APatt at SemEval-2023 Task 3: The Sapienza NLP System for Ensemble-based Multilingual Propaganda Detection

Antonio Purificato and **Roberto Navigli**
Department of Computer, Control and Management Engineering
Sapienza University of Rome
purificato.2019135@studenti.uniroma1.it, navigli@diag.uniroma1.it

Abstract

In this paper, we present our approach to the task of identification of persuasion techniques in text, which is a subtask of the SemEval-2023 Task 3 on the multilingual detection of genre, framing, and persuasion techniques in online news. The subtask is multi-label at the paragraph level and the inventory considered by the organizers covers 23 persuasion techniques. Our solution is based on an ensemble of a variety of pre-trained language models (PLMs) fine-tuned on the propaganda dataset. We first describe our system, the different experimental setups we considered, and then provide the results on the dev and test sets released by the organizers. The official evaluation shows our solution ranks 1st in English and attains high scores in all the other languages, i.e. French, German, Italian, Polish, and Russian. We also perform an extensive analysis of the data and the annotations to investigate how they can influence the quality of our systems. We release our code at https://github.com/antoniopurificato/apatt_at_semeval.

1 Introduction

Due to its diffusion on social media platforms, information is being shared in real-time in the modern digital world, in particular following the COVID-19 pandemic. The detection of propaganda has emerged as an important research topic with the recent interest in fake news (Da San Martino et al., 2020). In today's information age, anyone can take advantage of the diffusion potential of Internet to spread propaganda surreptitiously, and this is in fact done by activist groups, businesses, religious institutions, the media, and even ordinary people, reaching vast audiences. Many different strategies are used to "hide" propaganda messages inside standard text, ranging from appealing to the audience's emotions to utilizing logical fallacies.

SemEval-2023 Task 3 (Piskorski et al., 2023) Subtask 3 offers a different way to investigate this problem: given a news article selected from a media source that could potentially spread disinformation, the system is tasked to identify the persuasion techniques adopted in each paragraph. This is a multi-label task at the paragraph level and also multilingual, with news articles in 6 different languages: English, Italian, Polish, German, French and Russian.

In our approach, we adopt transfer learning and, instead of using one single pre-trained language model, we create an ensemble of multiple state-of-the-art architectures. To output the final prediction two different approaches are used. For all the languages except English it is possible to use a maximum of two models, so for every ensemble we take the probabilities of these models and average them to make the final prediction.

In English, instead, where it is possible to use more than two models for the ensemble, we do not average but weight the predictions of each model. We notice that some models have better results on a subset of classes, so we assign them a higher weight on the corresponding classes in the final output. Our contribution is summarized as follows:

- We employ transfer learning by fine-tuning Transformer-based language models on the propaganda dataset.
- We improve the classification results of the ensemble by weighting the predictions of each model.

The remainder of this paper is organized as follows: In Section 2 we review some recent work in the current literature. In Section 3 we provide the details of our method, while in Section 4 we describe the experimental setup. In Section 5 we demonstrate the effectiveness of the proposed approach by showing and discussing its results. We provide conclusions in Section 6.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 382–388

2 Related Work

Propaganda detection is a crucial topic for research in the field of Natural Language Processing because it may be useful for additional applications like spotting fake news on social media. Jiang et al. (2016) used the representation power of graphs to detect strange behaviours of groups of people. They noticed that propaganda techniques are applied when there are strange patterns on the adjacency matrix and in the spectral subspaces of the corresponding graph. Garimella et al. (2018) tried to capture the main style of propaganda with a graph-based three-stage pipeline: they first built a conversation graph about a topic, then they partitioned the graph to identify potential controversies and finally they measured the amount of controversy from features of the graph.

Barrón-Cedeño et al. (2019) developed a technique to determine the "degree" of propaganda in a piece of writing and integrated their model into a website that groups recent articles about the same event based on their propagandistic content. With a similar approach Vorakitphan et al. (2022) implemented a system, called PROTECT, to automatically detect propagandist messages and classify them along with the propaganda techniques employed. PROTECT allows users to input text and retrieve the propagandist spans in the message as output. They used the RoBERTa PLM and then performed a post-processing step to automatically join tokens labelled with the same propaganda technique into the same textual span. Da San Martino et al. (2019) focused on more in-depth analysis, putting forward a more theoretical paradigm for identifying and categorizing propaganda campaigns. They first created a corpus of manually annotated news articles at the fragment level and then they proposed a suitable evaluation measure for this task.

