NamDang at MultiPride@EVALITA 2026: Multilingual Classification of Reclaimed Language in LGBTQ+ Discourse using Transformer-based Models

Nam Dang[1,2], Vo Tuan Kiet[1,2] and Dang Van Thin[1,2]

[1]University of Information Technology-VNUHCM, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper describes the system developed for Task A of the MultiPride shared-task at EVALITA 2026. The task focuses on the identification of reclaimed language used by the LGBTQ+ community, where potentially offensive terms are reappropriated and used as expressions of identity and pride rather than discrimination. This task is particularly challenging due to limited and imbalanced data, as well as the semantic ambiguity between reclaimed and insulting language, where similar derogatory terms may appear in both labels. To address this problem, we adopt a transfer learning approach based on the power of pre-trained multilingual language models. We fine-tune the XLM-R and mDeBERTa models on a multilingual training dataset that combines three languages: English, Italian, and Spanish. Additionally, we investigate the impact of data augmentation by comparing models trained with and without augmented data. The experimental results demonstrate that our approach achieves competitive performance compared with other baselines and methods in the overall MultiPride ranking.

Warning for Explicit Content: *This paper contains examples of language that may be offensive, including slurs and derogatory terms related to the LGBTQ+ community. These examples are included solely for the purpose of scientific analysis.*

Keywords

Multilingual NLP, Pre-trained Language Models, Text Classification, MultiPride

1. Introduction

The automatic detection of abusive, offensive, and discriminatory language has become an important research topic in Natural Language Processing, particularly in the context of online platforms and social media[1, 2]. Language used in these environments often reflects broader societal issues such as sexism, homophobia, and discrimination against marginalized communities, motivating the development of computational methods for identifying harmful content.

Among these challenges, language targeting the LGBTQ+ community[3] presents specific complexities. Members of the community are frequently subjected to hate speech and derogatory expressions, which can reinforce social exclusion and discrimination. At the same time, LGBTQ+ speakers may intentionally reappropriate slurs and offensive terms as reclaimed language, using them to express identity, solidarity, or pride. This phenomenon of reclamation has been widely discussed in sociolinguistic literature[4] and highlights the importance of context and speaker intent in language interpretation.

From a computational perspective, reclaimed language poses a significant challenge for text classification systems. The same lexical items may appear in both insulting and self-referential contexts, making it difficult to rely on surface-level features alone. Distinguishing between discriminatory and reclaimed uses therefore requires models to capture contextual, semantic, and pragmatic information beyond individual words[5, 6].

The MultiPride shared task at EVALITA 2026[7] directly addresses this problem by focusing on the identification of reclaimed language related to the LGBTQ+ community. In Task A, systems are required to distinguish between reclaimed and offensive uses of potentially derogatory expressions. The task

EVALITA 2026: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Bari, Italy, February 2026

24521098@gm.uit.edu.vn (N. Dang); kietvt@uit.edu.vn (V. T. Kiet); thindv@uit.edu.vn (D. V. Thin)

involves multilingual data in English, Italian, and Spanish, and is characterized by limited training data and class imbalance, further increasing the difficulty of the classification problem.

In this paper, we present the system developed for Task A of the MultiPride shared task. Our approach is based on fine-tuning pre-trained transformer-based language models[8], which have demonstrated strong performance across a wide range of Natural Language Processing tasks. Specifically, we employ multilingual transformer models to leverage shared representations across languages and improve generalization in low-resource settings. We experiment with XLM-RoBERTa-base[9] and mDeBERTa-v3-base[10], training the models on a combined multilingual dataset and evaluating the impact of data augmentation strategies[11].

2. Methodology

The experimental framework of this study is structured around two distinct experimental runs, both utilizing the datasets described in Section 2.

- ***run-1***: utilizes the augmented dataset containing synthetic examples to evaluate the impact of data expansion
- ***run-2***: serves as a baseline, trained exclusively on the original MultiPride dataset without augmentation.

To address the classification task, this research employs the transformer-based architecture provided by the Hugging Face ecosystem[12]. Transformer models are particularly suited for text classification due to their self-attention mechanisms, which capture long-range contextual dependencies across diverse linguistic structures. The selection of Hugging Face transformers ensures a robust, scalable, and reproducible framework, leveraging state-of-the-art pretrained weights that have demonstrated significant efficacy in cross-lingual transfer learning.

2.1. Models

The methodology incorporates two distinct pretrained models, selected for their specific architectural strengths and suitability for the experimental objectives.

