DigiS-FBK at SemEval-2026 Task 9: Multi-task Learning for Multilingual and Cross-cultural Polarization Classification

Veronica Orsanigo[1,2], Alan Ramponi[1], Elisa Leonardelli[1]

[1]Fondazione Bruno Kessler, [2]University of Trento

Correspondence: vorsanigo@fbk.eu, alramponi@fbk.eu, eleonardelli@fbk.eu

Abstract

Online polarization promotes social fragmentation, misinformation, hate, and toxic language. Polarization has been studied from social and communication perspectives, but it can also be addressed computationally as a text classification task. Due to the variety of polarization targets and manifestations, polarization is a complex phenomenon to study, and both detecting and characterizing it are challenging tasks. In this paper, we present the systems submitted by the DigiS-FBK team to SemEval-2026 Task 9 (POLAR) aimed at detecting polarization in textual content (subtask 1) and identifying its type (subtask 2) and manifestation (subtask 3) in a multilingual, multicultural, and multievent context. Considering the strong link between subtasks, we propose an approach that leverages a multi-task learning paradigm. Our results reveal that, despite the variability in scores across languages, the overall performance when using multi-task learning is higher than when adopting a single task approach in all subtasks.[1]

1 Introduction and Background

Online polarization is a growing concern today. Through the creation of separated and opposed groups, polarized content naturally incentivizes social fragmentation (Vasist et al., 2024), hate speech, and toxic language (Liu et al., 2024). Moreover, polarization may promote the spread of misinformation (Vicario et al., 2019). For these reasons, identifying polarization at an early stage is crucial for developing interventions that foster healthier online environments.

While polarization has been studied mainly from social and communication perspectives (Sunstein, 2002), it can also be addressed computationally. In natural language processing (NLP), it can be framed as a text classification task that aims to detect divisive and segregating discourse. However, previous work has mainly tackled it indirectly by relying on sentiment analysis (Martínez-España et al., 2024) and stance detection (Mohammad et al., 2016) systems, which inherently fail to capture the broader dynamics of inter-group hostility and identity-driven polarization.

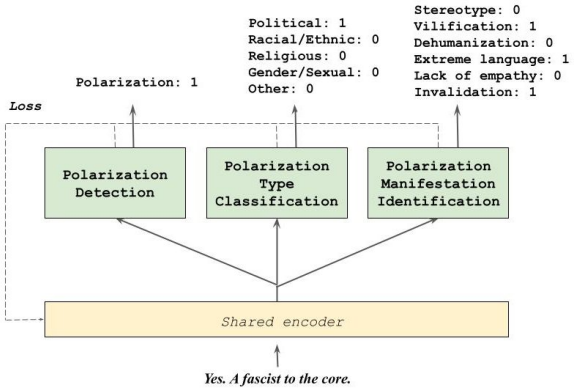

Figure 1: High-level overview of our multi-task learning framework for polarization detection and classification, with an example from the `train` set and associated output labels for the three POLAR subtasks.

POLAR (Naseem et al., 2026a), a shared task organized as part of SemEval-2026 (Task 9) and based on the homonymous dataset (Naseem et al., 2026b, Section 2.1), aims to promote research to address textual polarization online with computational means. With its three subtasks (i.e., polarization detection, polarization type classification, and polarization manifestation identification; Section 2.2), POLAR not only aims to detect polarized textual content, but also to characterize it in a multilingual, multicultural, and multievent context, thereby addressing a gap in current literature.

This paper presents the **DigiS-FBK** systems submitted for the POLAR shared task. Given the interdependent nature of polarization detection and type and manifestation identification, we propose a multi-task learning approach in which the different subtasks are learned jointly to improve per-

[1]Code is available on GitHub at: https://github.com/vorsanigo/polar.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2838–2851

formance on each individual subtask (Figure 1). By relying on encoder-based models instead of computationally-demanding large language models (LLMs) and combining different subtasks into a unified framework, we obtain a lightweight model to be seamlessly used across subtasks. This makes our approach especially valuable in scenarios in which large computational resources are not available or computational time is a requirement. Our results show that our multi-task learning approach consistently improves the performance across POLAR subtasks compared to a single task setup. However, we also observe that there is still room for additional tuning to further improve performance.

2 Data and Task Description

In this section, we describe the POLAR dataset (Section 2.1) and the associated shared task that has been organized upon it (Section 2.2).

2.1 POLAR Dataset

The annotated data for SemEval-2026 Task 9, POLAR (Naseem et al., 2026b), contains text from sources including news websites, commentary forums, and social media (e.g., Reddit, Facebook, X, Bluesky, Threads, YouTube, Weibo, Zhihu) and covers events such as elections, conflicts, gender rights, and migration, among others. POLAR is a large-scale, multilingual, multicultural, and multi-event dataset intended for polarization detection. The POLAR shared task includes three subtasks, described in Section 2.2. For the subtasks of polarization detection and polarization type classification, data is provided in 22 languages from seven language families across different cultural and geographical contexts,[2] while for polarization manifestation identification, a subset of 18 languages is available.[3] Train subsets for each language contain between 1,700 and 7,000 annotated instances (with most of them around 3,000), as reported in Appendix B, making POLAR quite balanced across different languages.

