JUAGE at SemEval-2023 Task 10: Parameter Efficient Classification

Jeffrey Sorensen Google Jigsaw sorenj@google.com Katerina Korre University of Bologna aikaterini.korre2@unibo.it John Pavlopoulos

Athens University of Economics and Business annis@aueb.gr

Katrin Tomanek Google Research katrintomanek

@google.com

Nithum Thain Google Research nthain@google.com Lucas Dixon Google Research ldixon@google.com **Léo Laugier** EPFL leo.laugier@epfl.ch

Abstract

Using pre-trained language models to implement classifiers from small to modest amounts of training data is an area of active research. The ability of large language models to generalize from few-shot examples and to produce strong classifiers is extended using the engineering approach of parameter-efficient tuning. Using the Explainable Detection of Online Sexism (EDOS) training data and a small number of trainable weights to create a tuned prompt vector, a competitive model for this task was built, which was top-ranked in Subtask B.

1 Introduction

Misogyny is a growing problem online: inflicting harm on the women who are targeted, making online spaces inaccessible and unwelcoming, and perpetuating social asymmetries and injustices. Automated tools are now widely deployed to find, and assess sexist content at scale but most only give classifications for generic, high-level categories, with no further explanation. The Explainable Detection of Online Sexism (EDOS) SemEval Task 10 was introduced to address this problem (Kirk et al., 2023), challenging English-language systems to flag what is more generally sexist content and also explaining why it is sexist.

The JUAGE team participated by building one ensemble with BERT models for the most general subtask (A) and for the more detailed subtasks with taxonomies for sexism (B and C) submissions were made by tuning a PaLM derived model (Chung et al., 2022a), a 62-billion parameter large Pre-trained Language Models (PLM). The PaLM model variant was instruction-tuned (Chung et al., 2022b) and optimized for inference using quantization (Prato et al., 2020).

While it is possible to fine-tune these large language models directly on user tasks, parameterefficient tuning techniques keep the large language model weights fixed and enable sharing of resources across tasks and massively reduce the number of parameters to be updated and stored. The smaller number of parameters that are updated also enable the rest of the model to act as a regularizer, allowing the model to learn from much smaller datasets (Mozes et al., 2023).

Prompt engineering (Brown et al., 2020) is another popular way to use PLMs. However, given the constraints of the model input buffer, only a few training examples may be used (i.e., few-shot prompts for in-context learning). A recently proposed parameter-efficient alternative is to prepend pseudo-tokens to the input and then adjust the embedding weights of these tokens ("soft prompt learning"). This combines the resource sharing of a large, frozen model with a procedure for optimizing weights over a training set. This technique has come to be known as prompt-tuning (Lester et al., 2021). We opted for prompt-tuning in our submissions, except for Subtask A, where we opted for fine-tuning BERT. Follow-up experiments showed, however, that prompt tuning would also have been better for this subtask.

2 Background

Hateful language, such as hate speech and misogyny, is very present nowadays also due to its propagation through social media. Its subjective nature makes hateful language detection a multi-faceted task. Therefore, most existing datasets are taskspecific while the lack of common benchmarks leads the researchers to compile their own datasets (Schmidt and Wiegand, 2017). Detection models, however, trained on these datasets, do not generalize well (Ludwig et al., 2022), and unintended biases can impair the models' performance (Waseem, 2016; Wiegand et al., 2019; Al Kuwatly et al., 2020). In this paper, we discuss these aspects with regard to misogyny.

2.1 Misogynistic language detection

Misogynistic (and more broadly sexist) comments are very common in social media and their automatic detection can be a challenging task as they can appear in various forms (Schütz et al., 2022). Sexism and misogyny detection in written language has been the focus of many shared tasks; in the EVALITA edition of the Automatic Misogyny (AMI) shared task (Fersini et al., 2018) participants had to classify misogynistic comments in English and Italian for Subtask A, and further classify them into seven discreet categories for Subtask B: discredit, stereotype, objectification, sexual harassment, threats of violence, dominance, and derailing. In the IberEval version of AMI (Anzovino et al., 2018) the focus was on English and Spanish comments. Similarly, the sEXism Identification in Social neTworks (EXIST) shared task (Rodríguez-Sánchez et al., 2021-09, 2022-09), focused on English and Spanish comments from Twitter and Gab.com and involved two main subtasks: sexism identification and sexism categorization, in which participants had to categorize the sexist comments into the following categories: ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and nonsexual violence. Finally, misogyny in multimodal settings was explored with the Multimedia Automatic Misogyny Identification (MAMI) shared task (Fersini et al., 2022), which attempted the automatic detection of misogynous memes online. The categories used for the second task of the challenge were s stereotype, shaming, objectification, and violence.

