Team JAT at SemEval-2026 Task 9: Enhancing Polarization Detection with Cross-Lingual Transfer and Feature Fusion

Aleksandra Matkowska a.matk09@gmail.com
Taya Lin tayalin2018@gmail.com
Yu-Chun Chao wxes41028@gmail.com
Eberhard Karls Universität Tübingen

Abstract

We describe our system for SemEval-2026 Task 9 (POLAR), Subtask 1 - binary polarization detection. Our approach investigates polarization detection through monolingual and cross-lingual experimental settings. We first utilize a RoBERTa-based architecture enhanced with feature fusion, combining contextual sentence representations with handcrafted sentiment and intensity cues. As for multilingual joint training, we explore it within the Indo-European family to test whether cross-lingual transfer can elevate performance in data-scarce scenarios. Our final fine-tuned model achieves average F1-score of 0.763 on the test set, compared to 0.491 for the random baseline and 0.769 for the official POLAR baseline. We also report ablations for augmentation, feature fusion, and class weighting to quantify each component's contribution.

1 Introduction

Polarization is defined by Cambridge Dictionary as the "act of dividing something, especially something that contains different people or opinions, into two completely opposing groups" (Cambridge University Press, n.d.). This concept has become increasingly more prominent in recent years as political landscapes in various countries have become more divisive, as has been reported to be the case in the USA, Poland, Columbia, Britain, or Indonesia (Simchon et al., 2022). This rise of political fragmentation coincides with the rapid development of social media platforms, which play a crucial part in the current sociopolitical landscape. Specifically, platforms that rely on social networks and news-feed algorithms (like Facebook or Twitter) contribute to the emergence of echo chambers among its users (Cinelli et al., 2021), which has been linked to an increase in polarization (Allcott et al., 2020).

As the organizers of Task 9 in SemEval-2026 note, polarizing speech is often a precursor to hate and offensive speech and can lead to a highly fragmented society (Naseem et al., 2026a). As such developing a reliable polarization detection tool should be seen as a very pressing challenge for Natural Language Processing (NLP) community. Being able to correctly identify and mitigate polarization is crucial for fostering safety and inclusivity in online spaces, which could lead to a more tolerant society where open and constructive dialogue is encouraged.

Within the shared task, participants can choose among three subtasks: (1) Polarization Detection, (2) Polarization Type Classification and (3) Polarization Manifestation Identification. This paper focuses on Subtask 1 for 13 Indo-European languages (Bengali, German, English, Farsi, Hindi, Italian, Nepali, Odia, Punjabi, Polish, Russian, Spanish, Urdu) and is structured as follows: in section 2 we review related work; section 3 gives the system overview; in section 4 we detail the experimental set up and report our results in section 5; finally, section 6 concludes the paper.

2 Background

2.1 Task Description

POLAR (SemEval-2026 Task 9) provides a multilingual benchmark for recognizing polarization in online text, where authors present issues as zero-sum struggles between groups and signal intolerance or hostility toward out-groups (Naseem et al., 2026b). The organizers compile short posts and comments from diverse public sources—news websites, Reddit, blogs, Bluesky, and regional forums—and intentionally span multiple cultural contexts and event types (e.g., elections, conflicts, migration, and gender rights). The benchmark covers 22 languages and typically offers roughly 3,000–5,000 annotated instances per language, so systems must cope with both cross-lingual variation and domain shifts across events. Subtask 1 is

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2402–2408

binary polarization detection, predicting whether a text is polarized or not. Macro-F_1 score was used as the evaluation metric as it treats each class/label equally when averaging, making performance on minority polarization categories directly affect the final score and encouraging balanced modeling across languages (Naseem et al., 2026a).

2.2 Related Work

Prior work relevant to our setup can be grouped into three themes: multilingual encoder adaptation, imbalance-aware learning, and feature fusion. First, transformer adaptation choices strongly affect downstream classification quality. Sun et al. (2019) show that optimization and fine-tuning strategy matter substantially, while Conneau et al. (2020) demonstrate strong cross-lingual transfer with XLM-R and analyze the trade-off between transfer gains and multilingual capacity dilution.

