LMU at HaSpeeDe3: Multi-Dataset Training for Cross-Domain Hate Speech Detection

Viktor Hangya[1,2,*], Alexander Fraser[1,2]

[1] *Center for Information and Language Processing, LMU Munich*
[2] *Munich Center for Machine Learning*

Abstract

We describe LMU Munich's hate speech detection system for participating in the cross-domain track of the HaSpeeDe3 shared task at EVALITA 2023. The task focuses on the politics and religion domains, having no in-domain training data for the latter. Our submission combines multiple training sets from various domains in a multitask prompt-training system. We experimented with both Italian and English source datasets as well as monolingual Italian and multilingual pre-trained language models. We found that the Italian out-of-domain datasets are the most influential on the performance in the test domains and that combining both monolingual and multilingual language models using an ensemble gives the best results. Our system ranked second in both domains.

Keywords

hate speech detection, multitask learning, prompt-training

1. Introduction

Due to the sheer amount of social media content, manual filtering for hate speech is impossible which makes building high performance and reliable hate speech classifiers important. To promote research in the field various datasets were built [1, 2], and shared tasks were organized [3, 4, 5], where the best performing systems are based on pre-trained language models (PLMs) [6, 7].

The *HaSpeeDe3* shared task [8] is the third iteration of the series on hate speech detection in Italian social media posts (tweets) organized at EVALITA 2023 [9], focusing on strongly polarized debates in political and religious topics. Two subtasks were organized: Task A – *Political Hate Speech Detection* which on top of textual inputs, allows for the use of contextual information, such as metadata of tweets and authors. Task B – *Cross-domain Hate Speech Detection* involves only textual inputs, however the main objective is to explore cross-domain hate speech detection in the politics and religious domains, by allowing the use of external datasets (open track). In contrast to the politics domain where in-domain training data is given, in the religious domain such data is not provided. Our team participated only in Task B.

Cross-domain training is a crucial problem in machine learning, aiming to build high quality models for the target domain by leveraging labeled samples from out-of-domain sources as well [10]. For hate speech detection, [11] experimented with training classifiers using out-of-domain training examples and showed a significant performance drop on the test sets compared to in-domain training. By simply combining multiple datasets of different domains, including the target domain, they achieved only slight improvements. In a similar work, one source- and one target-domain were explored [12], but the authors showed mixed results, i.e., improvements on some domains but decrease on others. Similarly, [13] applied the general domain adaptation technique of [10] and showed improvements when incorporating some out-of-domain datasets into the final model, even though the approach seemed sensitive to the chosen out-of-domain dataset. In addition, [14] showed negative performance on the target-domain in German by using additional source-domain English training examples.

Following previous work, we rely on transfer learning to leverage out-of-domain (external) datasets to build our classifiers for the political and religious domains. We experiment with various external datasets containing both Italian and English hate speech inputs. Additionally, in contrast to previous work which used datasets with matching label sets, we use corpora annotated with different label sets, e.g., stereotype. To avoid negative results, we combine multiple datasets in a multitask training fashion in order to build robust systems. Additionally, we train our systems in a two-step process, where we first pre-finetune our models on the external datasets, followed by fine-tuning them to the target task. As the basis of our models, we take various PLMs based on the BERT [15] and RoBERTa [16] architectures, including

The Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), September 07–08, 2023, Parma, Italy

[*] Corresponding author.

hangyav@cis.lmu.de (V. Hangya); fraser@cis.lmu.de (A. Fraser)
https://www.cis.uni-muenchen.de/~hangyav (V. Hangya);
https://www.cis.uni-muenchen.de/~fraser (A. Fraser)
0000-0002-5144-3069 (V. Hangya); 0000-0003-4891-682X (A. Fraser)

both Italian only and multilingual models. Furthermore, in order to facilitate information sharing across the used datasets, we perform prompt-training which eliminates dedicated classification heads for each dataset.

Our experiments show that using only Italian external datasets is more beneficial compared to leveraging English as well. In contrast, we find that both monolingual and multilingual PLMs perform comparably well, and that they can support each other when combining them using model ensembling.[1]

2. Approach

Our approach consists of two steps where we first pre-finetune a given PLM on external datasets (see Section 3.1), followed by in-domain fine-tuning in case of the political domain where such data is provided. Instead of classification heads, we leverage model prompting.