2.2 The Task

The SemEval-2023 Task 10: Explainable Detection of Online Sexism (Kirk et al., 2023) defined sexism as "any abuse or negative sentiment that is directed towards women based on their gender, or based on their gender combined with one or more other identity attributes (e.g. Black women, Muslim women, Trans women)". The task introduced three hierarchical subtasks: sexism detection (A), sexism classification (B), and fine-grained sexism classification (C). Subtask A was binary concerning whether a post is sexist or not. Subtasks B and C were multi-class, the former with four and the latter with eleven classes, summarised in Table 1.

The shared dataset consisted of one training, one development, and one unlabeled test set, all with

	Task B		Task C
1.	threats	1.1	harm
		1.2	incitement
2.	derogation	2.1	attacks
		2.2	aggression
		2.3	dehumanization
3.	animosity	3.1	insult
		3.2	stereotypes
		3.3	belittling
		3.4	patronization
4.	prejudice	4.1	mistreatment
		4.2	discrimination

Table 1: Simplified single-word category labels ("verbalizers") used during prompt-tuning.

posts in English.

3 System overview

Our method is primarily based on agile classification via prompt tuning on large PLMs. We did not use any misogyny-specific resources beyond those provided by the task organizers.

3.1 Subtask A

For Subtask A, we used KTRAIN,¹ a lightweight Keras wrapper that can be used to fine-tune models such as BERT.

3.2 Subtasks B & C

When using large PLMs, we were concerned that the existence of multi-word category labels might create a potential bias against the categories that required a few tokens. As such, we summarized the multi-word categories using single-word appropriate names. For example, instead of Casual use of gendered slurs, profanities & insults, we used "insult". All the names we used for the categories and subcategories are shown in Table 1, but we note that this change did not produce measurable differences in practice. While Gu et al. (2022) found that the choices of verbalizers mattered when training prompt-tuning on sentiment analysis with few-shot examples, the particular choice of SemEval-2023 Task 10 categories was probably not a factor, as we observed little performance difference when some of the category names were misspelled.

¹https://github.com/amaiya/ktrain

Our final submissions for subtasks B and C were ensembles of six prompt-tuned models, whose parameters only differ at the level of the soft-learned prompts. Pseudoprompt tokens are a sequence of "words" that have randomly assigned initial embeddings, and it is these embeddings alone that are adjusted via gradient back-propagation. In addition to the 5-token pseudoprompt that was used during tuning, we prepended to the input both natural language instructions and few-shot prompting, specifically using the examples provided on the competition website, but neither of these demonstrated improvement over using just the pseudoprompt tokens.

For the models where we used instructions, the following prompt was inserted before the few shot examples:

Label the following examples of sexist language according to the following categories and labeled subcategories threats (harm or incitement), derogation (attacks, aggression, or dehumanization), animosity (insult, stereotypes, belittling or patronizing), or prejudice (mistreatment or discrimination).

And the few shot examples were chosen based on the annotated examples provided by the task hosts.

4 Experimental setup

4.1 Subtask A

We fine-tuned BERT for 7 epochs, using patience of 5, and a max-length of 50 tokens. We trained 3 different models by changing the random state value during the train, validation, and test split, and built an ensemble by majority voting.

4.2 Subtasks B & C

For soft prompt-tuning, we set the tuneable prompt token length to 5. The embeddings of the soft prompt tokens have the dimensionality of the underlying PaLM model, leading to 8,192-dimensional vectors. Each prompt was initialized with a random sample of vocabulary token embeddings from the model's 5,000 most frequent tokens. Tuning of the prompt was held based on the accuracy achieved on the development set that was provided by the organizers. **The ensemble** for Subtask C was based on four models (i.e., M3 to M6). The predicted class of M3, M4, and M5 was defined to be the one with the maximum probability (i.e., using argmax). For M6, we also tuned a threshold per class, employing the class with the maximum probability score when several classes were returned. For Subtask B, we combined the labels returned by M1 and M2 with ones returned by models M3 to M6 (e.g., from class 2.1 predicted for C, we infer class 2 for B). The final label was defined using majority voting.