As recent studies have shown, polarization is a cross-lingual phenomenon influenced by language and cultural background (Naseem et al., 2026a). In this setting, multilingual transformer models such as XLM-R provide a strong starting point, as they create a shared cross-lingual semantic space and support effective zero-shot transfer across typologically diverse languages (Conneau et al., 2020). Recent analyses of cross-linguistic transfer (Bankula, 2025) further demonstrate that transfer performance correlates with language-family proximity and morphological similarity, with substantially higher scores for intra-family transfer than for cross-family transfer.

Hence, we start with the focus on the Indo-European subset, as the shared syntactic and morphological structures, along with partially aligned discourse patterns, can be exploited by aggregating training data across related languages to alleviate the sparsity issue of per-language data (Bankula, 2025; Naseem et al., 2026a). On the other hand, we apply LoRA fine-tuning methods to further complement this data-centric strategy. Previous work on parameter-efficient fine-tuning (PEFT) shows that adapter- and LoRA-style methods can adapt large pretrained encoders to new text-classification domains under data-scarce conditions while preserving or even improving macro-F_1 relative to full fine-tuning, especially by avoiding overfitting and maintaining performance on minority classes (Hu et al., 2021; Nwaiwu, 2025).

Second, class imbalance is central for POLAR because macro-F_1 penalizes weak minority-class performance (Naseem et al., 2026a; Henning et al., 2023) summarize practical remedies, including sampling, augmentation, and loss-level adjustments. This motivates our use of data augmentation and class-weighted training.

Third, fusion architectures suggest that contextual embeddings and explicit lexical cues can be complementary. Nagar et al. (2019) shows gains from combining local and global text representations. This is relevant for polarization, where intensity markers and group-referencing patterns may add signal beyond contextual semantics. Our system therefore combines a RoBERTa backbone with handcrafted linguistic features.

Setting	Macro-F_1
13-language mean (random, $p = 0.5$)	0.491

Table 1: Mean random baseline result across the 13 Indo-European languages.

3 System Overview

We mostly focus on XLM-RoBERTa (XLM-R) large, which was trained on 2.5TB of filtered CommonCrawl data across 100 languages (Conneau et al., 2020), among which were the languages we focus on. We apply fine-tuning in different linguistic configurations to inspect whether cross-lingual transfer can elevate performance in closely related languages. Additionally, we train specialized monolingual models for English and Polish to serve as a comparison for the multilingual setups.

3.1 Random Baseline

To document a simple lower bound, we include a random baseline that predicts the polarized class with probability $p = 0.5$. This baseline is independent of the input text and provides a sanity check for the task-specific models. Table 1 reports its mean score, while Table 3 reports the random baseline for each Indo-European language individually.

3.2 Monolingual Feature Fusion

Focusing only on English for this variant of the model, we extracted a 5-dimensional feature vector as auxiliary input to capture sentiment and structural cues. The features included: VADER Compound Score (Hutto and Gilbert, 2014) to represent sentiment intensity, uppercase character ratio to total text length, exclamation mark count and emotion counts of tokens associated with anger

and fear based on the NRC Emotion Lexicon (Mohammad and Turney, 2013).

Let h_{CLS} be the RoBERTa sentence representation and f the handcrafted 5-dimensional feature vector. To ensure numerical stability alongside the dense embeddings, f is first transformed using standard scaling and clamped to the range $[-5, 5]$. We fuse them by concatenation, $z = [h_{\text{CLS}}; f]$, apply dropout ($p = 0.4$) to the fused representation, and predict the probabilities via $\hat{y} = \text{softmax}(Wz + b)$. In plain terms, the classifier jointly uses contextual semantics from RoBERTa and explicit intensity/emotion cues from f.

To increase minority coverage, we used Easy Data Augmentation (EDA) with four operations: synonym replacement, random insertion, random swap, and random deletion. We set $N_{\text{aug}} = 8$ augmented sentences per original minority-class sentence, then mix augmented and original examples. To additionally reduce the majority-class bias, we applied proportional class weights in cross-entropy (Class 0: 1.0; Class 1: 1.67) alongside a label smoothing factor of 0.1.

Finally, the models were trained using early stopping with patience of two epochs, evaluated on the validation macro-F_1 score to prevent overfitting. To improve generalization and robustness, our final predictions are derived by averaging the output logits from an ensemble of models trained across three different random seeds.