Prompt-Training Prompt-training was shown to be effective and more reliable for various NLP tasks, including classification [17]. Instead of using classification heads on top of PLMs which add additional parameters to the model, it relies on the masked language modeling task (MLM). Using pattern-verbalizer-pairs (PVPs), an input sentence is first transformed using the pattern, e.g., *I hate you.* $\rightarrow$ *Is this hate speech? I hate you. [MASK]*, and the task is to predict the masked token. Finally, the verbalizer maps the highest probability token, out of a set of valid tokens, to labels of a given dataset, e.g., *Yes* $\rightarrow$ *Hate* or *No* $\rightarrow$ *nonHate*. During training all model parameters are fine-tuned using the MLM objective.

Step 1 Given a set of external training corpora ($D_E = \{D_{e_i} : i = 1..N\}$), we randomly select a single dataset D_{e_i} and a batch of samples from it in each training step. For each dataset we apply a dedicated PVP (see Section 3.2) in order to handle datasets of different label sets, and use cross-entropy loss to perform a single model update. This way we mix the available external datasets during pre-finetuning instead of performing a sequential model update which could lead to catastrophic forgetting. Additionally, we make sure that we exhaust all datasets in D_E in each epoch, i.e., the model is trained on each input sample once per epoch.

Step 2 In case of the political test domain, we apply a second round of model fine-tuning given the in-domain training dataset. We follow the same training procedure as in Step 1 but using only a single training corpus instead of multiple corpora. The goal of this step is to specialize our model to the target domain, given the pre-finetuned base model which is already aware of general hate speech language phenomena. This is in contrast to standard multitask training where i) the goal is to build a single model supporting multiple target tasks (datasets) and more importantly ii) which is trained by optimizing a joint objective function across all datasets.

No in-domain training data was provided for the religious test datasets. In this case, we omit Step 2 and apply our model resulted from Step 1 in a zero-shot transfer learning fashion, i.e., the model is only trained on the external (source) datasets but not on the target corpus.

Ensembling To further improve the robustness of our final models, we employ model ensembling to combine the output of multiple models. We ensemble models in two dimensions: we combine models of the same setup but using 3 different random seeds, and models based on different PLM architectures as defined below. We simply take the mean of the probabilities of the considered models for a given input sample.

3. Experiments

3.1. Datasets

Next, we list our external dataset setups followed by the introduction of the official shared task data. We define the following groups of external datasets:

HaSpeeDe We leverage Italian datasets from previous HaSpeeDe iterations. More precisely, we take i) the training data containing 2 400 Facebook posts annotated with binary hate speech labels from HaSpeeDe1 [18], ii) 5 470 binary hate speech annotated Twitter posts from HaSpeeDe2 [5] and iii) the same Twitter posts but annotated for binary stereotype detection.

It Additionally, to the datasets mentioned in the HaSpeeDe set, we used further Italian abusive language related datasets. Tweets from the AMI18 misogyny detection shared task [19]: i) 3 200 binary and ii) 1 460 fine-grained (discredit, stereotype, dominance, harassment, derailing) training sets as well as iii) 1 454 binary target detection set (individual, group). Furthermore, we took binary iv) hate (3 271) and v) stereotype (441) annotated training sets from the IHSC corpus [20] containing tweets related to immigrants.

Mixed Finally, to test the effect of leveraging English training data as well, in addition to the datasets contained in the HaSpeeDe set we used 7 078 politics related tweets annotated for binary hate speech detection released in [21].

[1] Our code is available at https://cistern.cis.lmu.de/multi_hs

Table 1

PVPs. For each pattern the input sentence in depicted as **X**. For the verbalizers the left-hand side of → indicates the predicted tokens by the PLMs which are assigned to the label on the right-hand side. We note that we used English to Italian machine translation to build the Italian PVPs.

	Pattern
$p1$	**X** → Questo è un discorso di odio? **X** [MASK]
$p2$	**X** → Is this hate speech? **X** [MASK]
$p3$	**X** → Questo è stereotipato? **X** [MASK]
$p4$	**X** → **X** Era [MASK]
$p5$	**X** → **X** Era preso di mira [MASK]
	Verbalizer
$v1$	*Sì* → *Hate*; *No*→ *nonHate*
$v2$	*Yes* → *Hate*; *No*→ *nonHate*
$v3$	*stereotipato* → *stereotype*; *predominante* → *dominance*; *deragiante* → *derailing*; *molesto* → *sexual_harassment*; *screditante* → *discredit*
$v4$	*individuale* → *individual*; *gruppo* → *group*

Official HaSpeeDe3 datasets The HaSpeeDe3 shared task focuses on strongly polarized debates in two domains. For the politics domain, the binary hate speech labeled PolicyCorpusXL was made [22], containing 5 600 train and 1 400 test tweets. In the religious domain, the ReligiousHate [23] corpus contains 3 000 test tweets and no training set.

