy setup and new label

- Which parts were changed in this iteration?
  - Named multi-user sessions were added to the Prodigy setup, which changed the annotators' URLs in this document.
  - The label lower court was added as a new annotation category and subsequently to the prodigy setup. Explanation of how and when to use it was added to sections 2 and 3.
  - Directions on how to skip already annotated cases were added. To section 4.2
- Why was this part changed?

  The named multi-user session was a pending part of the prodigy setup, which is now resolved. The URLs of the annotators had to be adapted accordingly. After a meeting with the lawyer and annotator Thomas Lüthi, we decided on adding the new label "lower court", to highlight it as a separate entity additionally to the existing two labels. Correct sessions were not yet implemented in the first setup of Prodigy used for some annotations, for this reason, directions on how to skip a case of an already annotated case were added.

12.05.2022 – Changes to Prodigy setup, introduction revision, explanation of the annotation cycle

- Which parts were changed in this iteration?
  - The Prodigy setup was extended to enable the iterative work on the annotations. Therefore, an explanation on when to use which link was added. In addition explanations on the annotation cycle itself were added.
  - Updated images because Prodigy Interface changed
  - After writing the proposal of the thesis corresponding to these guidelines the introduction was adapted accordingly.
  - Ignored Case example was added
- Why was this part changed

  Enabling the iterative process is an important step to provide quality annotations and guidelines. Therefore after the setup was implemented the guidelines had to be adapted. The images had to be updated because they where no longer up to date and to provide consistency in these guidelines. To give the annotator a better understanding of the task the introduction was updated with some input from the proposal. After analyzing the currently done annotation an example of an ignored case could be added to these guidelines.

09.08.2022 – Neutral label addition, instruction gold standard, grammarly

- Which parts were changed in this iteration?
  - Neutral label was added in the gold standard iteration
  - Added explanation, instruction, and screenshots about the gold standard annotation prodigy setup
  - Language correction using Grammarly
- Why was this part changed

  After an intense meeting with Joel Niklaus, we decided on adding the neutral label in the gold standard iteration. this will help with the section splitting for the occlusion later. For this reason, some instructions and an example on how to use this label had to be added. In addition, an overview and some explanation on how to handle the gold standard setup were also added to the implementation section. Lastly, the language was revised using Grammarly.

21.07.2022 – Language extensions, clarification of the instructions

- Which parts were changed in this iteration?
  - Extensions to French and Italian in the iterative annotation cycle in the implementation part of these guidelines.
  - Added some clarification to the lower court label which also specifies how often it should be annotated.
  - Added the section dividing capital letters as a new neutral element.
- Why was this part changed

  Enabling the iterative process is an important step to provide quality annotations and guidelines. Therefore after the setup was implemented in Italian and French the guidelines had to be adapted. After reviewing the first results of the annotation done using these guidelines some clarification for a more consistent annotation where added. The new neutral element and clarification in the lower court section will help to prevent distortion in the annotator agreement caused by minor shifts at the start of the label. The clarification that all lower court instances appearing in complete form should be annotated, was added so that the annotation was most similar to the models' output (the model will extract all instances of the lower court).

References

from April 8th 2020, C. . (2020). *Bgu 8c 641/2019.* Retrieved from https://www.bger.ch/ext/eurospider/live/de/php/aza/http/index.php?highlight_docid=aza%3A%2F%2F08-04-2020-8C_641-2019&lang=de&type=show_document&zoom=YES&

from September 8th 2017, C. . (2017). *Bgu 9c 424/2017.* Retrieved from https://www.bger.ch/ext/eurospider/live/de/php/aza/http/index.php?lang=de&type=highlight_simple_query&page=1&from_date=20.08.2017&to_date=08.09.2017&sort=relevance&insertiondate=&top_subcollection_aza=all&query_words=&rank=9&azaclir=aza&highlight_docid=aza%3A%2F%2F08-09-2017-9C_424-2017&number_of_ranks=456

Leitner, E., Rehm, G., & Moreno-Schneider, J. (2019, 9). Fine-grained Named Entity Recognition in Legal Documents. In M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, & Y. Sure-Vetter (Eds.), *Semantic systems. the power of ai and knowledge graphs. proceedings of the 15th international conference (semantics 2019)* (pp. 272–287). Karlsruhe, Germany: Springer. (10/11 September 2019)

Niklaus, J., Chalkidis, I., & Stürmer, M. (2021). *Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark.* arXiv. Retrieved from https://arxiv.org/abs/2110.00806 DOI: 10.48550/ARXIV.2110.00806

Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning* (Nos. Bd. 9,S. 878). O'Reilly Media, Incorporated.

Reiter, N. (2020). Anleitung zur erstellung von annotationsrichtlinien. In N. Reiter, A. Pichler, & J. Kuhn (Eds.), *Reflektierte algorithmische textanalyse: Interdisziplinäre(s) arbeiten in der creta-werkstatt* (pp. 193–202). De Gruyter. Retrieved from https://doi.org/10.1515/9783110693973-009 DOI: doi:10.1515/9783110693973-009

Wiegreffe, S., & Marasovic, A. (2021). Teach me to explain: A review of datasets for explainable NLP. *CoRR*, *abs/2102.12060*. Retrieved from https://arxiv.org/abs/2102.12060

List of Figures