A major limit of all the above cited previous work is their focus on the English language. With this work we aim to study propaganda techniques in a multilingual setup by employing ensemble methods.

3 Method

3.1 Pre-trained Language Models (PLMs)

We worked with six types of Transformer-based (Vaswani et al., 2017) PLMs, with the goal of creating an ensemble of them whenever this was possible. In fact, in some languages, we were not able to find more than one PLM, therefore the predictions were made using the only available one. In all cases, we used the pre-trained models with a layer on top to perform classification. In the following we provide a brief description of the models we used in our proposed solution:

- BERT (Devlin et al., 2019) is a popular Transformer-based PLM, which enables bidirectional training using a "masked language model" (MLM) pre-training objective. It also uses next-sentence prediction (NSP) objective during pre-training in order to understand relationships between sentences.

- RoBERTa (Liu et al., 2019) is an architecture based on BERT, but improved using dynamic masking and a different byte-level Byte-Pair Encoding.

- alBERT (Lan et al., 2020) replaces the next sentence prediction (NSP) loss with a sentence order prediction (SOP) one to better model inter-sentence coherence. In order to reduce memory consumption and computational time it introduces two parameter reduction techniques.

- XLNet (Yang et al., 2019) is based mainly on pre-training and a different parametrization from that of the other models but it also introduces concepts from Transformer-XL (Dai et al., 2019), such as the segment recurrence mechanism and the relative encoding scheme.

- distilBERT (Sanh et al., 2019) is a simpler Transformer model trained by distilling BERT base. It has 40% fewer parameters than BERT and runs much faster, while in many cases keeping performance in the same ballpark.

- HerBERT (Mroczkowski et al., 2021) is a Transformer-based model that was trained for the Polish language, and outperformed multilingual BERT on average on a set of tests.

3.2 Ensemble

Each PLM is fed with the `input_ids` that are the indices of the input sequence tokens in the vocabulary. They also take as input the `attention_masks`, used to prevent the model from looking at padding tokens, and which also result from the tokenization

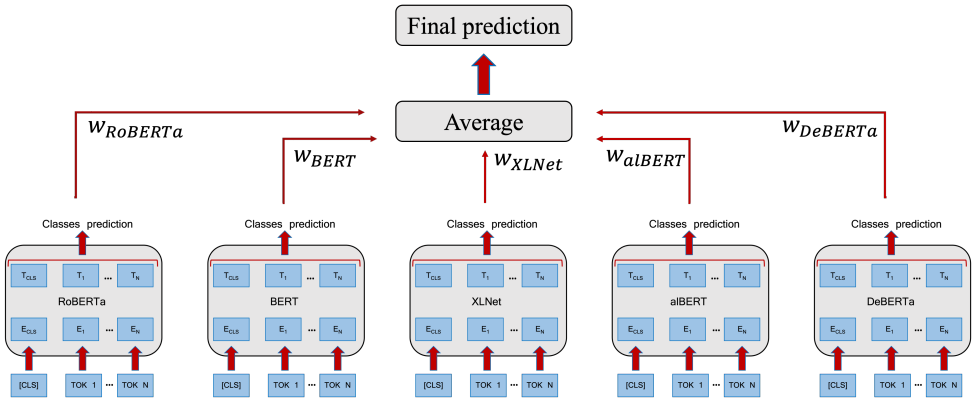

Figure 1: Our ensemble for this task. For English we select the value of the weights as described in Section 4. For the other languages, we use a simple average and we set the weights to 1.

step that is responsible for preprocessing text into the corresponding array of indices.

