A New Dataset for Topic-Based Paragraph Classification in Genocide-Related Court Transcripts

Miriam Schirmer[a,b], Udo Kruschwitz[b], Gregor Donabauer[b]

†University of Regensburg, Germany
‡TUM School of Social Science and Technology, Technical University of Munich, Germany
{miriam.schirmer, udo.kruschwitz, gregor.donabauer}@ur.de

Genocide Transcript Corpus

Recent progress in natural language processing has been impressive in many different areas with transformer-based approaches setting new benchmarks for a wide range of applications. This development has also lowered the barriers for people outside the NLP community to tap into the tools and resources applied to domain-specific applications. The bottleneck however still remains the lack of annotated gold-standard collections as soon as one's research or professional interest falls outside the scope of what is readily available. One such area is genocide-related research, also including the work of experts who have a professional interest in accessing, exploring and searching large-scale document collections on the topic, such as lawyers.

We present GTC (Genocide Transcript Corpus), the first annotated corpus of genocide-related court transcripts which serves three purposes:

(1) to provide a first reference corpus for the community,
(2) to establish benchmark performances (using state-of-the-art transformer-based approaches) for the new classification task of paragraph identification of violence-related witness statements,
(3) to explore first steps towards transfer learning within the domain.

BERT Text Classification

In order to assess the utility of the corpus and the difficulty of the underlying classification task we will adopt the commonly applied baseline approach of fine-tuning BERT. One of the goals is to show whether or not BERT also serves as an efficient tool for this type of text data and whether it can help simplify classification of paragraphs in court data.

Material & Label Annotation

The dataset used in this study consists of 1475 text passages from three different genocide tribunals. Transcripts from the three biggest ad-hoc genocide tribunals, the Extraordinary Chambers in the Courts of Cambodia (ECCC), the International Criminal Tribunal for Rwanda (ICTR), and the International Criminal Tribunal for the former Yugoslavia (ICTY) were selected. All transcripts are publicly available online on the respective courts' websites.

Between 4 and 7 transcripts were selected per tribunal and divided into equally large text chunks of 250 words each. In the final dataset, the number of samples is roughly equally distributed across tribunals (ECCC: 465, ICTY: 530, ICTR: 480).

Identifying Violence-Related Paragraphs

The paragraph labeling is aimed at identifying those parts of the text that refer to violence experienced by witnesses. All samples were therefore labeled according to whether they contain a witness's description of experienced violence (0 = no violence, 1 = violence). Violence in this context is interpreted broadly and includes accounts of experienced or directly witnessed torture, interrogation, death, beating, psychological violence, experienced military attacks, destruction of villages, looting, and forced displacement.

To ensure that the categorization is reliable, a random selection of approximately 200 samples were independently labeled by a second researcher (with an inter-rater reliability κ = 0.86) according to the above-mentioned facets of experienced violence.

Label 0

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.

A. We were living in Taba commune.

Q. Is that in Rwanda?

A. It's a commune in Rwanda, in Gitarama prefecture.

Q. Around the beginning of April did you ever receive news of the crash of the president's plane?

A. Yes, I heard this. […]

ICTR-96-4-I, October 23rd 1997, p. 17-18.

Label 1

Q. What happened next?

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 bureau communal, 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 this 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. […]

ICTR-96-4-I, October 23rd 1997, p. 60.

Sample abstracts from the corpus demonstrating two clear-cut examples for a text passage that does not contain accounts of violence in a witness statement (top example – Label 0) and one that does (bottom example – Label 1). The examples were shortened, and both format and punctuation were adapted for readability.

Experimental Setup

For all experiments, the 12-layer $BERT_{base}$ architecture for sequence classification (Devlin et al., 2019) was used to classify text passages of genocide tribunal transcripts. The dataset consists of 3 subsets with data from different tribunals. 5-fold crossvalidation was applied to each subset and to the full version of the dataset.

In a first step, BERT was trained on the full dataset to classify samples of all three tribunals, but also to classify tribunal-specific text chunks. Secondly, we applied the same setup to all three subsets. More specifically, training was performed using tribunal-specific samples to see if BERT is still able to predict class labels of both the mixed dataset (excluding training class), as well as the remaining tribunal-specific subsets. Overall, BERT was trained on all possible train, validate and test constellations, leading to a total of 16 different combinations.

For training and validation a batch-size of 16 samples and an epoch-number of 3 was used. Precision, recall, micro and macro F1 scores for each train/validate/test constellation are provided, with macro F1 scores being the reference score when comparing results.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

d'Sa, A. G., Illina, I., and Fohr, D. (2020). BERT and fastText embeddings for automatic detection of toxic speech. In 2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies"(OCTA), pages 1–5. IEEE.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. (2021). Revisiting few-sample BERT fine-tuning. In International Conference on Learning Representations, Vienna, Austria.

Results & Discussion

Our results show that a binary classification based on BERT yields very reliable results across text data from different tribunals. A macro F1 score of 0.81 when training, testing and validating with the complete, mixed dataset that includes all three tribunals shows that BERT can be applied to this type of data and provides reasonably good predictions across the different subsets. Considering the individual tribunals, using a tribunal-specific dataset for training and validating provided varying test results (ECCC-ECCC macro F1=0.70; ICTY-ICTY macro F1=0.68; ICTR-ICTR macro F1=0.80).

Overall, using the mixed dataset for training and validating resulted in the highest F1 scores throughout the tribunal variations (min macro F1=0.78, max macro F1=0.85), independently of the dataset that was used for testing. The highest individual F1 score in our experiments was obtained when predicting data from ICTR transcripts with trained and validated data from the mixed dataset ("ALL") (macro F1=0.85).

	ALL				ECCC				ICTY				ICTR			
	P	R	mac. F1	mic. F1	P	R	mac. F1	mic. F1	P	R	mac. F1	mic. F1	P	R	mac. F1	mic. F1
ALL	0.81	0.83	0.81	0.83	0.81	0.82	0.82	0.82	0.78	0.78	0.78	0.83	0.85	0.85	0.85	0.85
ECCC	0.77	0.77	0.77	0.79	0.77	0.72	0.70	0.75	0.73	0.71	0.71	0.78	0.81	0.79	0.79	0.80
ICTY	0.77	0.78	0.77	0.78	0.77	0.78	0.77	0.78	0.70	0.73	0.68	0.74	0.81	0.81	0.81	0.81
ICTR	0.74	0.74	0.74	0.78	0.79	0.77	0.78	0.79	0.69	0.74	0.70	0.75	0.83	0.78	0.80	0.85

Results for macro precision (P), macro recall (R) and macro/micro F1 scores on test data (columns) with respect to different training/evaluation set (rows) combinations.

The results indicate that the mixed dataset is most successful when predicting if a certain text passage from one of three genocide tribunals contains accounts of experienced violence by a witness. The ballpark figures we obtained are comparable to state-of-the-art (BERT-based) performance on some other commonly used binary classifications such as MRPC (Zhang et al., 2021), but fall short of performance levels expected for other settings (d'Sa et al., 2020).

Overall, this dataset has the potential of serving as a basis for a variety of research approaches in the field of genocide research in the future. For example, more in-depth comparisons between linguistic or content-based characteristics between the three tribunals could be made, building bridges between the interdisciplinary field of genocide research and NLP approaches. From an NLP perspective, next steps could include further fine-tuning of BERT and conducting the experiments with more recent transformer architectures or machine learning techniques.

Data and code are available to the community: https://github.com/MiriamSchirmer/genocide-transcript-corpus

LREC 2022