3.2. Setup

PVPs We aimed at keeping our used patterns and verbalizers simple and uniform across datasets. Both patterns and verbs are presented in Table 1. For binary hate and misogyny datasets we used patterns $p1$ and $p2$ for the Italian and English datasets respectively. Similarly, we used $p3$ for the binary stereotype datasets. As verbalizers, we used $v1$ and $v2$ for the two languages. For the AMI18 misogyny fine-grained and target sets we used patterns number $p4$ and $p5$ respectively, with verbalizers $v3$ and $v4$.

Models As the base PLMs we experiment with two monolingual Italian and two multilingual models. AlBERTo was trained purely on Italian social media texts (Twitter in particular), based on the BERT base architecture [24]. We selected this model since it performs well on social media texts. Similarly, we experiment with UmBERTo [25] which is based on the RoBERTa base architecture, and was trained with whole word masking on Italian CommonCrawl corpus. As for the multilingual models, we used the highly popular mBERT [15] and XLM-R [16] PLMs.

We used the OpenPrompt toolkit for implementation [26], and used standard hyperparameter values. Due to memory limitation of Nvidia GTX 1080 Ti however, we used batch size 4 with gradient accumulation steps 4 for BERT based models, while we used batch size 1 with gradient accumulation steps 16 for RoBERTa based models. We train our models for a single epoch in Step 1 of our approach, while we perform early stopping in Step 2 based on the performance on the development set. During the development of our system, we split the official political training set to train/dev/test splits. Since no labeled sets were provided for the religious domain for development, we simulated zero-shot transfer experiments on the politics domain.

Preprocessing We also experimented with two sets of data manipulation methods. To clean tweets, we applied standard Twitter preprocessing steps: user mention and hashtag removal, HTML and repeated character unification. Since hate speech datasets often suffer from label imbalance, we tested random oversampling, class weighting and focal loss. However, none of these approaches led to consistent improvements, thus we omitted these steps from our final systems.

4. Results

We evaluate our systems using macro averaged F_1 scores as it is the official score used in the shared task. First, we present the comparison of various external dataset setups (Table 2), followed by the comparison of different PLMs and their combination with ensembling (Table 3). Finally, we present our official results in Table 4.

External datasets As the baseline system to measure the effectiveness of the external datasets, we only perform Step 2 of our approach, i.e., we fine-tune the off-the-shelf PLM (mBERT) using only the HaSpeeDe3 politics training corpus without any pre-finetuning steps on the external datasets. As mentioned, no in-domain data is provided for the religious domain, thus we perform zero-shot transfer learning, i.e., we only perform pre-finetuning on the external datasets. Additionally, since not even a development set was provided for this domain, we simulate zero-shot transfer on the politics dataset. The gold labels of the religious test set were released after the shared task deadline, thus we are able present (oracle) results for comparison. The results in Table 2 show the positive impact of the external datasets, as the baseline systems were outperformed by a large margin. Comparing the different external dataset setups, we found that they perform comparably. On the politics domain the HaSpeeDe setup performed the best, although both It and Mixed lagged behind with less than half a percentage point in the two-step setting, while on the simulated zero-shot experiments the gap between HaSpeeDe and

Table 2
Macro F_1 scores (%) comparing different external dataset setups using mBERT as the base PLM. The baseline system uses the HaSpeede3 training dataset only, while the HaSpeeDe, It and Mixed incorporate the external datasets as well. Pol. depicts the results of our systems for the politics domain, zero Pol. our zero-shot experiments on the politics domain simulating the missing train set of the religious domain and gold Rel. shows results on the gold religious test set after its release. We bold the **best** setup.

	Pol.	zero Pol.	gold Rel.
baseline	84.48	-	52.34
HaSpeeDe	**86.61**	**63.99**	62.02
It	86.43	61.82	61.69
Mixed	86.21	-	**62.36**

Table 3
Macro F_1 scores (%) comparing different monolingual and multilingual PLMs, as well as model ensembles (mono: AlBERTo and UmBERTo; mix: AlBERTo, UmBERTo, mBERT and XLM-R). We use the HaSpeeDe external dataset setup. We highlight the best individual and **ensemble** models.