As we can see from Figure 1, after we train each PLM we create a list containing their predictions. Once all the predictions are available we apply a weighted average. Our APatt architecture works in the following way:

$$x = \sum_{p} w_p \sigma_p \quad (1)$$

where σ_p is the output of the p-th PLM, w_p is the weight given to the corresponding prediction and x is the final result. For all the languages other than English $w_p = 1$, while for the English language the value of the weights is described in Section 4.

3.3 Training

At training time, we minimize the binary cross-entropy (BCE) objective $\mathcal{L}$ as follows:

$$\mathcal{L}(x_c, y_c) = -y_c log x_c - (1 - y_c) log(1 - x_c)$$

where y_c is the label of class c and x_c is the predicted value for class c.

At test time, to predict the labels we classify each class based on the following formula:

$$\tilde{y_c} = \chi(x_c > \tau)$$

where τ is a probability threshold and χ is an indicator function.

4 Experimental setup

4.1 Ensemble setup

For Russian, Polish and Italian we decided to avoid ensembling and use only the output of the PLM, for one or other of the following reasons:

Model	Micro-F1
alBERT	0.293
RoBERTa	0.322
BERT	0.343
DeBERTa	0.345
XLNet	**0.365**

Table 1: Comparison of the performance of the different models for the English subtask on the dev set.

- It was not possible to find more than one PLM for the specific language, or
- The creation of the ensemble did not improve the performance.

For all the other languages we decided to use an ensemble. For English the PLMs used were BERT, RoBERTa, DeBERTa, alBERT, XLNet. For French, we used BERT and RoBERTa. For German we used BERT and distilBERT.

To select the values of the weights for the English language we performed an ablation study for each PLM on the dev set, as shown in Table 1.

XLNet achieves the best results. For this reason, the predictions of XLNet are the most weighted in the ensemble. alBERT has the worst results but is not penalized too much in the final ensemble because in some classes it also attains very high values of micro-F1. In the final ensemble, we used the normalized F1 scores as weights of the corresponding PLMs: BERT 0.15, alBERT 0.17, RoBERTa 0.19, DeBERTa 0.22 and XLNet 0.27.

4.2 Hardware and Hyperparameters

All the experiments were performed on an NVIDIA RTX 3090 with 10752 CUDA cores. We selected

Team	Micro-F1	Macro-F1
APatt (weighted) [1]	**0.39823**	**0.29660**
NLUBot101	0.39737	0.29273
APatt (not weighted)	0.36895	0.26413
NL4IA	0.37937	0.32936
PersuasionNLP4SG	0.37790	0.26514
Baseline	0.16125	0.21735

Table 2: Results on the dev set for the subtask 3 for the English language. Our solution ranks 1st over 18 teams[1].

Team	Micro-F1	Macro-F1
APatt (ours)	**0.37562**	**0.12919**
SheffieldVeraAI	0.36802	0.17194
Appeal for attention	0.36299	0.16621
KInITVeraAI	0.36157	0.13324
Baseline	0.19517	0.06925

Table 3: Official results on the test set for the subtask 3 for the English language. Our solution ranks 1st over 22 teams.

Language	Micro-F1	Macro-F1	Ranking
German	0.48375	0.17692	4/20
Polish	0.36570	0.14969	6/20
Russian	0.30602	0.11666	7/19
French	0.38414	0.19125	8/20
Italian	0.44094	0.16626	9/20

Table 4: The results of our APatt for the other languages.

the *base* version of the PLMs to reduce the computational time. We decided to use the cased models because casing might convey expressions of emotion: for example, if we write one or more words in capital letters it might be because we want to express anger. We tried different learning rates, but the best results were obtained with 3×10^{-5}. We selected a batch size equal to 16 and we trained our systems for 10 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019). We selected a value $\tau = 0.2$ as classification threshold because, after multiple experiments with the dev set, we obtained the best results with this value. The hyperparameters were equal for all the languages.

4.3 Data

The input for all tasks was news and Web articles in plain text format. Articles were given in six languages (English, French, German, Italian, Polish, and Russian) and were collected from 2020 to 2022. They were gathered from various sources and cover a variety of popular subjects, such as COVID-19 or the Russo-Ukrainian War, as well as abortion and migration. They were chosen primarily from the mainstream media, but also from websites that media credibility experts have flagged as possibly disseminating false information.