For data annotation, Naseem et al. (2026b) provided detailed guidelines in English and then translated and culturally adapted them for each target language. However, even though the guidelines have been standardized, the dataset authors highlighted that different cultural and political contexts could have had a minor influence on annotators' interpretation. A summary of aggregated statistics about the POLAR dataset is in Table 1, where we also report the size of the `train` and `dev` splits that we create for our experiments, as described in Section 4.1. Per-language statistics with number of instances and label distribution are reported in Appendix B. More details on the data and annotation process can be found in Naseem et al. (2026b).

train set		**dev set**		**test set**
original	*ours*	*original*	*ours*	*original*
73,681	68,513	3,687	8,853	33,288

Table 1: Aggregated statistics across `train`, `dev`, and `test` splits computed on the full POLAR dataset by merging all the language subsets. For `train` and `dev` sets, in addition to the *original* size we also report the size of the custom split we create for our experiments (*ours*).

2.2 POLAR Shared Task

The POLAR shared task is organized into three related subtasks (S1, S2, and S3) aimed at polarization detection and characterization.

[S1] Polarization detection A binary classification task aimed at identifying whether a text contains polarized content or not;

[S2] Polarization type classification A multi-label classification task aimed at identifying the type or target of polarized content, if any. Possible polarization types or targets are: *Political*, *Racial/Ethnic*, *Religious*, *Gender/Sexual*, *Other*;

[S3] Polarization manifestation identification A multi-label classification task aimed at identifying how polarization is expressed, namely which rhetorical tactics are used among the following: *Stereotype*, *Vilification*, *Dehumanization*, *Extreme language*, *Lack of empathy*, and *Invalidation*.

Since polarized content often exhibits conceptual and contextual overlap, multiple types and manifestations may occur in the same text. In Table 2, we report an extract from the English data subset. Moreover, in Appendix A, we report the definitions of the polarization categories mentioned above, according to Naseem et al. (2026b).

[2] Languages in subtask 1 and 2: Amharic (`amh`), Arabic (`arb`), Bengali (`ben`), Burmese (`mya`), Chinese (`zho`), English (`eng`), German (`deu`), Hausa (`hau`), Hindi (`hin`), Italian (`ita`), Khmer (`khm`), Nepali (`nep`), Odia (`ori`), Persian (`fas`), Polish (`pol`), Punjabi (`pan`), Russian (`rus`), Spanish (`spa`), Swahili (`swa`), Telugu (`tel`), Turkish (`tur`), and Urdu (`urd`).

[3] Languages that are not included in subtask 3: Burmese (`mya`), Italian (`ita`), Polish (`pol`), and Russian (`rus`).

Text	Subtask 1	Subtask 2					Subtask 3					
	Polarization	P	R/E	Re	G/S	O	S	V	D	EL	LE	I
17 political cartoons tackle Donald Trumps	0	0	0	0	0	0	0	0	0	0	0	0
All because of Democrat open borders. FJB	1	1	0	0	0	0	1	0	1	0	0	0
Yes. A fascist to the core.	1	1	0	0	0	0	0	1	0	1	0	1

Table 2: Examples of texts and corresponding labels (0: *No*; 1: *Yes*) from POLAR. Categories for subtask 2: P (*Political*), R/E (*Racial/Ethnic*), Re (*Religious*), G/S (*Gender/Sexual*), O (*Other*). Categories for subtask 3: S (*Stereotype*), V (*Vilification*), D (*Dehumanization*), EL (*Extreme language*), LE (*Lack of empathy*), I (*Invalidation*).

3 Methods

In this section, we present the general framework on which our systems are based (Section 3.1), the specific model decisions taken for the languages in the dataset (Section 3.2), and the different configurations of our systems (Section 3.3).

3.1 Multi-task Learning Framework

Our approach relies on multi-task learning (Caruana, 1997), motivated by the observation that the three POLAR subtasks are strongly related. Multi-task learning is a paradigm that enables a model to learn shared representations across tasks, so that information from one task can support the learning of the other ones. We experiment with different combinations of POLAR subtasks when training our systems to understand if and when multi-task learning is beneficial to each subtask.

In all our configurations, we use widespread encoder-only transformer-based (Vaswani et al., 2017) models as shared encoders, motivated by their higher computational efficiency compared to LLMs, and a separate layer for each subtask. Different subtasks can therefore benefit from mutual signals encoded by shared contextualized representations that are fine-tuned during training. The encoder receives in input a text instance (i.e., a short sentence) and encodes it using byte-pair encoding. The output label is returned by the task-specific layers (i.e., one for each class of each subtask; see Section 4.1) that operate on the contextual embeddings of the special `[CLS]` token and consist of a linear classification layer. In Figure 1, we present an overview of our framework.

3.2 Language-specific Model Choices

We adopt two different strategies for choosing encoder-based models to balance performance and experimental consistency across languages. For English, Italian, Spanish, and Polish, we employ well-established monolingual encoder-based models. For the remaining languages, we instead use a single multilingual encoder-based model as a backbone to ensure architectural consistency across diverse linguistic settings. Our hypothesis is that language-specific models can lead to better performance, however, training a different model for each language can be expensive. This double approach allows us to perform a first comparison between the monolingual and multilingual approaches to understand pros and cons of both. For this shared task, due to time and computational constraints, we only apply the monolingual approach on a small subset of languages for which we have well-established models. Details about the specific encoder-based models can be found in Section 4.1.

3.3 Model Configurations

Our hypothesis is that the information from each subtask can benefit other subtasks. For instance, knowing whether a sentence contains polarized opinions (subtask 1) can facilitate the identification of the polarization type (subtask 2) and the way in which polarization is expressed (subtask 3). Indeed, if a sentence does not present polarized content, there can be neither polarization type nor polarization manifestation. The type and manifestation of polarization are also often related (e.g., ethnic polarization is often expressed through stereotyping, and religious polarization through vilification, as shown in Table 7 in Appendix B). Moreover, both subtasks indirectly signal whether polarization is actually present in a text. Motivated by these subtask inter-relations, we therefore designed our model configurations as follows:

Single task learning configurations We fine-tune encoder-based models on data for each subtask separately as our baselines. We refer to these single task baselines as `[S1]`, `[S2]`, and `[S3]`.

