# Decomposing Complex Text Classification Tasks through Error Analysis: A Study on Genocide-related Court Hearings

David Preda[1], João Baltazar[1], Luís Viegas[1], and Henrique Lopes Cardoso[1,2]

[1] Faculty of Engineering of the University of Porto
[2] Artificial Intelligence and Computer Science Lab (LIACC)
Rua Doutor Roberto Frias, s/n, 4200-465 Porto, Portugal
{up201904726,up201905616,up201904979}@up.pt, hlc@fe.up.pt

**Abstract.** Recent advances in natural language processing (NLP) with Transformer-based models have reduced many classification tasks to a pretrain and fine-tune problem. However, several datasets are too complex for such a simplistic approach. We analyse the Genocide Transcript Corpus (GTC) to show that, by carefully analysing the errors made by classifiers on such datasets, one can make more informed choices on what pretrain domains best fit a given task. We start off from a wide array of classical techniques and make our way up to BERT-based models through error analysis, beating the state-of-the-art for classification on the GTC.

**Keywords:** Natural Language Processing · Error Analysis · Transformers · Fine-Tuning.

## 1 Introduction

With the introduction of the transformer architecture [10], pretrain and fine-tune became the general approach to most natural language processing (NLP) tasks. However, certain datasets are hardly approachable with such a straightforward strategy, as their label assignment is far too specific for there to be any other similar tasks. Furthermore, by being semantically dissimilar from any other datasets, these tasks are less keen to transfer learning from pre-trained language models [7].

The Genocide Transcript Corpus (GTC) [9], comprised of court hearings on genocide cases from 3 different tribunals, is one such example. In this corpus, *all samples were labeled according to whether they contain a witness's description of experienced violence* [9, p. 4]. This means that a classifier has to take into account whether the contents of a paragraph are violent, first-person, and described orally during the hearing itself.

It is thus difficult to pinpoint a model pre-trained on similar tasks, that are not only semantically similar but also labelled in a similar fashion. Therefore, we depart from various classical NLP techniques and analyze when and why these

approaches fail to capture the correct class for a given sample. Specifically, we look for the semantic aspects that confuse simpler models in order to identify language models that have been specifically pre-trained for better dealing with those same aspects.

We showcase that, by following this approach, it is possible to surpass the state-of-the-art for the GTC corpus within the same training and testing conditions without simply resorting to larger-scale models. Additionally, we analyze the misclassifications of the new improved model, comparing them with classical classifiers in order to understand what progress has been achieved. Finally, we analyse in which situations our best model underperforms and propose solutions for future work.

## 2 Related work

**GTC** The Genocide Transcript Corpus (further detailed in Section 3) is an underexplored dataset – only the author's models [9] for this dataset can be found online (as of the time of writing). Marginally similar problems, such as hate speech detection, which requires similar attention to context before classification, have been further explored [6].

**Form vs Meaning** Our approach to tackling the complex labelling process of the GTC relies on a thorough analysis of our models' shortcomings. For each of the examples we analyze, we infer what the model is capturing and paying attention to – as the entries of the dataset get increasingly complex, it is arguable whether the model understands its contents or just looks for common indicators of one class or the other. This is a recurrent problem in NLP [1,2], especially with the recent increase in the capacity and complexity of "black-box" Deep Learning models.

**Feature Engineering vs Fine-Tuning** Before the rise of Deep Learning, most NLP models relied on carefully hand-crafted features engineered from the original dataset. While the current trends follow an end-to-end neural network approach, feature engineering can still provide valuable insights to choose the most appropriate pre-trained models for a given task.