This study builds upon two pretrained Transformer-based architectures, RoBERTa[13] and DeBERTa[14], which differ primarily in how positional information is modeled within self-attention. RoBERTa follows a standard encoder-only Transformer design in which token embeddings and absolute positional embeddings are summed at the input layer and then passed through stacked self-attention and feed-forward layers. In contrast, DeBERTa introduces a disentangled attention mechanism that represents content and relative positional information separately, allowing them to interact explicitly within the attention computation. By decoupling token semantics from positional encoding, DeBERTa provides a more expressive mechanism for modeling word order and contextual relationships, motivating its inclusion alongside RoBERTa for a comparative evaluation of positional modeling strategies.

To support a controlled comparison in the multilingual setting, we evaluate XLM-RoBERTa (XLM-R)[9] and mDeBERTa-v3[10]—multilingual extensions of RoBERTa and DeBERTa—under two independent experimental runs. ***run-1*** is conducted on the augmented dataset, while ***run-2*** uses the original MultiPride dataset. In each run, the two models are trained and evaluated in parallel, and a single best-performing model is selected independently for that run based on validation performance, rather than assuming consistent superiority of one architecture across data conditions. The model chosen in ***run-1*** is thus the one that empirically demonstrates greater robustness when exposed to augmented data, potentially reflecting architectural properties that better accommodate increased diversity or noise. Conversely, the model selected in ***run-2*** is the one that more effectively captures the characteristics of the original data distribution, favoring precise contextual or positional representations. Importantly, these interpretations are grounded solely in observed validation outcomes, and architectural differences are discussed only to contextualize empirical results, without imposing prior assumptions about model advantage.

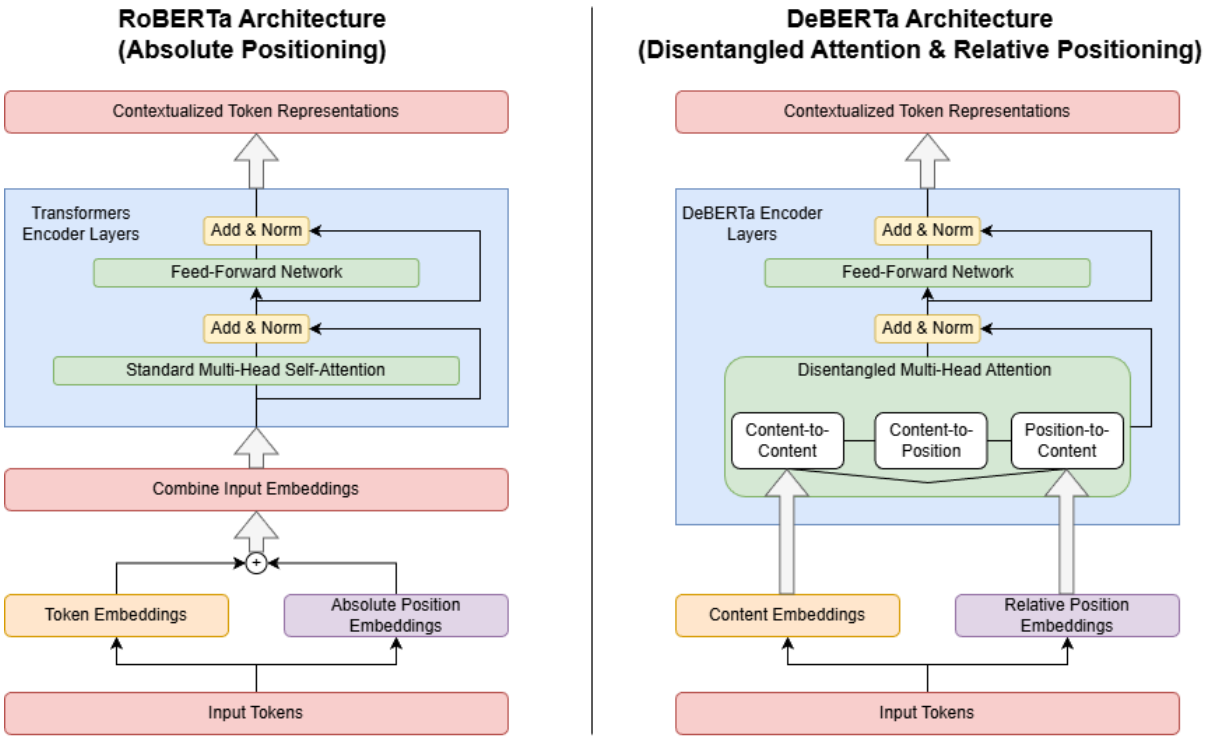

Figure 1: Structural comparison between RoBERTa and DeBERTa. Unlike RoBERTa, which sums absolute position and token embeddings at the input, DeBERTa processes content and relative positions independently through a disentangled attention mechanism.