Multi-task learning configurations We experiment by fine-tuning multi-task learning models on different combinations of subtasks and corresponding data: `[S1, S2]`, `[S1, S3]`, and `[S1, S2, S3]`. We include `S1` in each configuration as the main task, since we argue that it can provide the greatest support for learning the other related subtasks.

4 Experiments and Results

In this section, we describe the experimental setup (Section 4.1), then we report results (Section 4.2) and analyze and discuss them (Section 4.3).

4.1 Experimental Setup

We conduct our experiments using the MaChAmp v0.4.2 toolkit (van der Goot et al., 2021). In the multilingual case, we use XLM-RoBERTa (Conneau et al., 2020) as shared encoder and aggregate the subsets for all available languages to create `train` and `dev` sets. For English, Italian, Spanish, and Polish we instead rely on widely-used language-specific models – i.e., RoBERTa (Liu et al., 2019), AlBERTo (Polignano et al., 2019), BETO (Cañete et al., 2023), and Polbert (Kłeczek, 2020), respectively – and fine-tune our multi-task learning systems on the `train` data subset of the corresponding language.[4] For specific versions of encoder-based models, we refer to Appendix C.

We use the default MaChAmp hyperparameters (Appendix C) and fine-tune all models for up to 10 epochs except for RoBERTa-large, for which we increase the maximum number of epochs to 15.[5] We select each model at the epoch in which the macro F_1 score (i.e., the official evaluation metric for POLAR) on the `dev` set reaches the highest value. To deal with under-represented classes in the data and give equal importance to all of them during fine-tuning, we use a cross-entropy loss with balanced class weights. The multi-task learning loss is calculated as $L = \sum_t \lambda_t L_t$, where L_t is the loss for the task t, while λ_t is the corresponding weighting parameter. Multi-label classification with n non-mutually exclusive classes is implemented by decomposing the task into n independent binary classification problems. This formulation makes it possible to assign a specific loss weight λ to each class. We empirically set the loss weights for each configuration as reported in Appendix C.

Lastly, we observed that the original `dev` set portions for each language were rather small for the sake of model selection (95%-5% `train`-`dev` ratio). Therefore, for each language, we shuffled the `train` data and moved part of them into the `dev` set. This led us to obtain a ≈ 90%-10% `train`-`dev` ratio, as shown in Table 1. For our final models for the POLAR shared task, we merged the `train` and `dev` sets and used the resulting set to fine-tune the best model configurations according to results obtained on the `dev` set, using the model at the same best epoch for the final `test` set prediction.

4.2 Results

In this section, we present the official results on the `test` set for our final models and additional results on the `dev` set to motivate our choices. Moreover, we summarize the final configurations chosen for the submitted systems.

In Table 3, we report the average of the macro F_1 scores of all the different languages, both on our `dev` set and on the `test` set for the single task and the selected multi-task configurations. Per-language results for each subtask on the `test` set are reported in Appendix D, while those on the `dev` set across different model configurations are in Appendix E.

Subtask	Configuration	Avg macro F_1	
		`dev` set	**`test` set**
S1	`[S1]`	0.775	0.765
	`[S1, S2, S3]`	**0.783**	**0.776**
S2	`[S2]`	0.495	**0.516**
	`[S1, S2, S3]`	**0.522**	0.506
S3	`[S3]`	0.416	0.416
	`[S1, S3]`	**0.421**	**0.418**

Table 3: Average of the macro F_1 scores of the different languages for the single and multi-task configurations.

The configuration of the final submitted system for subtask 1 and subtask 2 is `[S1, S2, S3]`,

[4] Note that multi-task learning systems that include subtask 3 (i.e., `[S1, S3]` and `[S1, S2, S3]`) are trained and tested only on the 18 languages for which the corresponding data is available (see Section 2.1). When we refer to the configurations `[S1, S3]` and `[S1, S2, S3]` for the remaining languages (i.e., Italian, Polish, Russian, and Burmese), we mean `[S1]` and `[S1, S2]` with the same hyperparameters of the complete multi-task setting, since annotated data for `S3` is not available.

[5] This choice is motivated by its larger parameter size.

in which all the three subtasks are learnt jointly through multi-task learning. The final task weights (decided empirically and reported in Table 9 in Appendix C) are the following: 1 for S1, 0.2 for each class in S2, and 0.166 for each class in S3.[6] We chose this configuration after comparing the average macro F_1 score obtained by this configuration with those of the other ones (i.e., [S1], [S1, S2], and [S1, S3] for subtask 1, and [S2] and [S1, S2] for subtask 2) on our dev set (see Appendix E). For subtask 3, the final configuration is [S1, S3], which, unlike the previous cases, performs better on our dev set compared to the single task configuration (i.e., [S3]) and the one in which all three subtasks are learned jointly (i.e., [S1, S2, S3]), as shown in Appendix E. The learning weights are the following: 1 for S1 and 1 for each class in S3.[7]

4.3 Analysis and Discussion

In this section, we provide an analysis and discussion of our results to give insights for future work in polarization detection and characterization.