	Pol.	zero Pol.	gold Rel.
AlBERTo	89.92	63.00	64.16
UmBERTo	88.77	62.16	62.81
mBERT	86.61	63.99	62.02
XLM-R	86.21	55.45	61.98
mono-ens.	**91.44**	**61.57**	64.58
mix-ens.	90.95	60.71	**64.61**

It[2] is around 2 percentage points. These findings indicate that the misogyny detection tasks in the It setup could be slightly detrimental to the binary hate speech detection task. Furthermore, the additional English politics related dataset in the Mixed setup does not lead to further improvements on the politics domain, although they are from the same domain, indicating that leveraging only Italian external datasets is an important factor. Looking at the results on the gold religious test set, we found similar trends. The use of additional training datasets on top of the HaSpeeDe3 politics training set improves the performance[3]. Although the HaSpeeDe set performs well, interestingly the best performance was achieved by Mixed which includes English politics tweets, which needs further investigations. Nonetheless, based on these findings, we used the HaSpeeDe setup in our final system submission and in the following experiments.

Table 4
Our final results as reported by the shared task organizers.

	XPoliticalHate	XReligiousHate
Run 1 (mono-ens.)	**90.14**	64.58
Run 2 (mix-ens.)	89.84	**64.61**

Model variations In Table 3 we compare the mentioned 4 PLMs and their combinations. In the mono-ens. ensemble setup we combine the monolingual Italian models (AlBERTo and UmBERTo), while in mix-ens. all PLMs (AlBERTo, UmBERTo, mBERT and XLM-R). We found that the monolingual models outperform multilingual models in most cases, especially on the politics domain. AlBERTo has the best performance on average which is due to its pre-training on social media content. Interestingly, comparing BERT (AlBERTo and mBERT) and RoBERTa (UmBERTo and XLM-R) architectures, the former outperform the latter, which is a somewhat contradictory result as the latter often performs better. The ensemble results, however, show that although the results of different PLMs vary, they can support each other and by ensembling their outputs the performance can be further increased. Similarly, as for the individual models, the monolingual ensemble performed the best during our system development, however the combination of all models does not lag much behind. Furthermore, mix-ens. outperformed mono-ens. on the gold religious test set.

Final Submission The shared task allowed two submitted runs for each domain. Based on our findings during development, our official systems were mono-ens. (Run 1) and mix-ens. (Run 2) using the HaSpeeDe external dataset setup. We note that in the case of the religious domain, we also include the HaSpeeDe3 politics training set as an external dataset. Our official results are shown in Table 4. We achieved the second-best result in both domains.

5. Conclusions

We presented the LMU Munich team's systems at the HaSpeeDe3 shared task, participating in the cross-domain hate speech detection task. Our approach involves a two-step method for the politics domain: pre-finetuning using external datasets followed by a second step of fine-tuning on the target domain. In case of the religious domain, we used a zero-shot transfer setup involving training on the external datasets only. Additionally, we performed prompt-training instead of the use of classification heads in order for a more seamless combination of external datasets of different label sets. By comparing various external datasets, including both Italian and English, we found that Italian datasets are

[2] Due to the inclusion of politics related training data in the baseline and Mixed setups, these are not applicable in the simulated zero-shot case.

[3] Note that we also included the politics HaSpeeDe3 train set in the HaSpeeDe, It and Mixed sets when training our models for the religious domain.

more beneficial. Similarly, by comparing various PLMs we found that individually monolingual models perform better than multilingual models. On the other hand, combining multiple PLMs with model ensemble, we found that different models can support each other leading to improved performance. Our best result on the political domain was achieved by combining monolingual PLMs only, while combining all PLMs performed the best on the religious domain.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback and the Cambridge LMU Strategic Partnership for funding for this project.[4] The work was additionally supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (No. 640550) and by the German Research Foundation (DFG; grant FR 2829/4-1).

References

[1] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013.

[2] O. de Gibert, N. Perez, A. García-Pablos, M. Cuadros, Hate speech dataset from a white supremacy forum, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2018, pp. 11–20. URL: https://aclanthology.org/W18-5102.

[3] T. Mandl, S. Modha, P. Majumder, D. Patel, M. o. Dave, C. Mandlia, A. Patel, Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages, in: 11th meeting of the Forum for Information Retrieval Evaluation, 2019, pp. 14–17. URL: http://ceur-ws.org/Vol-2517/T3-1.pdf.