2.2. Fine-tuning Approach

To prepare for the training process, we processed raw textual data collected from social media through a dedicated cleaning pipeline designed to remove noisy elements while preserving the contextual information necessary for effective modeling. Specifically, we normalized URL links and user mentions into placeholder tokens URL and @USER, respectively, while intentionally retaining hashtags in their original form to preserve salient topical cues. An empirical analysis of the sequence length distribution revealed that 95% of the samples contained fewer than 92 tokens; accordingly, we set the maximum sequence length (MAX_LENGTH) to 128 tokens. This choice ensures coverage of the vast majority of textual variations while maintaining computational efficiency.

To address the class imbalance (particularly in ***run-2***), we implemented a custom error-penalization mechanism based on a weighted loss function. We computed class weights from the label frequencies using the `compute_class_weight` function from the `scikit-learn` library[15], thereby assigning greater importance to the minority class during optimization. In addition, we developed a custom `WeightedTrainer`, inheriting from the Hugging Face `Trainer`, to replace the default loss function with a Weighted Cross-Entropy Loss. This design explicitly encourages the model to learn more discriminative features for the minority class while reducing its tendency to bias predictions toward the. The loss function for a single sample is defined as:

$$\mathcal{L} = -\sum_{i=0}^{C-1} w_i \cdot y_i \log(\hat{y}_i) \tag{1}$$

Where:

- C: The total number of target classes, which is 2 for Task A (Reclaimed vs. Non-reclaimed).
- y_i: The ground truth binary indicator for class i.
- $\hat{y}_i$: The probability for class i predicted by the model (after the Softmax layer).

- w_i: The penalty weight assigned to class i to balance the gradient updates.

The class weights w_i were computed using the inverse frequency method via the `scikit-learn` library to ensure that the minority positive class receives higher importance during optimization:

$$w_i = \frac{N}{C \cdot n_i} \tag{2}$$

In this equation:

- N: Represents the total number of samples in the training set.
- n_i: Denotes the number of samples specifically belonging to class i.

We controlled the optimization process using a learning rate set to 2×10^{-5}, combined with a cosine decay learning rate schedule and a warm-up phase spanning the first 10% of the total training steps. Besides, we applied a weight decay[16] coefficient of 0.01 to enhance generalization performance and employed an early stopping mechanism with a patience of three epochs to prevent overfitting. We selected the final model checkpoint based on the highest macro-averaged F1 score on the validation set, as this metric represents the most appropriate and objective evaluation criterion for imbalanced classification tasks.

3. Data Augmentation Strategy

The empirical basis for this study consists of the datasets provided by the MultiPRIDE task organizers at EVALITA 2026. For Task A, which focuses on the classification of textual content, the provided material comprises social media data—specifically microblogging posts—annotated for the presence of reclamatory intent within the LGBTQ+ context.

While the primary data was originally distributed as language-specific subsets, a unified multilingual approach was adopted for this research. All individual language datasets were concatenated into a single corpus to facilitate the training of a robust multilingual model. This architectural decision was motivated by the objective of leveraging cross-lingual transfer learning, allowing the model to develop shared representations across different linguistic structures and cultural nuances associated with the target phenomenon.

3.1. MultiPride Dataset

The MultiPride dataset serves as the core corpus for the experimental framework, consisting of merged data from three languages: Italian, Spanish, and English. The dataset is characterized by a relatively balanced distribution among these languages, ensuring that the model is not unduly biased toward a single linguistic domain.

A significant feature of the dataset is the pronounced class imbalance[17] between the two target labels. While the majority of instances belong to the negative class (non-reclamation), the positive class (reclamation) is substantially underrepresented (see Table 1). Such an imbalance is representative of real-world data distributions in social media monitoring, where specific linguistic phenomena like reclamation occur with lower frequency than general discourse. This disparity presents a notable challenge for supervised learning, as models may develop a bias toward the majority class without specific intervention.