**Transfer Learning** The introduction of the transformer architecture has revolutionised natural language processing, leading to the generalized adoption of pre-training and fine-tuning approaches. These pre-trained language models, such as BERT, significantly simplified classification problems by leveraging large-scale language modelling pre-training followed by task-specific fine-tuning. While the pretraining and fine-tuning approach has become widely adopted, there are multiple methods available to fine-tune a model for a new task. Initially, a model pretrained with generic data can be fine-tuned for objectives similar to the target operation [8] (these objectives can differ semantically from the initial one)

using cross-entropy or masked language modelling. This process is then followed by fine-tuning towards the specific goal. On the other hand, another strategy is to tailor the model explicitly to the domain at hand [5]. The choice of strategy hinges on the intended outcome of the model, whether it is optimized for performance on a particular type of operation or a variety of operations within a given domain.

Thus, one can select a pre-trained model based on the semantic similarity between the pre-train data and the dataset to be classified. This eases the transfer learning process between tasks [7] and allows for better performance.

## 3 The Genocide Transcript Corpus (GTC)

The Genocide Transcript Corpus is a corpus composed of transcripts from the hearings of genocide cases in 3 different international tribunals [9]. Data distribution is detailed in Table 1. Each entry is made of a paragraph from a given transcript, additional information such as the tribunal it originated from and a binary label. This label indicates whether that paragraph contains an oral account, by one of the witnesses, of *directly experienced violence*. Table 2 showcases examples presented by the GTC authors.

| | n0 | n1 | ntotal |
|---|---|---|---|
| All tribunals | 946 | 529 | 1475 |
| Extraordinary Chambers in the Courts of Cambodia | 286 | 179 | 465 |
| International Criminal Tribunal of the Former Yugoslavia | 401 | 129 | 530 |
| International Criminal Tribunal for Rwanda | 259 | 221 | 480 |

**Table 1.** GTC samples for both classes, as presented by its authors [9, p. 4].

Therefore, in order for an entry to have a positive label, its' paragraph has to meet all of the following criteria:

- It has to contain a witness description of experienced violence.
- That same description must be a direct experience – meaning that the witness must have suffered directly from it or have been present at the time of a violent act against another person. This rules out the retelling of third-party experiences.
- That same description must also be given by the witness during the court hearing.

| Negative Class Sample | Positive Class Sample |
|---|---|
| Q. As we discussed before, I will ask you some questions concerning your experiences in Rwanda back in 1994. Back in April of 1994 where did you live? And please you can just specify by commune.<br>A. We were living in Taba commune.<br>Q. Is that in Rwanda?<br>A. It's a commune in Rwanda, in Gitarama prefecture.<br>Q. Around the beginning of April did you ever receive news of the crash of the president's plane?<br>A. Yes, I heard this. [...] | Q. What happened next?<br>A. He took me and he had a very long knife that he was wearing in his belt and also a small ax in his hand. We arrived near the primary school. The classrooms are very close to the place where we were before and it's very close to the road, as well, and when we arrived at that location **this child put down his ax, he also put down the long knife, near me, and you see these things are not very easy to see, a young child like that rape me.** I hope you understand that this is something that is very, very painful. [...] |

**Table 2.** Exemplifying samples from both of the GTC classes, as presented by the dataset's authors [9, p. 4].

While the first point is simpler – violence-related vocabulary differs quite significantly from the remainder of the lexicon, thus making it easy to detect violence – the two other points are too complex to approach this problem only at a semantic level. If one only takes into account the semantic value of the text, it becomes impossible to understand the role played by the witness in a description - whether they were present or not, whether they suffered from it or not - or even who is telling what. Therefore, the syntax of the samples must also be taken into account in order to retrieve this key context for classification.

## 4 Classical Approach

In order to better understand how to tackle the complexities of the GTC, we extract various features from the dataset, mainly related to first *vs* third-person discourse and positive *vs* negative class vocabulary. We also train several classifiers paired with different input vectors and choose the best pair for error analysis.

### 4.1 Feature Engineering

Two main types of features were explored, each with a different underlying motivation. We begin by exploring first *vs* third-person discourse. The positive class contains first-party accounts of experienced violence. Therefore, the detection of first-person discourse should help with pinpointing which entries are a direct

testimony and which are retellings of another person's experience. This is an underexplored area of NLP, with some work developed on third to first-person rephrasing [3]. This work could be used to train a classifier to detect samples containing first-person discourse, which would help rule out third-party accounts. We were unable to obtain the original dataset and thus had to develop another approach.