3.3 LoRA Finetuning

In the fine-tuning approach, we wanted to explore both monolingual and multilingual settings. Therefore, we used XLM-R as the base model for multilingual analysis, and RoBERTa and HerBERT, a BERT-based LM trained on Polish corpora (Mroczkowski et al., 2021), as monolingual reference points.

We conducted a hyperparameter search for learning rate, batch size, dropout, weight decay, the number of epochs, and LoRA rank and alpha (Hu et al., 2021) to find the best configuration for each iteration of our models. Additionally, during training we use Early Stopping based on the F1-score with patience = 3.

To account for imbalanced data we implemented a custom cross entropy weighted loss with class weights given by $w_j = \frac{n}{k \cdot n_j}$, where n is the total number of samples in the dataset and k is the number of unique classes. We do not use any data augmentation techniques anymore.

Lang.	0	1	0:1 ratio
ben	1909	1424	1.34
deu	1668	1512	1.10
eng	2047	1175	1.74
fas	855	2440	0.35
hin	398	2346	0.17
ita	1966	1368	1.44
nep	997	1008	0.99
ori	1685	683	2.47
pan	860	840	1.02
pol	1388	1003	1.38
rus	2325	1023	2.27
spa	1645	1660	0.99
urd	1087	2476	0.44
Total	18830	18958	0.99

Table 2: Raw counts showing per-language class distribution in the training set. Class 0 denotes non-polarized instances, class 1 - polarized.

4 Experimental Setup

We devise five multilingual configurations for fine-tuning: all thirteen Indo-European languages together; English, German, Polish, and Russian, representing languages from the same subfamilies – Germanic and Slavic, respectively; English and Polish to inspect the result of including two Indo-European languages from different sub-families; and English and Chinese in order to evaluate the influence of an entirely unrelated language on performance of an Indo-European language. We further fine-tune on English, Polish, and Russian in monolingual settings with a multilingual backbone (XLM-R), and on English and Polish with specialized monolingual backbones (RoBERTa, HerBERT).

4.1 Datasets

We use the datasets provided by the organizers of SemEval Task 9 (Naseem et al., 2026b). Although the combined training split of all 13 languages we are interested in is balanced, the same is not true for individual language datasets, as shown in Table 2. The most extreme cases include Hindi, where there are almost six times as many polarized instances compared to unpolarized, or Odia where there are over twice as many unpolarized examples as polarized ones.

4.2 Threshold optimization

To further combat class imbalance in the datasets, in addition to a custom loss function, we also optimize individual decision thresholds for each language. We begin by setting the threshold to 0.5 at the beginning of training. After each epoch we evaluate the model on the development set by iteratively searching through the possible threshold values and calculating the F1 score for each one, saving the best score with the corresponding threshold. We then apply the winning decision threshold for each language for test set predictions.

4.3 Training process

Having run the hyperparameter search for each linguistic set-up, the following configurations repeatedly emerged as the best ones: learning rate of 10^{-4}, batch size of 16, dropout value of 0.1 or 0.05, weight decay of 0.1, the number of epochs between 5 and 10, LoRA rank of 4 or 8, and alpha of 16 or 32. Benchmark baseline settings used batch size 16, maximum sequence length 128, and 1–2 epochs, For a comprehensive overview see Table 4 in Appendix A.

Baseline runs were executed on CUDA on our CRETE server using a single NVIDIA GeForce RTX 5090 GPU (32 GB VRAM), with mixed precision (bf16 AMP) for training baselines. The fine-tuning of the models was performed in the Baden-Württemberg high performance computing cluster, using HuggingFace Transformers (Wolf et al., 2020), PyTorch (Ansel et al., 2024) and scikit-learn (Pedregosa et al., 2011).

5 Results

In this section, we report our results for the feature fusion, as well as the LoRA fine-tuned model (see Table 3). Due to a better performance of the fine-tuned architecture on the development set, we submitted that model as our entry to the official competition, achieving the following places in the official rankings: 2nd for Odia, 11th for German, 15th for Spanish, 21st for Nepali and Russian, 22nd for Farsi, 23rd for Urdu, 27th for English and Polish, 28th for Italian, 30th for Bengali and Punjabi, and 31st for Hindi.