4.4 Evaluation metrics

This subtask is a multi-label classification task. The official evaluation measure for this task is micro-F1. We also report macro-F1.

5 Results

We first report results on the English language. The introduction of weighting the predictions improved our results. As we can see from Table 2, we ranked 2nd on the dev set when using a standard average, while we achieved the 1st place by applying a weighted average.[1]

Turning to the test set, according to the official results, shown in Table 3, our weighted average system ranked 1st.

For the other languages we obtained good results in German and Polish, as we can see from the performance and the corresponding ranking shown in Table 4.

5.1 Output

From an analysis of the output of our system it can immediately be seen how the results are affected by the distribution of the samples in the dataset. This is shown clearly in Table 5. If we sort the class distribution in the dev set from the least frequent (1) to the most frequent (23), we can see that results are heavily biased towards more frequent classes. Interestingly, however, we notice that we obtain a high F1-Score for *Appeal To Time* without training samples for this class. On the other hand, the *False Dilemma-No Choice* class is in position 16 of the ranking but has an F1-Score of only 0.128. More in general, it is clear that, for a number of classes, there are insufficient training samples to fit the network parameters satisfactorily.

5.2 Data analysis

The training set for the English language is composed of 446 articles, while the dev set of 90 articles. As we can see from the label distribution

[1]This latter result with weighted average was obtained after the competition was closed.

Class	Ranking	F1-Score
Red Herring	10	0.295
Appeal To Time	4	0.462
Repetition	20	0.570
Loaded Language	**23**	**0.574**

Table 5: Comparison between the distribution of samples and the performance of the model on some classes.

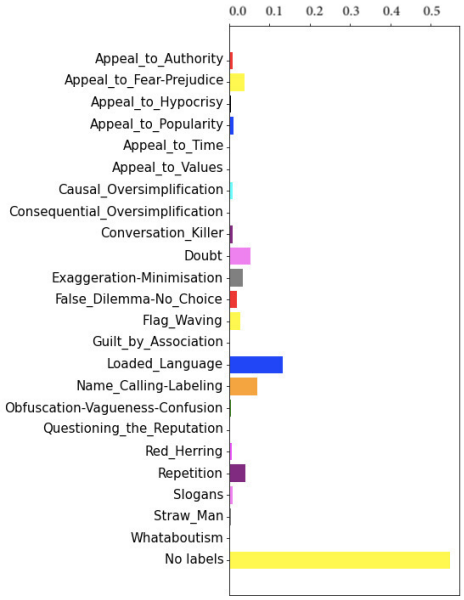

Figure 2: Data distribution of the labels in the English train set.

in Figure 2, 54% of the paragraphs have no labels, 13% contain the *Loaded Language* propaganda technique and 7% contain the *Name calling labelling* propaganda technique. We noticed that for the English language, there are techniques that are not in the training set, such as *Appeal to time* and *Appeal to values*.

Obviously, the distribution of the data depends on the language under consideration. As also discussed in other works (Wu et al., 2019), the dataset imbalance is a problem that needs to be addressed in this task. However, there is also a more serious problem: while the labels for the propaganda techniques were assigned using annotation guidelines, there are multiple cases where there is an ambiguity as to which label to choose, also for a human. Here below we show and discuss two examples in English:

He vowed that London would remain the same after March 29 2019, and said the fireworks display was about "showing the world, while they're watching us, that we're going to carry on being open-minded, outward looking, pluralistic".

The prediction of our system is *Repetition* while the labels are *Repetition, Name Calling Labelling*. From the annotation guidelines, we have the following definition of *Name Calling Labelling*: "a form of argument in which loaded labels are directed at an individual or group, typically in an insulting or demeaning way". The reason why this propaganda technique is assigned to this paragraph is not clear. In fact, it contains no words that recall any type of insult or accusation. Another example is the following:

There is a chance; as unfortunately there are many MPs who don't respect the vote and may just turn on it, but short of that I don't see any way the Conservatives would vote for it, and the majority is slender as it is, as the DUP is bitterly against it, and I can't see the Lib Dems voting for it, so it will only be if there are enough, what I can describe as remoaner MPs, that the deal won't be dead in the water.