Two different features were constructed using personal pronouns: a first-person pronoun counter and a third-person pronoun counter. We theorize that witnesses will focus on retelling what they personally experienced – what they saw, how they felt, what they did, what was done to them, how they reacted – therefore referring to themselves or a party they are in many times as they retell their experiences, with occasional mentions of the perpetrator. Feminine third-person pronouns were left out of this experience, as all the accused were men.

We then explored the differences between the vocabularies of the positive and negative classes. Violence often has a very strong vocabulary associated with it. Therefore, we hypothesize that the frequency of some words in the positive class must be much greater than in the negative class, as violent words won't be used as much outside of witness testimonies. For this, we use the TF-IDF vector of each entry's text and take the mean value per class for each word across all entries. We then subtract the negative label frequency from the positive one, which results in a TF-IDF difference, telling how much more frequently a word happens in the positive class than in the negative one. Finally, for each entry, we take the average of all (except stopwords) its words' TF-IDF difference. This makes it so that the final per-entry metric is solely based on the words that bring semantic value to the text.

Also based on this metric, we assessed that violence-related words, such as *kill* or *rape*, are more commonly found on the positive class than on the negative side, thus empirically confirming our hypothesis.

| Feature | Description | Correlation with Label |
|---|---|---|
| First Person Pronouns | Count of occurrence of pronouns 'I', 'me', 'my', 'we', 'us', 'our' | 0.233853 |
| Third Person Pronouns | Count of occurrence of pronouns 'he', 'his', 'they', 'their' | 0.276952 |
| TF-IDF Class Difference | Difference between TF-IDF values for positive and negative classes | 0.613511 |

**Table 3.** Features extracted for dataset analysis (Correlation with Label indicates the Pearson correlation coefficient between the label and each feature).

In Table 3 we report the Pearson correlation coefficient for the correlation between the binary label and the newly created features. To our surprise, third-

person pronouns ended up having a greater influence on the entries' class than first-person ones. We believe this is due to the witness describing the actions of other people, therefore focusing more on tertiary actors than on themselves, but no further exploration was done in this direction.

The TF-IDF difference revealed itself to be the most influential feature, with a correlation of 61.35% with the label. Therefore, the semantic aspect of the source text plays a big role in whether or not a paragraph contains a description of violence but doesn't fully explain an entry's class as it doesn't provide the syntactic context required to capture the most complex aspects of the dataset.

### 4.2 Experimental Evaluation

Before evaluating other techniques we define as a baseline model a Multinomial Naive Bayes classifier with a vector of word counts as input. The model is trained on an 80/20 train/test split of the original dataset. Our baseline model attains a **macro F1-score of 0.7692**. Other metrics can be found in Table 5.

We test a set of 6 algorithms combined with four different input formats. We report the macro F1-scores in Table 4. Each model undergoes the same training process as the baseline model but with the addition of conducting a hyperparameter search prior to training.

| | **Bag of Words** | Bigrams | TF-IDF | Word Embeddings |
|---|---|---|---|---|
| **Multinomial Naive Bayes** | **0.7692** | 0.7000 | 0.7434 | —— |
| Decision Tree | 0.6186 | 0.6161 | 0.5758 | 0.5888 |
| Logistic Regression | 0.7431 | 0.7024 | 0.3677 | 0.7569 |
| K-Nearest Neighbours | —— | 0.5667 | 0.6566 | —— |
| Random Forest | 0.0536 | 0.0551 | 0.0541 | —— |
| AdaBoost | 0.6566 | 0.6537 | 0.6200 | —— |

**Table 4.** Macro F1-scores for all classical technique combinations.

Although a wide array of techniques was tested, the baseline model still attains the best score out of all the models. We attribute this to the simplicity of the approach – as mentioned in Section 3, violence-related keywords play a major role in separating the two classes, making it easy for a crude, Bag-of-Words based model to reach such scores. However, the finer details are lost in the process of vectorizing the input, thus leaving a big gap when compared to the GTC authors' F1-score of 0.81 for the same training conditions [9].