We observe that, in all the subtasks, the multi-task learning approach leads to a higher macro F_1 score on the dev set than the single task strategy. This is also confirmed on the test set, with the exception of subtask 2. This could be due to a major misalignment between train and test sets; for instance, in the case of Italian, we see that in subtask 2 there are no positive instances for the *Political* category in the train (Table 4) and dev sets (Table 5), while these are present in the test set (Table 6). In general, the effect of data imbalance can be amplified when combining more tasks, and for this reason, additional strategies to address class imbalance, such as targeted data augmentation, can be explored.

This issue could also affect the performance of language-specific and multilingual approaches. When adopting language-specific models for specific languages, we see that for English and Spanish this choice leads to a high overall performance, while for Italian and Polish – for which no annotations are available for subtask 3 – we observe worse results in subtask 2 compared to using the XLM-RoBERTa multilingual model, especially for Italian. The misalignment between our train set and the test set is reflected in our predictions on the test set, where no instances have been labeled as *Political* for Italian. The Italian one is an exceptional case, but also other languages show an underrepresentation of some categories, which may be mitigated by employing a multilingual model and using data from multiple languages for fine-tuning. Additional analyses can be conducted to better understand the advantages and disadvantages of the multilingual approach compared to the language-specific one as part of future work. Further experiments can also be done by adopting different encoder-based models as shared encoders, both multilingual and language-specific, and assessing whether the overall performance across languages and subtasks improves.

Besides multi-task learning, in the experimental phase we also explored sequential fine-tuning, in which the model is trained on one task after the other, and its weights are modified accordingly at each fine-tuning stage. We tried starting from subtask 1 – our main task – however, the preliminary results on the dev set did not lead to performance improvements; therefore, we decided to stick to the multi-task learning approach in our experiments.

Another experiment that we performed was to force all the labels for subtask 2 and 3 to 0 when the predicted label for subtask 1 was 0, since with no polarization there can not be any polarization target and type of polarization manifestation. However, we observed no improvement or decrease in the F_1 scores.

Our results show some variability across languages and subtasks, as revealed both by the macro F_1 scores and the positioning in the leaderboard. Training a model jointly on all three subtasks (i.e., the configuration [S1, S2, S3]) is beneficial for subtask 1 and 2, while for subtask 3 the configuration with only two subtasks (i.e., [S1, S3]) performs better than other alternatives. As mentioned in Section 2.1, the interpretation of both the annotation guidelines and the text by the annotators could have been influenced by their cultural and political contexts, making multi-label classification in subtask 2 and 3 even more complex. Moreover, as shown in Appendix B, only a small fraction of the data belongs to each category of subtasks 2 and 3, making these classes more unbalanced than the polarization detection ones – for which about half the text instances present polarized content.

Table 8 in Appendix B shows that most classification mistakes across categories are false negatives.

[6] The other weight combination tested in the [S1, S2, S3] configuration was 1 for S1, 1 for each class in S2, and 1 for each class in S3.

[7] The other weight combination tested in the [S1, S3] configuration was 1 for S1 and 0.166 for each class in S3.

In future work, further fine-tuning can be done to better deal with this unbalanced class distribution.

5 Conclusion

In this paper, we presented our system submitted to the POLAR shared task. We proposed an approach that tackles multiple subtasks jointly, with the goal of having a single system for all subtasks while reducing the use of computational resources. We observed that with a single model we can tackle two subtasks together (i.e., subtask 1 and 2), while for subtask 3 a different model is needed for better performance. Our multi-task learning system represents a viable approach for textual polarization detection and characterization, even though performance can vary across different languages and subtasks. Whether using specific models for specific languages can improve the performance needs further exploration, since some languages (e.g., Italian) could benefit from training on a multilingual dataset. Moreover, the results of different fine-tuning strategies (i.e., multi-task learning and sequential fine-tuning) can be further analyzed to find the best configuration for all subtasks, taking into account their complexity. Future work could also explore the use of additional models, both language-specific and multilingual, to assess if performance can be further improved.

Limitations

For a more systematic comparison between the monolingual and the multilingual approach further experiments should be performed, extending them also to languages other than English, Italian, Spanish, and Polish.

Ethical Considerations

Despite the detailed guidelines, polarization annotation in POLAR data remains inherently subjective and may reflect annotators' perspectives. Models fine-tuned on social media and news data may inherit societal and political biases, which can potentially lead to uneven performance across social groups. Our systems are only intended for research on online polarization dynamics and discourse analysis in an aggregate and anonymized form, and are not used for other purposes such as surveillance or automated moderation. We use publicly-available data released in the context of the POLAR shared task, and no personally-identifiable information was intentionally collected or analyzed.

Acknowledgments

This work has been partially funded by the European Union's Horizon Europe research and innovation programmes under grant agreement No. 101178061 (TWIN4DEM).

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained BERT model and evaluation data. *arXiv preprint arXiv:2308.02976*.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

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.

Dariusz Kłeczek. 2020. Polbert: Attacking polish NLP tasks with transformers. In *Proceedings of the PolEval 2020 Workshop*, Warsaw, Poland. Institute of Computer Science, Polish Academy of Sciences.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zikun Liu, Chen Luo, and Jia Lu. 2024. Hate speech in the internet context: unpacking the roles of internet penetration, online legal regulation, and online opinion polarization from a transnational perspective. *Information Development*, 40(4):533–549.

Raquel Martínez-España, Julio Fernández-Pedauye, José Giner-Pérez de Lucía, Jose Miguel Rojo-Martínez, Kaoutar Bakdid-Albane, and Juan José García-Escribano. 2024. Methodology for measuring individual affective polarization using sentiment analysis in social networks. *IEEE Access*, 12:102035–102049.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016.

SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. 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, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. 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, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 312–317, Bari, Italy. CEUR Workshop Proceedings.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Cass Sunstein. 2002. The law of group polarization. *Journal of political philosophy*.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Pramukh Nanjundaswamy Vasist, Debashis Chatterjee, and Satish Krishnan. 2024. The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Information systems frontiers*, 26(2):663–688.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):1–22.

Appendix

A Definition of Polarization Categories

A.1 Categories for Subtask 2

The list of categories for the polarization type classification subtask is the following:

- **Political polarization (P)**: Focuses on division, intolerance, and conflict between political parties and followers;
- **Racial/Ethnic polarization (R/E)**: Focuses on ethnic identity or racial origin and incites division, intolerance, and conflict between ethnic groups or races;
- **Religious polarization (Re)**: Focuses on religious identity and incites division, intolerance, and conflict between religious followers;
- **Gender/Sexual polarization (G/S)**: Refers to the exclusion, discrimination, and marginalization of individuals based on their gender or sexual orientations;
- **Other (O)**: Polarized texts targeting other groups or identities not covered above, such as economic class, technology, or media.

A.2 Categories for Subtask 3

The list of categories for the polarization manifestation identification subtask is the following:

- **Stereotype (S)**: A generalized belief that attributes specific characteristics to all members of a group, often neglecting individual differences, thereby reducing complex personalities to simplistic and uniform representations;
- **Vilification (V)**: The act of defaming or demonizing a particular group, person, or entity by inciting fear, often through exaggeration, misrepresentation, or biased framing that portrays the subject negatively and harmfully;
- **Dehumanization (D)**: The process of depriving a group or individual of their human qualities or personality by comparing them to animals, machines, or objects, or otherwise denying their humanity, dignity, or individuality;

- **Extreme Language and Absolutism (EL)**: The use of language that is extreme or makes definitive, all-encompassing statements, often involving words like “always”, “never”, “worst”, or “best”, and presenting issues in a dichotomous manner such as “us versus them” or “right versus wrong”;
- **Lack of Empathy or Understanding (LE)**: The absence of compassion or recognition for other viewpoints or experiences in the text;
- **Invalidation (I)**: The act of denying or dismissing the identity and existence of individuals or groups, thereby rejecting their sense of self and their presence.

B Data Details and Analyses

In Table 4, we show statistics on the original `train` set, in Table 5 those on the original `dev` set, while in Table 6 those on the `test` set. In Table 7 and 8 we instead present our classification analyses.

C Hyperparameters and Encoders

We present hyperparameter values and specific encoder-based model versions in Table 9 and 10, respectively. All models were fine-tuned on a single NVIDIA Tesla V100-SXM2-32GB GPU.

D Per-language Results: Submitted System on the Test Set

In Table 11, we present per-language results for our final submitted system on the `test` set.

E Per-language Results: All Configurations on the Dev Set

In Table 12, 13, and 14, we present per-language results of the different configurations tested for subtask 1, 2, and 3, respectively, on the `dev` set.

Lang	Size	Pol	P	R/E	Re	G/S	O	S	V	D	EL	LE	I
amh	3,332	75.6	66.8	25.9	2.0	0.6	24.8	54.6	48.5	13.1	30.6	17.6	16.0
arb	3,380	44.7	23.1	17.2	8.4	10.9	16.7	33.3	37.2	10.9	30.4	17.0	8.1
ben	3,333	42.7	34.0	0.8	2.0	0.5	10.1	6.0	24.1	10.7	4.7	1.9	1.8
deu	3,180	47.5	40.7	18.5	11.1	5.9	13.8	35.8	30.1	14.9	21.8	26.7	16.3
eng	3,222	36.5	35.7	8.7	3.5	2.2	3.9	15.1	26.6	12.1	23.9	11.1	18.2
fas	3,295	74.1	43.9	2.4	9.6	6.0	24.2	13.1	57.5	4.3	16.9	9.9	8.0
hau	3,651	10.7	4.9	3.1	2.5	0.8	0.4	4.3	1.2	3.5	3.0	0.9	0.2
hin	2,744	85.5	73.7	12.1	58.7	11.5	13.1	49.7	65.2	18.2	50.6	56.7	65.7
ita	3,334	41.0	0	22.4	8.5	11.4	0	—	—	—	—	—	—
khm	6,640	90.8	18.3	1.5	3.4	1.7	65.9	68.3	1.5	1.2	2.3	11.0	6.5
mya	2,889	58.2	25.3	5.3	3.1	10.6	45.1	—	—	—	—	—	—
nep	2,005	50.3	17.2	14.0	7.9	5.2	11.8	26.8	31.4	6.6	27.1	10.6	15.0
ori	2,368	28.8	20.9	5.0	6.3	3.3	3.7	10.0	11.7	0.7	13.4	1.6	3.4
pan	1,700	49.4	30.8	5.9	7.9	11.2	8.9	16.2	40.4	22.0	23.9	12.4	24.4
pol	2,391	41.9	36.6	9.0	3.6	4.6	6.5	—	—	—	—	—	—
rus	3,348	30.6	13.9	9.8	4.1	5.6	2.4	—	—	—	—	—	—
spa	3,305	50.2	27.3	18.9	15.9	13.4	13.4	27.5	30.6	8.9	24.2	23.9	10.6
swa	6,991	50.1	2.7	35.5	3.5	2.2	7.9	39.7	41.2	12.8	23.9	29.8	23.4
tel	2,366	53.8	21.6	17.0	9.0	13.3	23.7	11.2	22.7	2.5	13.4	26.3	22.8
tur	2,364	48.9	44.7	16.9	15.2	4.8	4.8	40.8	32.4	10.9	43.1	9.6	4.0
urd	3,563	69.5	67.2	54.4	55.3	51.2	50.7	62.3	64.7	55.6	62.2	56.2	57.2
zho	4,280	49.6	5.9	22.6	2.0	16.9	8.6	30.1	18.5	5.0	8.1	7.9	4.8
Total	**73,681**	**53.1**	**27.4**	**15.9**	**10.3**	**8.5**	**18.6**	**33.6**	**31.1**	**11.5**	**21.9**	**18.8**	**16.4**