3.2. Synthetic Data

To address the limitations inherent in the original dataset distribution, particularly the sparsity of the minority class, a synthetic data generation phase was implemented. Large Language Models (LLMs)[18] were utilized to expand the training set through a few-shot prompting strategy[19]. We provided the models with representative examples from the original MultiPride corpus, and the generation process

Table 1
Distribution of samples in the MultiPride dataset

Split	Non-reclaimed	Reclaimed	Total
Training	2,048	342	2,390
Validation	512	86	598
Total	**2,560**	**428**	**2,988**

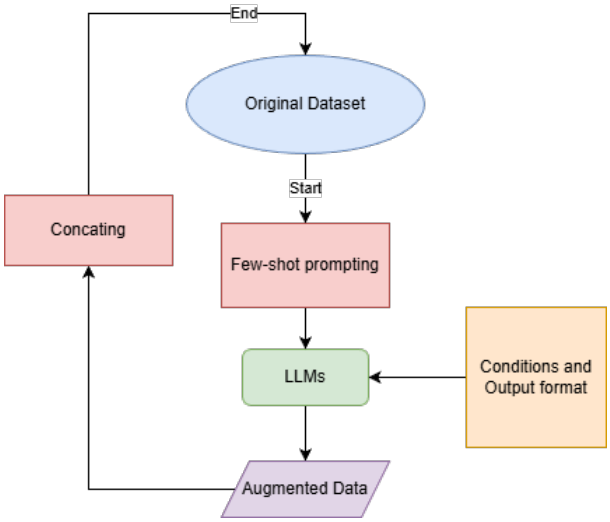

Figure 2: Workflow of the iterative data augmentation process using Few-shot prompting and Large Language Models (LLMs)

Table 2
Distribution of samples in the MultiPRIDE dataset after data augmentation.

Split	Non-reclaimed	Reclaimed	Total
Training (Augmented)	2,505	2,440	4,945
Validation	512	86	598
Total	**3,017**	**2,526**	**5,543**

ensured that the synthetic outputs maintained semantic and stylistic consistency with the source data - the workflow is visualized in Figure 2.

Synthetic augmentation primarily targeted the underrepresented label 1 to mitigate class imbalance and prevent majority-class bias. Furthermore, this strategy served to scale the training corpus for more stable weight updates during fine-tuning while simultaneously diversifying linguistic contexts to enhance model generalization on unseen instances. The use of advanced LLMs enabled the creation of high-quality synthetic examples that accurately reflect the complex pragmatic nature of reclaimed language, while preserving the overall distribution of the task's domain[20].

Following this workflow, we applied the augmentation process to the MultiPRIDE dataset. First, we performed data augmentation for each target language by prompting Large Language Models (LLMs) to generate synthetic samples in .csv format. These generated files were then merged with our existing training set, while the validation set remained untouched to ensure the integrity of the evaluation. The final composition of the augmented dataset is presented in detail in Table 2:

Table 3
Classification results across different languages and runs

Language	Run	F1 (Class 0)	F1 (Class 1)	Macro F1
English (en)	**run-1**	0.9396	0.3214	0.6305
	run-2	0.9402	0.2400	0.5901
Spanish (es)	**run-1**	0.9312	0.5868	0.7590
	run-2	0.9365	0.6049	0.7707
Italian (it)	**run-1**	0.9396	0.7418	0.8407
	run-2	0.9475	0.7704	0.8589

Table 4
Our results in MultiPRIDE Task A at EVALITA 2026 Final Ranking

Subtask	Team	Run	Rank	Macro F1
A1 (Italian)	Ghavidel-Rajabi	1	1	0.8981
	MilaNLP	1	2	0.8959
	Baseline	1	10	0.8731
	NamDang (Ours)	**2**	**12**	**0.8589**
	NamDang (Ours)	**1**	**17**	**0.8407**
A2 (Spanish)	The Hate Busters	2	1	0.7776
	NamDang (Ours)	**2**	**2**	**0.7707**
	NamDang (Ours)	**1**	**3**	**0.7590**
	Baseline	1	14	0.7000
A3 (Italian)	Avahi	1	1	0.6416
	SaFe Tweets	1	2	0.6329
	NamDang (Ours)	**1**	**3**	**0.6305**
	NamDang (Ours)	**2**	**7**	**0.5901**
	Baseline	1	8	0.5760

4. Result

The system performance was evaluated through two official runs on the MultiPRIDE test set at EVALITA 2026. The objective of these experiments was to compare the generalization capabilities of different Transformer architectures under the influence of data augmentation strategies versus training on the original dataset. ***run-1*** corresponds to a configuration using the microsoft/mdeberta-v3-base model, which incorporates architectural enhancements and is trained on the augmented dataset. In contrast, ***run-2*** serves as a reference baseline, employing the multilingual xlm-roberta-base model trained directly on the original dataset.