### 4.3 Error Analysis

Let us now take a look at what exactly are these details that are lost by the classical techniques. For this, we take our best model (Multinomial Naive Bayes

| Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|
| 0.8170 | 0.7143 | 0.8333 | 0.7692 |

**Table 5.** Detailed results for the Multinomial Naive Bayes + Bag of Words classifier.

+ Bag of Words) and run predictions on the whole dataset. We will examine some misclassified examples in order to try to pinpoint the root causes of the classifier errors. Figure 1 highlights two key examples that we break down.

"[...] Q What does the painting reflect the types of **torture** A Mr President a victim but he already passed away [...] He told me that **his nails were pulled** in that way So **he lost some of the fingernails** as the result of this types of **torture** so I painted this image for him as a souvenir So this one **reflects the actual torture done** by the S staff on Oeng Bech Q After you did the painting was Oeng Bech still alive to and whether he receive it"

"[...] I could not hold my tear My tear runs now immediately whenever **I recall the past sorrowful experience** in our family particularly **the tragic death of my children during the Khmer Rouge period** [...]"

**Fig. 1.** Misclassifications from the best classical model. Left: negative example predicted as positive; Right: positive example predicted as negative.

As expected, explicit descriptions of violence are flagged as a part of the positive class by the classifier, which is most likely due to its simple approach of counting words. The left example of Figure 1 contains a conversation between a witness and another person, where the first explains to the second the inspiration behind one of their paintings. Although there are gruesome descriptions of torture, these weren't inflicted on the witness nor did it happen in their presence – they are an account of another man, already deceased. With this example, we can infer that the model does not capture whether or not a violent experience relates directly to the witness.

On the other hand, the second example showcases how vocabulary plays a key role in this classifier. While this example does contain a traumatic violent experience, it is described in a euphemistic way, which avoids the typical explicit lexicon the model usually picks up as positive label indicators.

Summing up, the classifier is mainly looking for specific vocabulary which it associates with one of the two classes. But, more important than that, is what the model is missing: context. As seen through the examples of Figure 1, keywords alone are not enough to capture the positive class – it also depends on who's associated with those words and whether or not that is being described during the hearing.

Therefore, going forward, improved models should be pretrained on domains that enable them to better understand the context of the paragraphs they try to classify. This shifts the problem away from merely finding the violence-related words and their frequency to matching the task with a model pre-trained on similar domains.

## 5 Transformer-Based Models

Language Models are more easily transferable between NLP tasks when their semantic domains are similar [7]. This means that we need to find domains similar to the GTC's. However, the motivation for this work is the fact that the dataset is one of its kind and, thus, so is its domain. Therefore, we explore 3 different options, each with a different motivation.

### 5.1 Baseline

The GTC authors propose a series of models, each with a different train/test combination [9], based on the tribunal the data originated from. We decide to focus on training the model with a train/test split from the whole dataset instead of a subset, as varying this split as the Schirmer et al. [9] did does not contribute to our goals - we are trying to capture relations between the GTC and other pre-train domains, not intra-dataset relations.

Using the whole dataset as well, Schirmer et al. [9] train a $\text{BERT}_{base}$ model for sequence classification with an 80:10:10 train/test/validation split with cross-validation. Such a model attains a macro F1-score of 0.81. We use this model as a baseline comparison for our new improved models.

### 5.2 Pretrained models

**Sentiment Analysis** The first model we use is very similar to the GTC authors' baseline – both are BERT-based. However, this particular transformer is pretrained on a large corpus of multilingual data from a sentiment analysis task. With this model, we aim to tackle false negatives by being more sensitive to the tone used in each paragraph.