Table 4: Original `train` set size for each language and percentage of instances classified as polarized (column *Pol*) and those assigned to each category of subtask 2 (P (*Political*), R/E (*Racial/Ethnic*), Re (*Religious*), G/S (*Gender/Sexual*), O (*Other*)) and subtask 3 (S (*Stereotype*), V (*Vilification*), D (*Dehumanization*), EL (*Extreme language*), LE (*Lack of empathy*), I (*Invalidation*)). "—" indicates languages that are not included in subtask 3 and therefore have no labels for it. The last row reports the aggregated values over all the languages.

Lang	Size	Pol	P	R/E	Re	G/S	O	S	V	D	EL	LE	I
amh	166	73.5	65.1	25.9	1.8	0.6	24.7	56.6	47.0	13.3	30.7	17.5	15.7
arb	169	44.4	24.9	17.2	8.3	10.7	16.6	34.3	39.1	10.7	30.2	17.2	8.3
ben	166	42.2	34.3	0.6	1.8	0.6	10.2	6.0	25.3	10.8	4.8	1.8	1.8
deu	159	47.8	45.3	18.2	11.3	5.7	13.8	37.1	30.2	15.1	22.0	27.7	16.4
eng	160	36.9	36.2	8.8	3.1	1.9	3.8	15.0	24.4	11.9	25.6	11.2	18.1
fas	164	71.3	43.9	2.4	9.8	6.1	24.4	12.8	54.3	4.3	17.1	9.8	7.9
hau	182	11.0	4.9	3.3	2.7	0.5	0.5	4.4	1.1	3.3	2.7	1.1	0.0
hin	137	81.8	68.6	12.4	58.4	11.7	13.1	49.6	69.3	18.2	50.4	56.9	69.3
ita	166	41.6	0	22.3	8.4	11.4	0	—	—	—	—	—	—
khm	332	90.7	18.4	1.5	3.3	1.8	65.7	68.4	1.5	1.2	2.1	10.8	6.6
mya	144	56.2	25.0	5.6	2.8	10.4	45.1	—	—	—	—	—	—
nep	100	51.0	17.0	14.0	8.0	5.0	12.0	27.0	31.0	7.0	27.0	11.0	15.0
ori	118	29.7	21.2	5.1	5.9	3.4	3.4	10.2	9.3	0.8	13.6	1.7	3.4
pan	100	47.0	31.0	6.0	8.0	11.0	9.0	19.0	39.0	22.0	24.0	12.0	24.0
pol	119	42.0	37.0	9.2	3.4	4.2	6.7	—	—	—	—	—	—
rus	167	31.1	13.8	9.6	4.2	5.4	2.4	—	—	—	—	—	—
spa	165	50.9	27.3	20.0	15.8	13.3	13.3	33.9	30.9	9.1	24.2	22.4	10.3
swa	349	49.9	2.6	35.5	3.4	2.3	8.0	38.4	41.3	12.9	23.8	29.8	23.5
tel	118	50.0	21.2	16.9	9.3	13.6	24.6	11.0	20.3	2.5	13.6	26.3	22.9
tur	115	48.7	43.5	16.5	14.8	4.3	5.2	40.9	32.2	10.4	45.2	9.6	4.3
urd	177	70.1	69.5	54.2	56.5	51.4	50.8	63.3	65.0	55.9	62.1	56.5	57.1
zho	214	50.9	5.6	24.3	1.9	16.8	8.4	29.9	18.7	5.1	7.9	7.9	4.7
Total	**3,687**	**52.7**	**27.5**	**16.0**	**10.2**	**8.4**	**18.6**	**34.1**	**30.9**	**11.6**	**22.0**	**18.8**	**16.6**

Table 5: Original dev set size for each language and fraction of instances classified as polarized (column *Pol*) and those assigned to each category of subtask 2 (P (*Political*), R/E (*Racial/Ethnic*), Re (*Religious*), G/S (*Gender/Sexual*), O (*Other*)) and subtask 3 (S (*Stereotype*), V (*Vilification*), D (*Dehumanization*), EL (*Extreme language*), LE (*Lack of empathy*), I (*Invalidation*)). "—" indicates languages that are not included in subtask 3 and therefore have no labels for it. The last row reports the aggregated values over all the languages.