5 Results

5.1 Ranking

The macro F1 of our BERT-based ensemble was 0.817, ranking our team at the 58th place out of the 84 teams of Subtask A. This rank can be explained by the fact that BERT-based models are well-established and commonly adopted by participants. Therefore, our choice to not undertake extensive hyperparameter searching experiments probably kept our team from a better score. Noticeable is the standard deviation of the results, with those above our rank being 0.017 while those below us being 0.095, more than five times higher.

Our prompt-tuning-based ensemble was ranked 7th for Subtask C and 1st for Subtask B. Figure 3 shows the confusion matrix for Subtask B. Although false classifications exist, the diagonal holds higher counts. The classifier confuses more the categories of derogation and animosity, more often predicting as derogation what was animosity. The confusion between the two classes can be further studied in Figure 4. The 'attacks' class (2.1) is confused with that of 'stereotypes' (3.2) while 'aggression' (2.2) is confused with 'insult' (3.1).

5.2 Decomposing the ensemble

The ensemble achieved the highest macro average F1 in both subtasks when compared to the individual member models. The F1 score per model is shown in Figure 1, where it is shown that the ensemble doesn't necessarily outperform all the members. For example, M4 scores higher for threats (a; class 1) and specifically for incitement (b; class 1.2). When focusing on the classification threshold tuning for the ensemble (Section 4.2), we can see that M6_tuned is better for some (e.g., 1.2) and worse for others (e.g., 4.1). In Subtask B, models with "_C" in the name infer their decision from the prediction for Subtask C (Section 4.2), a technique









Figure 1: F1 (%) per model (horizontally) per class









Figure 2: F1 (%) per class (horizontally) per model

that led to slightly better results compared to M3 and M5.

Figure 2 presents model performance aggregated per class. In Subtask B (a), the class of threats was easier to detect across models, followed by derogation. In Subtask C (b), the most difficult class was 3.4, followed by 3.3. This is surprising because the models handle equally well animosity and prejudice, but when focusing on specific subtypes of the two, prejudice is easier to capture.



Figure 3: Task B Confusion Matrix on the test set



Figure 4: Task C Confusion Matrix on the test set

5.3 Prompt-tuning for Subtask A

During the final weeks of this evaluation we switched from the BERT system used in Subtask A to the soft-prompt based classifier. This approach has shown to be flexible and robust to classification tasks (Mozes et al., 2023), especially in few-shot learning constraints. As this latter work mostly occurred after the Subtask A submission deadline, we followed prompt tuning also for this subtask, exactly as we did for subtasks B and C, but without building an ensemble. This resulted in an F1 macro score of 0.884 which would have been first place, since the first scored 0.8746, even without using any ensemble. The comparison with our submitted results is shown in Figure 5. Indeed, the parameter-efficient tuning-based classifiers, although resource-intensive by relying on larger PLMs, appear to provide similar benefits to the techniques of ensembles of models.



Figure 5: Subtask A: Comparison of our submitted BERT-based ensemble versus our prompt-tuning model

Annotator id	Samples num.	Task B	Task C
0	29	0.40	0.33
1	156	0.60	0.49
2	207	0.67	0.43
3	190	0.71	0.45
4	188	0.61	0.22
5	143	0.62	0.40
6	137	0.51	0.34
7	33	0.69	0.30
8	150	0.60	0.45
9	195	0.60	0.37
10	117	0.64	0.47
11	35	0.43	0.34
12	33	0.38	0.15
13	196	0.70	0.42
14	177	0.78	0.57
15	193	0.58	0.32
16	127	0.70	0.48
17	137	0.70	0.55
18	36	0.65	0.42

Table 2: Macro-averaged F1 scores per annotator per task (best results in bold)

5.4 Error analysis

Annotator disagreement

Upon the completion of the challenge, the individual annotations were released by the organizers. An initial observation of these annotations, by an expert member of our team, revealed that the annotators very often confused animosity with derogation and vice versa, something that also occurred with our models. In addition, some of the main themes throughout the data include politics and feminism which are inherently controversial. There are also instances of language used by incels, short for involuntary celibates, namely men unsuccessful in finding a sexual partner or significant other (Nagle, 2017). Incel language comprises of its own lexicon and can in many cases stay undetected unless someone is familiar with the jargon. One example is the following: "Even femcels+the hambeats on IT admitt that they could easily get sex if they wanted. Theres nothing you have to do to get sex as female. You dont even have to be female, even transwomen are enough in 2k18." Such discourses highlight the need for thorough annotation guidelines, as well as annotator training.