**Legal and Administrative Data** Secondly, we try to bring juridic context into the problem – as seen in Figure 1, there is a lot of legal jargon throughout the paragraphs, with many annotations imbued in dialogues and testimonies. We hope that, by being able to distinguish which words belong to passages that were actually spoken out loud and which ones are annotations, this DistilBERT-based model, pretrained on legal and administrative data [4], may achieve better results.

**Not-safe-for-work (NSFW) Internet Content** Finally, we explore once again the violent vocabulary spread throughout the positive class. This last model, also a DistilBERT-based one, is pretrained on data from Reddit. More specifically, this classifier is trained to detect whether or not a certain passage is safe for work (SFW). Not safe for work (NSFW) content includes violence among the potentially disturbing subjects it encapsulates, therefore bringing a promising semantic match for our dataset. We believe that this pretrain model, although very different in form, may bring the context awareness that our classical classifier lacked.

| Model | Base Model | Pre-Train Data | Performance[3] |
|---|---|---|---|
| $\text{BERT}_{Sentiment}$[4] | $\text{BERT}_{Base}$ | Multilingual Product Reviews | 61.0095 (Accuracy[5]) |
| EOIR[6] | DistilBERT | EOIR Privacy dataset [4] | 0.8088 (F1) |
| NSFW Classifier[7] | DistilBERT | Reddit Posts | — |

**Table 6.** Summary for each of the pre-trained models.

## 6 Experimental evaluation

For each model, we train with Huggingface's Trainer API for a maximum of 10 epochs on a 70/30 train/test split and keep the classifier with the best F1 score. All the remaining hyperparameters were kept as default.

### 6.1 Results

We report the detailed results for each of the classifiers in Table 7. Additionally, we also trained a $\text{BERT}_{base}$ model under the same circumstances to provide a better comparison with the model the GTC authors used. The best results were attained by the model pretrained on the NSFW Reddit data, which, in our opinion, bears better semantic similarity with the GTC domain. However, the EOIR [4] model came in as a close second, showing that juridic domain data can also play a key role in handling this complex dataset.

Furthermore, by taking a look at Table 8, we can see that our best model, the NSFW-based one, handles the negative class a lot better than its classical predecessor. Thus, the pretrain domain helps the classifier decide when the

[3] Performance on the original task as reported by the model's author.
[4] `https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment`
[5] Average of accuracy on each language.
[6] `https://huggingface.co/pile-of-law/distilbert-base-uncased-finetuned-eoir_privacy`
[7] `https://huggingface.co/michellejieli/NSFW_text_classifier`

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Classical Baseline | 0.8170 | 0.7143 | 0.8333 | 0.7692 |
| $\text{BERT}_{base}$ | **0.8578** | **0.7584** | 0.8710 | 0.8108 |
| $\text{BERT}_{Sentiment}$ | 0.8330 | 0.6919 | **0.9419** | 0.7978 |
| EOIR | **0.8578** | 0.7447 | 0.9032 | 0.8163 |
| **NSFW Classifier** | 0.8555 | 0.7487 | 0.9068 | **0.8202** |

**Table 7.** Detailed results for our improved models, our previous best model and the GTC's authors' baseline.

violence-related words are applied in the context we are looking for. This is further explored in Section 6.2

Another detail we believe is worth noting is how well the EOIR-pretrained model handles the positive class when compared to the other models while maintaining similar performance for the negative examples.

| **Classical Baseline** | True 1 | True 0 |
|---|---|---|
| Predicted 1 | 151 | 36 |
| Predicted 0 | 18 | 90 |

| **NSFW Classifier** | True 1 | **True 0** |
|---|---|---|
| Predicted 1 | 150 | 37 |
| **Predicted 0** | 5 | **103** |

| **$\text{BERT}_{Sentiment}$** | True 1 | True 0 |
|---|---|---|
| Predicted 1 | 152 | 35 |
| Predicted 0 | 12 | 96 |

| **EOIR** | **True 1** | True 0 |
|---|---|---|
| **Predicted 1** | **160** | 27 |
| Predicted 0 | 15 | 93 |

| **$\text{BERT}_{base}$** | True 1 | True 0 |
|---|---|---|
| Predicted 1 | 142 | 45 |
| Predicted 0 | 9 | 99 |