Lang	Size	Pol	P	R/E	Re	G/S	O	S	V	D	EL	LE	I
amh	1,501	73.8	67.0	25.9	2.0	0.6	24.8	54.4	46.9	13.1	30.5	17.6	16.0
arb	1,521	44.8	25.2	17.2	8.3	10.9	16.7	33.3	37.7	11.0	30.4	17.0	8.1
ben	1,501	42.2	34.0	0.8	1.9	0.5	10.1	5.9	23.7	10.7	4.7	1.9	1.8
deu	1,432	47.9	41.3	18.5	11.1	5.9	13.8	36.9	30.1	14.9	21.7	26.5	16.2
eng	1,452	36.7	35.7	8.7	3.5	2.3	4.0	15.1	25.8	12.1	23.8	11.1	18.2
fas	1,484	74.1	43.9	2.4	9.6	6.0	24.2	13.1	58.4	4.3	16.8	9.8	8.0
hau	1,644	10.6	4.9	3.2	2.5	0.9	0.4	4.3	1.3	3.6	3.0	0.9	0.2
hin	1,236	85.1	75.6	12.1	58.7	11.4	13.1	49.8	64.7	18.2	50.6	56.7	65.2
ita	1,538	47.3	26.8	9.3	4.5	4.0	14.2	—	—	—	—	—	—
khm	2,988	90.8	18.3	1.5	3.4	1.7	65.9	68.2	1.5	1.2	2.3	11.0	6.5
mya	1,301	57.3	25.3	5.2	3.1	10.6	45.1	—	—	—	—	—	—
nep	903	49.9	17.3	14.1	8.0	5.3	11.7	26.8	31.7	6.5	27.1	10.5	15.0
ori	1,066	28.4	20.9	5.1	6.4	3.4	3.7	9.9	9.1	0.7	13.4	1.6	3.4
pan	809	48.6	30.8	5.8	7.8	11.1	8.9	15.9	38.6	22.0	23.9	12.5	24.5
pol	1,077	41.9	36.6	9.0	3.7	4.6	6.4	—	—	—	—	—	—
rus	1,508	29.8	13.9	9.9	4.0	5.7	2.4	—	—	—	—	—	—
spa	1,488	49.4	27.2	19.4	15.9	13.4	13.4	26.3	30.6	8.9	24.2	24.1	10.6
swa	3,147	50.2	2.7	35.4	3.6	2.2	7.9	39.8	41.2	12.8	23.9	29.7	23.4
tel	1,066	51.8	21.6	17.0	8.9	13.2	23.6	11.3	20.7	2.4	13.4	26.3	22.8
tur	1,093	52.1	42.6	14.9	16.6	9.4	6.8	40.7	33.9	10.0	46.5	12.1	5.4
urd	1,606	69.4	67.6	54.4	55.1	51.2	50.7	62.1	64.6	55.5	62.2	56.2	57.2
zho	1,927	50.8	5.9	23.6	2.0	16.9	8.6	30.1	18.5	5.0	8.1	7.9	4.8
Total	**33,288**	**53.3**	**28.8**	**15.3**	**10.1**	**8.3**	**19.3**	**28.1**	**25.9**	**9.6**	**18.4**	**15.8**	**13.8**

Table 6: `test` set size for each language and percentage of instances classified as polarized (column *Pol*) and those assigned to each category of subtask 2 (P (*Political*), R/E (*Racial/Ethnic*), Re (*Religious*), G/S (*Gender/Sexual*), O (*Other*)) and subtask 3 (S (*Stereotype*), V (*Vilification*), D (*Dehumanization*), EL (*Extreme language*), LE (*Lack of empathy*), I (*Invalidation*)). "—" indicates languages that are not included in subtask 3 and therefore have no labels for it. The last row reports the aggregated values over all the languages.

	Stereotype	Vilification	Dehumanization	Extreme language	Lack of empathy	Invalidation
Political	0.577	0.683	0.273	0.517	0.388	0.371
Racial/Ethnic	0.794	0.708	0.356	0.543	0.513	0.434
Religious	0.717	0.750	0.417	0.656	0.597	0.554
Gender/Sexual	0.715	0.705	0.483	0.622	0.583	0.540
Other	0.645	0.489	0.249	0.359	0.366	0.296

Table 7: Percentage of instances classified in each category of subtask 3 (columns) for each category of subtask 2 (rows). Data refer to the full `train` set comprising all the languages.

Category	FN (%)	FP (%)	TN (%)	TP (%)
Polarization	0.105	0.080	0.388	0.427
Political	0.087	0.056	0.656	0.201
Racial/Ethnic	0.059	0.040	0.807	0.094
Religious	0.033	0.024	0.875	0.068
Gender/Sexual	0.036	0.026	0.891	0.047
Other	0.079	0.044	0.764	0.114
Stereotype	0.102	0.094	0.571	0.234
Vilification	0.094	0.091	0.600	0.215
Dehumanization	0.071	0.027	0.858	0.044
Extreme language	0.102	0.066	0.714	0.118
Lack of empathy	0.092	0.069	0.742	0.097
Invalidation	0.081	0.048	0.787	0.084

Table 8: Percentage of false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP) for all the polarization categories across the three subtasks on the full `test` set comprising all the languages. Percentages refer to predictions using the models in our final submission (i.e., `[S1, S2, S3]` for subtask 1 and 2, and `[S1, S3]` for subtask 3).

Configuration	Hyperparameter	Value
`[S1]`, `[S2]`, `[S3]`, `[S1, S2]`, **`[S1 S3]`**, **`[S1, S2, S3]`**	Optimizer	AdamW
	β_1, β_2	0.9, 0.99
	Dropout	0.2
	Epochs	10, 15
	Batch size	32
	Learning rate (LR)	0.0001
	LR scheduler	Slanted triangular
	Decay factor	0.38
	Cut fraction	0.2
`[S1]`, `[S2]`, `[S3]`	`S1` per-class loss weight	1
	`S2` per-class loss weight	1
	`S3` per-class loss weight	1
`[S1, S2]`, **`[S1, S3]`**	`S1` per-class loss weight	1
	`S2` per-class loss weight	1
	`S3` per-class loss weight	1
`[S1, S2, S3]`	`S1` per-class loss weight	1
	`S2` per-class loss weight	0.2
	`S3` per-class loss weight	0.166

Table 9: Hyperparameters' values used for the different model configurations. We indicate in bold the configurations corresponding to the final submitted system.