The prediction of our system is *Doubt, Loaded Language, Name Calling-Labeling* while the labels are *False Dilemma-No Choice, Loaded Language, Name Calling-Labeling*. Reading this sentence, from the point of view of a human, it is really difficult to understand why *Doubt* is not the right label. In fact, there are some sentences that express doubt, such as, for example, *if there are enough* or *there is a chance*. These types of ambiguous annotations can influence the performance of propaganda detection systems.

6 Conclusion

In this paper, we examined the ability of Transformer-based language models, and in particular weighting the predictions of multiple models, to carry out the task of detecting propaganda techniques in different languages. We showed how the class imbalance of a dataset can influence the performance of a model on this task. Future work includes examining more effective methods for solving the previous issue and also optimizing the value of the weights on the dev set to improve the performance.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

This work was also supported in part by the PNRR MUR project PE0000013-FAIR.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642.

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Inf. Process. Manage.*, 56(5):1849–1864.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *Trans. Soc. Comput.*, 1(1).

Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. 2016. Inferring lockstep behavior from connectivity pattern in large graphs. *Knowledge and Information Systems*, 48:399–428.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISÀ corpus of Italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden. Association for Computational Linguistics.

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.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Trieu H. Trinh and Quoc V. Le. 2019. A simple method for commonsense reasoning.

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.

Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. PROTECT: A Pipeline for Propaganda Detection and Classification. In *CLiC-it 2021- Italian Conference on Computational Linguistics*, Milan, Italy.

Zhenghao Wu, Hao Zheng, Jianming Wang, Weifeng Su, and Jefferson Fong. 2019. BNU-HKBU UIC NLP team 2 at SemEval-2019 task 6: Detecting offensive language using BERT model. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 551–555, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Adrian Zbiciak and Tymon Markiewicz. 2023. A new extraordinary means of appeal in the polish criminal procedure: the basic principles of a fair trial and a complaint against a cassatory judgment. *Access to Justice in Eastern Europe*, 6(2):1–18.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*.

A Pre-trained Language Models

All the selected PLMs were available on huggingface [2].

For the English language BERT was pretrained on the BookCorpus (Zhu et al., 2015) and the English Wikipedia, like alBERT. RoBERTa on BookCorpus (Zhu et al., 2015), CC-News, OpenWebText (Radford et al., 2019), Stories (Trinh and Le, 2019) and English Wikipedia. XLNet was pretrained on BookCorpus (Zhu et al., 2015), English Wikipedia, Giga5 (Zbiciak and Markiewicz, 2023), CC-News and ClueWeb. DeBERTa was pretrained on English Wikipedia, BookCorpus (Zhu et al., 2015), OpenWebText and Stories (Trinh and Le, 2019).

For the Italian Language BERT was pretrained on OPUS (Tiedemann, 2012) and OSCAR (Abadji et al., 2022), while DistilBERT on PAISÁ (Lyding et al., 2014) and ItWaC corpora.

For the Russian language both RoBERTa and BERT were pretrained on the Taiga corpus.

For the Polish language BERT was pretrained on Polish Wikipedia, a Polish Parliamentary corpus, OPUS (Tiedemann, 2012) and OSCAR (Abadji et al., 2022).

For the French language BERT was pretrained on the French Wikipedia corpus. RoBERTa was pretrained on the French Wikipedia corpus and on CC-News.

Finally for the German language BERT was pretrained on the German Wikipedia and OpenLegalData corpora, as DistilBERT.

B Ensemble

As shown in the paper the introduction of the ensemble improved our results in all the languages. While we showed the improvements only for the English language, we achieved better results also in French and German, as we can see from Table 6.

Language	No Ensemble	Ensemble
French	0.40881	**0.43783**
German	0.40900	**0.43288**

Table 6: Official results on the dev set for Subtask 3 for the French and German languages with and without ensemble. No Ensemble and Ensemble are the values of Micro-F1 without and with ensemble respectively.

[2] https://huggingface.co