The experimental results are summarized using a macro-averaged F1 score, with the latter serving as the primary evaluation metric for assessing classification performance on imbalanced datasets. Table 3 reports the detailed system performance across the three target languages (English, Spanish, and Italian) for both runs. Overall, the system achieved its best performance on Italian, with the highest macro-averaged F1 score reaching 0.8589 in ***run-2***. Spanish ranked second, attaining a peak macro F1 score of 0.7707, also in ***run-2***. In contrast, English proved to be the most challenging language for the system, with macro F1 scores ranging only from 0.5901 to 0.6305 across the two runs. This performance gap may reflect differences in contextual complexity or language-specific patterns in how reclaimed language is used by LGBTQ+ communities on social media across different countries.

For English, ***run-1*** yielded comparatively better results for the positive class and in terms of macro F1, suggesting that the combination of data augmentation and the disentangled attention mechanism of DeBERTa facilitated the learning of more diverse lexical variants. Conversely, for Spanish and

Italian,***run-2*** consistently outperformed ***run-1*** across all evaluation metrics. This observation may indicate that, for these two languages, the original data distribution was sufficiently informative for XLM-RoBERTa to exploit effectively, or that the data augmentation applied in ***run-1*** may have inadvertently introduced noise, leading to a degradation in local classification accuracy.

5. Discussion

Following the official evaluation phase, the organizing committee released detailed results for each submitted run, along with an *errors_lang.csv* file for every language and subtask, containing 20 representative misclassified instances. These error files correspond to the best-performing run according to the overall results. An in-depth inspection of these instances provides valuable insight into the underlying causes of the model's incorrect predictions. Based on the analysis of the error files across the three languages, we identify several recurring patterns that highlight the system's limitations.

For Italian and Spanish, the model errors primarily stem from the use of slur terms within socially and politically complex contexts, where the boundary between attack and reclamation becomes blurred. In Italian, the system struggles in particular with meta-discursive utterances, in which users quote or explicitly discuss the use of slurs rather than directly targeting an individual or group, leading to misinterpretation of the communicative intent. Moreover, highly polarized Pride-related contexts, characterized by a dense concentration of sensitive terms employed with positive or empowering meanings, tend to trigger default negative predictions due to the model's over-reliance on surface-level lexical cues. Similarly, in counter-speech against discrimination, users often deliberately reuse slur terms to criticize oppressive behaviors; however, the model lacks the pragmatic depth required to distinguish social critique from direct verbal abuse. In Spanish, the presence of hashtags such as *#OrgulloLGTBI* or *#Pride* introduces additional noise, occasionally biasing the model toward positive labels even when the surrounding content still contains strong or potentially offensive language. Furthermore, regional variants and colloquial nuances of certain slurs are not sufficiently captured by the model, particularly under the non-augmented training setting, resulting in systematic classification errors.

English is the language in which the model exhibits the weakest performance, largely due to the widespread and culturally entrenched nature of reclaimed language, which creates extremely fragile boundaries between offense and empowerment. In many instances, distinguishing between in-group and out-group usage depends almost entirely on the speaker's identity—an aspect that cannot be inferred by a text-only model. Additionally, forms of dark humor, sarcasm, and self-directed irony are pervasive in the data, leading the system to assign negative labels based solely on the presence of slur tokens, while ignoring narrative structure and speaker stance. The polysemy of terms such as "queer", which may function as a positive academic or political identifier or as an insult in conservative contexts, further exacerbates prediction instability and bias.

6. Conclusion

This paper reports our participation in Task A of the MultiPride shared task at EVALITA 2026, focusing on the identification of reclaimed language within the LGBTQ+ community across English, Italian, and Spanish. The system achieved its best performance (F1-macro) in Italian (0.8589) and Spanish (0.7707) using a fine-tuned XLM-RoBERTa-base model trained on the original dataset, while for English the highest score (0.6305) was obtained with mDeBERTa-v3-base enhanced by synthetic data augmentation, highlighting the benefits of disentangled attention for capturing lexical diversity.