Compared with the official POLAR baseline reported on the task leaderboard (POLAR SemEval Organizers, 2026), our submitted 13-language model improves macro-F_1 for 8 of the 13 Indo-European languages: German ($+0.050$), English ($+0.006$), Hindi ($+0.005$), Nepali ($+0.018$), Odia ($+0.040$), Polish ($+0.052$), Russian ($+0.028$), and Spanish ($+0.050$). Its average score across these languages is slightly below the official baseline (0.763 vs. 0.769), mainly because of larger drops for Italian (-0.138), Punjabi (-0.083), and Farsi (-0.054). This comparison suggests that our multilingual LoRA setup is competitive with the organizer baseline for several languages, but less robust for languages where the official baseline is already strong.

5.1 Feature Fusion

After performing permutation feature importance of our handcrafted features, it became apparent that they provide almost no predictive gain in a simple linear model (logistic regression).

When it comes to the effects of EDA, an ablation experiment showed that augmentation alone significantly worsened the performance of RoBERTa with fused handcrafted features on the development set ($0.809 \rightarrow 0.716$).

Overall, this model achieves a score of 0.788 on the test set. Given our analysis and the result achieved by fine-tuning RoBERTa for the task (0.791), we conclude that the integration of additional handcrafted features provides no additional benefit to the model, while EDA actively harms its performance.

5.2 LoRA Finetuning

Considering English, we observe that the best performing model for that language is fine-tuned with English and German data, and it performs on the same level as a specialized monolingual RoBERTa model. At the same time, while German benefits from this configuration in comparison to the monolingual XLM-R set-up ($0.685 \rightarrow 0.707$), it scores higher with our final multilingual model ($0.685 \rightarrow 0.721$). For English input from additional 11 languages is also beneficial (0.066 improvement), while fine-tuning with only Polish still elevates the score, but to a lesser extent (0.035). Lastly, a model fine-tuned on English and Chinese shows a decrease in English F1-score compared to a monolingual set-up ($0.720 \rightarrow 0.713$).

For Polish, we achieved the best score (0.827) with a specialized monolingual model. However, with XLM-R as the backbone performance in Polish seems to have benefited significantly from the Russian data, as its score increased to 0.806 from 0.768 when fine-tuned on its own. This is a consid-

Model	ben	deu	eng	fas	hin	ita	nep	ori	pan	pol	rus	spa	urd	zho
Random baseline ($p = 0.5$)	0.480	0.499	0.497	0.466	0.429	0.497	0.534	0.501	0.508	0.518	0.486	0.498	0.474	—
Official POLAR baseline (POLAR SemEval Organizers, 2026)	**0.853**	0.671	0.780	**0.842**	0.738	**0.677**	0.880	0.777	**0.790**	0.724	0.746	0.727	**0.789**	0.869
XLM-R monolingual fine-tuning	—	0.685	0.720	—	—	—	—	—	—	0.768	**0.790**	—	—	—
XLM-R multilingual LoRA fine-tuning														
All 13 languages	0.821	**0.721**	0.786	0.788	**0.743**	0.539	**0.898**	**0.816**	0.707	0.776	0.774	**0.777**	0.770	—
English & German	—	0.707	**0.792**	—	—	—	—	—	—	—	—	—	—	—
English & Polish	—	—	0.755	—	—	—	—	—	—	0.775	—	—	—	—
English & Chinese	—	—	0.713	—	—	—	—	—	—	—	—	—	—	**0.884**
Polish & Russian	—	—	—	—	—	—	—	—	—	0.806	0.780	—	—	—
Specialized monolingual training														
HerBERT LoRA	—	—	—	—	—	—	—	—	—	**0.827**	—	—	—	—
RoBERTa fine-tuned	—	—	0.791	—	—	—	—	—	—	—	—	—	—	—
RoBERTa + features	—	—	0.788	—	—	—	—	—	—	—	—	—	—	—

Table 3: Results from all of our experiments.

erable improvement from the F1-score achieved after training alongside English (0.775), which is not as closely related to Polish as Russian. On the other hand, no similar boost was observed in Russian – its best score (0.790) was achieved with monolingual fine-tuning of XLM-R and it decreased after fine-tuning with Polish (0.780), as well as with the other Indo-European languages (0.774).

Finally, we note that our final model (fine-tuned on all Indo-European languages) performed significantly better on the development set than on the test set. The average F1-score decreased by 0.053 (0.816 $\rightarrow$ 0.763), suggesting that the model is lacking in its ability to generalize to large amounts of unseen data (there were around 10 times more elements in test sets than in development sets; see Appendix A).