**Table 8.** Confusion matrices for an evaluation sample comprising 20% of the GTC. Examples are chosen at random.

### 6.2 Error analysis

**NSFW-pretrained Model** Let us take a look at what examples the NSFW-pretrained classifier has misclassified. By looking at the examples in Figure 2, we can see that the errors concern more intricate paragraphs than before.

"[...] **we learned of other people** in [...] who had been **murdered by the police** [...] you say again that you did not see it but heard it you learnt it from other people [...] allow me to say something **about what i experienced myself** because [...] simple answers the question is **whether you saw those killings or** [...]"

"[...] a painter we met only at night briefly so **he told me** that **he was made to be immersed into that bathtub** he didnt say that it was ice water he only told me that it was normal water in that bathtub and **his legs and hands were cuffed and his head were plunged** [...] or **choked with the water** and then they would **kick on his stomach** [...] **working in a constant atmosphere of terror and of fear** is that so a as the prisoners **we lived with constant fear its normal** q my last question [...]"

**Fig. 2.** Misclassifications from the NSFW-pretrained model. Left: negative example predicted as positive; Right: positive example predicted as negative.

In the left example, we have a discussion between a witness and another person questioning him or her. During this exchange, the questioner tries to clarify whether or not the witness saw the murder of other people by the police. The witness then asks to say something about her own experience. While this paragraph contains less violent vocabulary than others, it is still flagged as positive by our NSFW-pretrained model. We believe this to be due to the fact of the witness trying to talk about his or her experience, but this is further exemplified by the second example.

In the right example, there are two testimonies within the same paragraph - a retelling of the torture done to someone else, without the witness being present at the moment of the crime, and a direct account of terror and fear. While the second event clearly states the scarring conditions the witness went through, the example as a whole is predicted as negative.

Now, this paragraph is filled with violence-related vocabulary – which we have established previously, in Table 3, that easily triggers simpler models to evaluate samples as positive. We presume that, despite this, the classifier decides to label this entry as negative due to the clear indication ("he told me") that the first event is a retelling of someone else.

Coupling this information with the previous example, we believe that the model now pays attention to signals of direct accounts instead of merely looking for violence. This means that it can catch more intricate examples, but it still struggles with some confusing paragraphs (for example, it is debatable whether or not the right example in Figure 2 should be labelled as positive).

**EOIR-pretrained Model** Let us now turn our attention to the EOIR-pretrained model. Looking at Figure 3, we can see that the errors differ from the ones of the NSFW-pretrained classifier – they are made on entries that are much more dialogue heavy. We believe that this difference stems from the pretrain domains

"this was happening to the both of yes you other [...] we were near one another [...] q after this **attack** this **rape** how were you feeling physically a i **begged for death** [...] q what happened at this time [...] i a q a q a q a a q [...] where did the two others who came were coming from **attacks** but they were coming in hiding [...] the one who **raped me** where we were [...] did he do the same things [...] yes the same thing [...] **rape** was occurring did this person say anything to you them but [...] what happened at this time [...] q when you say this person **raped you** is a he said [...]"

"[...] witness ev do you recall what day it was [...] i think it was on the th you know that it was a long time ago but i think that it was the the [...] q what day do you recall what day of the week [...] i think it was a saturday night because i was upset because of the shooting and **i was so afraid i heard a lot of shooting so i cana€™t remember very well** but i think it was a saturday and you recall what day [...] that was on a Wednesday [...] but in fact i passed kabuye several times i passed there once and then i came back to kabuye [...] and thursday we came back to kabuye [...]"