Per annotator assessment and model bias

Our ensemble's F1 per annotator, shown in Table 2, reaches up to 0.78 for Subtask B and 0.57 for Subtask C, both concerning Annotator 14 (A14). To examine the possibility of bias in our models, we sampled 10 instances out of the test set where A14 disagrees with the rest of the annotators in Subtask B. Then we examined our ensemble's responses and checked whether a specific class was preferred. As expected, most cases concerned the animosity and derogation labels. In 9 out of 10 cases, A14 assigned the wrong label (4 being falsely derogation, 5 falsely animosity), while 7 out of 10 times our ensemble agrees with A14 indicating that our model might be biased, containing beliefs that are closer to A14. For instance, the following example: "Feminine traits in men can be exploited for gain, comfort, being in control; what kind of sexual benefit (the kind that men are generally interested in) can be gained from an androgynous woman?" was labeled as animosity by A14 and the model, while the gold label was derogation. The specific example lies in the border of both those categories, since it conveys a negative generalization for women (derogation) but also implies essential differences between men and women (animosity),

yet the model leans towards animosity possibly giving more weight to the gender differences. There are also examples where all the annotators disagree and the model agrees with A14, such as: "Well no fucking shit because women dont want to be ugly, sweaty, or a fucking CEO. Hmm what the fuck did I just lose a brain cell", where A14 and the model assigned the label animosity while the other annotators chose derogation and prejudiced discussions, with the gold label being derogation. Finally, the example "It is not insulting, it's **expressing love for your friends**... Unlike women that fake even the friendship towards other females... **We love our friends, they hate each other. **" was labeled as derogation by A14 and our model while the gold label was animosity, indicating that the negative generalization aspect of the sentence might be more important for the model and A14.

5.5 Limitations

While this SemEval task applies machine learning to the task of identifying content that is misogynistic in character, the authors have little grounding in this topic and are unable to judge the quality of task data provided, nor equipped to evaluate the taxonomy and annotation guidelines which were not provided at the time of the task. The underlying LLM likely has intrinsic biases that will affect the distribution of errors in ways that are not well captured by the aggregate performance statistics. Using the leaderboard results as an indication of fitness for tasks such as assisting in moderation would require extensive external verification, which is something the authors have not done.

6 Conclusions

The JUAGE team participated in the three subtasks of SemEval-2023 Task 10 regarding explainable detection of online sexism. Our submission was ranked first on the second subtask and 7th on the third. Our analysis shows that prompt-tuning would have won Subtask A also had it been submitted. We left for future work the investigations on the interpretability of the learned soft prompts. For each of their learned pseudo-token's embeddings, Lester et al. (2021) retrieved the nearest neighbor frozen vocabulary token's embdeddings. They suggest that prompt-tuning may learn natural language context specific to a task. We plan to use the same method to interpret our learned prompts as well as compare the learned prompts for each task.

References

- Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse* and Harms, pages 184–190, Online. Association for Computational Linguistics.
- Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022a. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022b. Scaling instruction-finetuned language models.
- Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.
- Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 task on automatic misogyny identification (AMI). In EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples, pages 59–66. Torino: Accademia University Press.

- Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning.
- Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In Proceedings of the 17th International Workshop on Semantic Evaluation, Toronto, Canada. Association for Computational Linguistics.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 29–39, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Maximilian Mozes, Jessica Hoffmann, Katrin Tomanek, Muhamed Kouate, Nithum Thain, Ann Yuan, Tolga Bolukbasi, and Lucas Dixon. 2023. Towards agile text classifiers for everyone.
- Angela Nagle. 2017. *Kill all normies: Online culture* wars from 4chan and Tumblr to Trump and the altright. Zero Books, Winchester, Hampshire, UK.
- Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. 2020. Fully quantized transformer for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1–14, Online. Association for Computational Linguistics.
- Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza Morales, Julio Gonzalo Arroyo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021-09. Overview of exist 2021: sexism identification in social networks.
- Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza Morales, Adrián Mendieta-Aragón, Guillermo Marco Remón, Maryna Makeienko, María Plaza, Julio Gonzalo Arroyo, Damiano Spina, and Paolo Rosso. 2022-09. Overview of exist 2022: sexism identification in social networks.
- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

- Mina Schütz, Jaqueline Boeck, Daria Liakhovets, Djordje Slijepčević, Armin Kirchknopf, Manuel Hecht, Johannes Bogensperger, Sven Schlarb, Alexander Schindler, and Matthias Zeppelzauer. 2022. Automatic sexism detection with multilingual transformer models.
- Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138– 142, Austin, Texas. Association for Computational Linguistics.
- Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.