6 Conclusion

Overall, we presented our cross-lingual approach to the detection of polarization, which relied on two different architectural setups. We started our research with a feature-fusion system combining RoBERTa representations with handcrafted sentiment/intensity cues, class rebalancing, and EDA-based augmentation. Having discovered that handcrafted features carried little predictive power and data augmentation harmed model performance, we reevaluated our approach. We then moved on to a fine-tuning strategy, an iteration of which became our submission to the SemEval-2026 Task 9 Subtask 1. The final fine-tuned system reached a mean F1-score of 0.763 on the test set, showing the benefit of task-specific adaptation.

Across experiments, we found that with a multilingual transformer backbone, performance can be boosted by supplying more data from a related language. Nonetheless, specialized monolingual models of high resource languages still show the best performance. These findings motivate more targeted adaptation and further research to improve performance of resource-scarce languages.

Acknowledgments

The authors acknowledge support by the state of Baden-Württemberg through bwHPC. We would like to offer special thanks to Dr. Çağrı Çöltekin for guidance on this project.

References

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The welfare effects of social media. *American Economic Review*, 110(3):629–676.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster machine learning through dynamic Python bytecode transformation and graph compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

A. Bankula. 2025. Cross-linguistic transfer in multilingual NLP: The role of language families and morphology. *CoRR*, abs/2505.13908.

Cambridge University Press. n.d. Polarization. Cambridge Dictionary. Accessed: 2025-12-21.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and

Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Ajay Nagar, Anmol Bhasin, and Gaurav Mathur. 2019. Text classification using gated fusion of n-gram features and semantic features. *Computación y Sistemas*, 23(3):1015–1020.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, Dheeraj Kodati, Sahar Moradizeyveh, Firoj Alam, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Nelson Odhiambo Onyango, Clemencia Siro, Ibrahim Said Ahmad, Lilian Wanzare, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, Arfeen Zeeshan, Dheeraj Kodati, Satya Keerthi, Sahar Moradizeyveh, Firoj Alam, Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026b. POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Steve Nwaiwu. 2025. Parameter-efficient fine-tuning for low-resource text classification: a comparative study of LoRA, IA3, and ReFT. *Frontiers in Big Data*, 8(1677331).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

POLAR SemEval Organizers. 2026. POLAR @ SemEval-2026: Official leaderboards. `https://github.com/Polar-SemEval/Leaderboards`. Accessed: 2026-05-01.

Almog Simchon, William J Brady, and Jay J Van Bavel. 2022. Troll and divide: the language of online polarization. *PNAS Nexus*, 1(1):pgac019.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

A Appendix

	learning rate	batch size	dropout	weight decay	epochs	rank	alpha
baseline	10^{-4}	16	—	—	1–2	—	—
XLM-R (all)	10^{-4}	16	0.05	0.01	5	8	16
XLM-R (eng-deu)	10^{-4}	8	0.05	0.01	10	4	32
XLM-R (eng-pol)	10^{-4}	16	0.05	0.0	8	4	32
XLM-R (eng-zho)	10^{-4}	8	0.1	0.0	5	4	16
XLM-R (pol-rus)	10^{-5}	16	0.05	0.01	8	16	16
XLM-R (eng)	10^{-5}	16	0.1	0.01	8	4	32
XLM-R (pol)	10^{-4}	16	0.1	0.01	5	4	32
XLM-R (rus)	10^{-4}	32	0.1	0.001	10	4	16
XLM-R (deu)	10^{-4}	32	0.1	0.01	8	8	16
HerBERT	10^{-4}	16	0.05	0.0	8	4	32
RoBERTa fusion	10^{-5}	32	0.4	0.3	5	—	—

Table 4: Hyperparameter settings for each model.

Language	**Train**	**Development**	**Test**
ben	3,333	166	1,501
deu	1,668	159	1,432
eng	2,047	160	1,452
fas	3,295	164	1,484
hin	2,744	137	1,236
ita	3,334	166	1,538
nep	2,005	100	903
ori	2,368	118	1,066
pan	1,700	100	809
pol	2,391	119	1,077
rus	3,348	167	1,508
spa	3,305	165	1,488
urd	3,563	177	1,606
Total	35,101	1,898	17,100

Table 5: Per-language dataset split.