As future work, we aim to incorporate identity-aware modeling to better distinguish in-group from out-group usage, a key limitation of text-only approaches. We also plan to explore contrastive learning and external knowledge integration to address pragmatic phenomena such as sarcasm, self-directed irony, and the polysemy of sensitive terms, as well as to enrich synthetic data with more culturally and pragmatically complex contexts.

Declaration on Generative AI

In this work, Generative AI tools were used exclusively for linguistic clarity. The conceptualization, data analysis, and conclusions remain the sole work of the authors, who accept full responsibility for the manuscript's content.

References

[1] M. S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural language processing, Neurocomput. 546 (2023). URL: https://doi.org/10.1016/j.neucom.2023.126232. doi:10.1016/j.neucom.2023.126232.

[2] A. Velankar, H. Patil, R. Joshi, A review of challenges in machine learning based automated hate speech detection, 2022. URL: https://arxiv.org/abs/2209.05294. arXiv:2209.05294.

[3] J. G. Parmenter, R. V. Galliher, A. D. A. Maughan, An exploration of lgbtq+ community members' positive perceptions of lgbtq+ culture, The Counseling Psychologist 48 (2020) 1016–1047. URL: https://doi.org/10.1177/0011000020933188. doi:10.1177/0011000020933188. arXiv:https://doi.org/10.1177/0011000020933188.

[4] A. D. Galinsky, C. S. Wang, J. A. Whitson, E. M. Anicich, K. Hugenberg, G. V. Bodenhausen, The reappropriation of stigmatizing labels: The reciprocal relationship between power and self-labeling, Psychological Science 24 (2013) 2020–2029. URL: https://doi.org/10.1177/0956797613482943. doi:10.1177/0956797613482943. arXiv:https://doi.org/10.1177/0956797613482943, pMID: 23955354.

[5] M. Melis, G. Lapesa, D. Assenmacher, A modular taxonomy for hate speech definitions and its impact on zero-shot LLM classification performance, in: A. Calabrese, C. de Kock, D. Nozza, F. M. Plaza-del Arco, Z. Talat, F. Vargas (Eds.), Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 490–521. URL: https://aclanthology.org/2025.woah-1.45/.

[6] M. Tonneau, D. Liu, N. Malhotra, S. A. Hale, S. P. Fraiberger, V. Orozco-Olvera, P. Röttger, Hateday: Insights from a global hate speech dataset representative of a day on twitter, 2025. URL: https://arxiv.org/abs/2411.15462. arXiv:2411.15462.

[7] C. Ferrando, L. Draetta, M. Madeddu, M. Sosto, V. Patti, P. Rosso, C. Bosco, J. Mata, E. Gualda, Multipride at evalita 2026: Overview of the multilingual automatic detection of slur reclamation in the lgbtq+ context task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Neural Information Processing Systems, 2017. URL: https://api.semanticscholar.org/CorpusID:13756489.

[9] 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, 2020. URL: https://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[10] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. URL: https://arxiv.org/abs/2111.09543. arXiv:2111.09543.

[11] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for nlp, 2021. URL: https://arxiv.org/abs/2105.03075. arXiv:2105.03075.

[12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing, 2020. URL: https://arxiv.org/abs/1910.03771. arXiv:1910.03771.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov,

Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[14] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Édouard Duchesnay, Scikit-learn: Machine learning in python, 2018. URL: https://arxiv.org/abs/1201.0490. arXiv:1201.0490.

[16] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. URL: https://arxiv.org/abs/1711.05101. arXiv:1711.05101.

[17] S. Henning, W. Beluch, A. Fraser, A. Friedrich, A survey of methods for addressing class imbalance in deep-learning based natural language processing, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 523–540. URL: https://aclanthology.org/2023.eacl-main.38/. doi:10.18653/v1/2023.eacl-main.38.

[18] A. Matarazzo, R. Torlone, A survey on large language models with some insights on their capabilities and limitations, 2025. URL: https://arxiv.org/abs/2501.04040. arXiv:2501.04040.

[19] T. Schick, H. Schütze, True few-shot learning with prompts—a real-world perspective, Transactions of the Association for Computational Linguistics 10 (2022) 716–731. URL: https://doi.org/10.1162/tacl_a_00485. doi:10.1162/tacl_a_00485.

[20] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A. T. Luu, S. Joty, Data augmentation using large language models: Data perspectives, learning paradigms and challenges, 2024. URL: https://arxiv.org/abs/2403.02990. arXiv:2403.02990.