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007.

[5] M. Sanguinetti, C. Gloria, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, C. Tommaso, V. Patti, R. Irene, et al., Haspeede2@ evalita2020: Overview of the evalita 2020 hate speech detection task, in: Proc. EVALITA, 2020, pp. 1–9.

[6] E. Lavergne, R. Saini, G. Kovács, K. Murphy, Thenorth@ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection, in: Proc. EVALITA, volume 2765, 2020.

[7] T. Tran, Y. Hu, C. Hu, K. Yen, F. Tan, K. Lee, S. R. Park, HABERTOR: An efficient and effective deep hatespeech detector, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7486–7502. URL: https://aclanthology.org/2020.emnlp-main.606.

[8] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), 2023.

[9] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), 2023.

[10] H. Daumé, Frustratingly Easy Domain Adaptation, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 256–263. URL: https://www.aclweb.org/anthology/P07-1033.

[11] G. Glavaš, M. Karan, I. Vulić, XHate-999: Analyzing and Detecting Abusive Language Across Domains and Languages, in: Proceedings of the 28th International Conference on Computational Linguistics, 2021, pp. 6350–6365. URL: https://www.aclweb.org/anthology/2020.coling-main.559/.

[12] M. A. Rizoiu, T. Wang, G. Ferraro, H. Suominen, Transfer learning for hate speech detection in social media, arXiv:1906.03829 (2019). URL: https://arxiv.org/pdf/1906.03829.pdf.

[13] M. Karan, J. Šnajder, Cross-Domain Detection of Abusive Language Online, in: Proceedings of the 2nd Workshop on Abusive Language Online, 2018, pp. 132–137. URL: https://www.aclweb.org/anthology/W18-5117.

[14] M. Wiegand, A. Amann, T. Anikina, A. Azoidou, A. Borisenkov, K. Kolmorgen, I. Kröger, C. Schäfer, Saarland University's Participation in the GermEval Task 2018 (UdSW) – Examining Different Types of Classifiers and Features, in: Conference on Natural Language Processing (KONVENS 2018), 2018, pp. 21–26. URL: https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/GermEval2018_Proceedings.pdf#page=27.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for

[4] https://www.cambridge.uni-muenchen.de

Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423.pdf.

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://www.aclweb.org/anthology/2020.acl-main.747/.

[17] T. Schick, H. Schütze, Exploiting cloze-questions for few-shot text classification and natural language inference, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 255–269. URL: https://aclanthology.org/2021.eacl-main.20.

[18] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, et al., Overview of the evalita 2018 hate speech detection task, in: Ceur workshop proceedings, 2018, pp. 1–9.

[19] E. Fersini, D. Nozza, P. Rosso, Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI), Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2018) 59–66. URL: https://pdfs.semanticscholar.org/05d5/17f3fa5f47b16265b378c81a0839ed760ba0.pdf.

[20] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An Italian Twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018. URL: https://aclanthology.org/L18-1443.

[21] C. Toraman, F. Şahinuç, E. Yilmaz, Large-scale hate speech detection with cross-domain transfer, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 2215–2225. URL: https://aclanthology.org/2022.lrec-1.238.

[22] F. Celli, M. Lai, A. Duzha, C. Bosco, V. Patti, Policycorpus XL: An Italian Corpus for the Detection of Hate Speech Against Politics., in: In Proceedings of the Eighth Italian Conference on Computational Linguistics, 2021. URL: https://ceur-ws.org/Vol-3033/paper38.pdf.

[23] A. Ramponi, B. Testa, S. Tonelli, E. Jezek, Addressing religious hate online: from taxonomy creation to automated detection, PeerJ Computer Science 8 (2022) e1128. URL: https://peerj.com/articles/cs-1128.

[24] M. Polignano, V. Basile, P. Basile, M. de Gemmis, G. Semeraro, Alberto: Modeling italian social media language with bert, IJCoL. Italian Journal of Computational Linguistics 5 (2019) 11–31. URL: https://journals.openedition.org/ijcol/472.

[25] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[26] N. Ding, S. Hu, W. Zhao, Y. Chen, Z. Liu, H. Zheng, M. Sun, OpenPrompt: An open-source framework for prompt-learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2022, pp. 105–113. URL: https://aclanthology.org/2022.acl-demo.10. doi:10.18653/v1/2022.acl-demo.10.