Language(s)	Encoder-based model version
amh, arb, ben, deu, fas, hau hin, khm, mya, nep, ori, pan rus, swa, tel, tur, urd, zho	xlm-roberta-base
eng	roberta-large
ita	m-polignano/bert_uncased_L-12_H-768_A-12_italian_alb3rt0
pol	dkleczek/bert-base-polish-uncased-v1
spa	dccuchile/bert-base-spanish-wwm-uncased

Table 10: Encoder-based model versions used in our single and multi-task learning models, divided by language.

Language	Subtask 1	Subtask 2	Subtask 3
amh	0.759	0.467	0.409
arb	0.799	0.542	0.500
ben	0.824	0.297	0.106
deu	0.675	0.446	0.389
eng	0.798	0.446	0.437
fas	0.813	0.553	0.372
hau	0.800	0.298	0.045
hin	0.782	0.713	0.729
ita	0.469	0.141	—
khm	0.741	0.642	0.301
mya	0.866	0.700	—
nep	0.901	0.766	0.666
ori	0.759	0.460	0.190
pan	0.762	0.396	0.464
pol	0.772	0.458	—
rus	0.754	0.505	—
spa	0.745	0.623	0.442
swa	0.793	0.470	0.486
tel	0.889	0.230	0.242
tur	0.750	0.519	0.414
urd	0.764	0.746	0.780
zho	0.859	0.726	0.559
Average	**0.776**	**0.506**	**0.418**

Table 11: Per-language macro F_1 scores on the test set for our final submitted system. "—" indicates languages that are not included in subtask 3 and therefore cannot be evaluated.

Language	[S1]	[S1, S2]	[S1, S3]	[S1, S2, S3]
amh	0.731	0.742	0.715	0.708
arb	0.770	0.758	0.798	0.808
ben	0.806	0.815	0.820	0.812
deu	0.636	0.724	0.691	0.696
eng	0.789	0.791	0.794	0.786
fas	0.866	0.830	0.849	0.874
hau	0.760	0.738	0.785	0.765
hin	0.808	0.791	0.817	0.767
ita	0.641	0.693	0.641	0.669
khm	0.599	0.695	0.652	0.720
mya	0.887	0.837	0.887	0.899
nep	0.863	0.882	0.873	0.864
ori	0.712	0.745	0.721	0.743
pan	0.824	0.795	0.805	0.814
pol	0.780	0.753	0.780	0.738
rus	0.797	0.720	0.797	0.753
spa	0.741	0.740	0.729	0.732
swa	0.796	0.812	0.812	0.844
tel	0.873	0.865	0.881	0.881
tur	0.738	0.795	0.729	0.770
urd	0.750	0.704	0.750	0.726
zho	0.871	0.854	0.880	0.862
Average	**0.775**	**0.776**	**0.782**	**<u>0.783</u>**

Table 12: Per-language macro F_1 scores on our dev set for subtask 1 with different model configurations. Averaged results are in bold, whereas the best configuration selected for our official submission on the test set is underlined.

Language	[S2]	[S1, S2]	[S1, S2, S3]
amh	0.458	0.449	0.354
arb	0.509	0.557	0.515
ben	0.304	0.198	0.432
deu	0.445	0.473	0.486
eng	0.438	0.475	0.462
fas	0.639	0.649	0.583
hau	0.168	0.200	0.183
hin	0.798	0.772	0.767
ita	0.389	0.385	0.407
khm	0.583	0.675	0.653
mya	0.607	0.590	0.536
nep	0.710	0.748	0.728
ori	0.434	0.552	0.595
pan	0.304	0.389	0.375
pol	0.527	0.516	0.545
rus	0.468	0.408	0.576
spa	0.642	0.625	0.619
swa	0.463	0.486	0.452
tel	0.127	0.311	0.286
tur	0.447	0.503	0.546
urd	0.765	0.687	0.719
zho	0.667	0.710	0.671
Average	**0.495**	**0.516**	**<u>0.522</u>**

Table 13: Per-language macro F_1 scores on our dev set for subtask 2 with different model configurations. Averaged results are in bold, whereas the best configuration selected for our official submission on the test set is underlined.

Language	[S3]	[S1, S3]	[S1, S2, S3]
amh	0.413	0.386	0.315
arb	0.404	0.436	0.427
ben	0.134	0.147	0.089
deu	0.410	0.385	0.330
eng	0.447	0.454	0.476
fas	0.365	0.362	0.276
hau	0.000	0.033	0.056
hin	0.753	0.794	0.755
khm	0.269	0.306	0.203
nep	0.644	0.662	0.555
ori	0.175	0.139	0.110
pan	0.522	0.531	0.463
spa	0.449	0.459	0.449
swa	0.513	0.523	0.511
tel	0.302	0.279	0.206
tur	0.336	0.344	0.375
urd	0.767	0.756	0.750
zho	0.583	0.578	0.427
Average	**0.416**	**<u>0.421</u>**	**0.376**

Table 14: Per-language macro F_1 scores on our dev set for subtask 3 with different model configurations. Averaged results are in bold, whereas the best configuration selected for our official submission on the test set is underlined.