**Fig. 3.** Misclassifications from the EOIR-pretrained model. Left: negative example predicted as positive; Right: positive example predicted as negative.

of these models - the NSFW-pretrained model uses data from Reddit, which is much more conversational than the EOIR [4] dataset.

Furthermore, by looking at the left example of Figure 3, we see that its capability of detecting when violence-related keywords are used outside of first-person accounts, when compared to that of the NSFW-based model, isn't as good. This is also reflected in Table 8, where this model reports the second-worst False Negative count, only behind the Classical Baseline model.

## 7 Discussion

As discussed in Section 1, despite the positive results obtained, it is still difficult to pinpoint a pre-trained model as ideal for the GTC.

We have seen that the models pretrained on the juridic and NSFW domains can reach better performance with a very simple training scheme. We have also seen each has its flaws, through Table 8 and Section 6.2.

However, both these models complement each other - one deals well with the negative class, while the other deals well with the positive one. They also bring knowledge from two very different subjects - the Reddit NSFW dataset contains both conversational and violent dimensions within its text while the EOIR [4] dataset brings expertise on the juridic and administrative domain.

Thus, we believe that simultaneously taking into account the pre-train domains of each of these models might help reach even better results. This might be achieved through an ensemble, built from the presented models, or through a single model, pre-trained on both domains.

## 8 Conclusions

We analyze the Genocide Transcript Corpus (GTC) [9] and determine that violence-related words are greatly related to each entry's class, but they are not enough to capture the whole complex labelling process. To better understand and fit the GTC into a semantically similar task, we experiment with various classical NLP techniques and analyze the mistakes they make. We then identify key pretrain domains to which we can approximate the GTC to improve the quality of the predictions. We repeat the error analysis process for our best transformer-based model and discover that, while the classifier now pays attention to the direct account indicators, it still fails to capture some intricate details of the most complex examples.

We believe that combining the two pretrain domains of the best transformer-based models – the NSFW data and the EOIR [4] data – can help improve the results on this particular dataset. This is because the domains of those two datasets are quite different from each other while still being similar to that of the GTC. Future work would involve pretraining similar models ($\text{BERT}_{base}$ or DistilBERT) on the juridic and NSFW domains and fine-tuning them on the GTC.

## References

1. Bender, E.M., Koller, A.: Climbing towards NLU: On meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5185–5198. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.463, https://aclanthology.org/2020.acl-main.463
2. Branco, R., Branco, A., António Rodrigues, J., Silva, J.R.: Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 1504–1521. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.113, https://aclanthology.org/2021.emnlp-main.113
3. Granero Moya, M., Oikonomou Filandras, P.A.: Taking things personally: Third person to first person rephrasing. In: Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI. pp. 1–7. Association for Computational Linguistics, Online (Nov 2021). https://doi.org/10.18653/v1/2021.nlp4convai-1.1, https://aclanthology.org/2021.nlp4convai-1.1
4. Henderson*, P., Krass*, M.S., Zheng, L., Guha, N., Manning, C.D., Jurafsky, D., Ho, D.E.: Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset (2022), https://arxiv.org/abs/2207.00220
5. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1031, https://aclanthology.org/P18-1031

6. Jahan, M.S., Oussalah, M.: A systematic review of hate speech automatic detection using natural language processing. Neurocomputing **546**, 126232 (2023). https://doi.org/https://doi.org/10.1016/j.neucom.2023.126232, https://www.sciencedirect.com/science/article/pii/S0925231223003557
7. Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., Jin, Z.: How transferable are neural networks in NLP applications? In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 479–489. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1046, https://aclanthology.org/D16-1046
8. Phang, J., Févry, T., Bowman, S.R.: Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks (2019)
9. Schirmer, M., Kruschwitz, U., Donabauer, G.: A new dataset for topic-based paragraph classification in genocide-related court transcripts. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 4504–4512. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.479
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf