TeamAmpa at SemEval-2023 Task 3: Exploring Multilabel and Multilingual RoBERTa Models for Persuasion and Framing Detection

Amalie Brogaard Pauli[1] **Rafael Pablos Sarabia**[1] **Leon Derczynski**[2] **Ira Assent**[1]
[1]Department of Computer Science, Aarhus University, Denmark
[2]IT University of Copenhagen, Denmark
{ampa,rpablos,ira}@cs.au.dk, ld@itu.dk

Abstract

This paper describes our submission to the SemEval 2023 Task 3 on two subtasks: detecting persuasion techniques and framing. Both subtasks are multi-label classification problems. We present a set of experiments, exploring how to get robust performance across languages using pre-trained RoBERTa models. We test different oversampling strategies, a strategy of adding textual features from predictions obtained with related models, and present both inconclusive and negative results. We achieve a robust ranking across languages and subtasks with our best ranking being nr. 1 for Subtask 3 on Spanish.

1 Introduction

Task 3 at SemEval-2023 (Piskorski et al., 2023) promotes research on propaganda detection by providing a shared task with annotated data for detecting framing (Subtask 2) and persuasion techniques (Subtask 3) in online news from various languages; namely, English, French, German, Italian, Polish, and Russian. In addition, three additional surprise languages were included in the test phase: Spanish, Greek, and Georgian.

We have submitted predictions for each of the nine languages on both Subtask 2 and Subtask 3, with both subtasks being multi-label classification problems. We obtain robust results across the different languages by fine-tuning RoBERTa models (Liu et al., 2019). Our best rankings on Subtask 3 are nr. 1 on Spanish and nr. 2 on French and Russian. On Subtask 2, we rank nr. 2 on English and Greek.

In our paper, we study how to get good performance on multi-label classification using fine-tuning of transformer-based models on both subtasks and across languages. More concretely, we conduct experiments and report results on the following three questions. First, we explore when to use an English backbone with less annotated data and when to use a multilingual backbone to exploit a larger set of multilingual annotated data. Second, we study different oversampling strategies. Third, we try to directly include textual features in training for persuasion detection of the broader categories of the fallacy of *ethos*, *pathos* and *logos*. We report both positive and negative results. We made our final training scripts available on GitHub[1].

2 Background

2.1 Related work

Detecting persuasion techniques has recently gotten increased attention in the community; It has been addressed in different forms in former Shared Tasks (Martino et al., 2020; Dimitrov et al., 2021). But also in related work e.g. detecting logical fallacies (Habernal et al., 2017; Jin et al., 2022) and personal attacks in debates/communication (Habernal et al., 2018; Sheng et al., 2021). The study of Pauli et al. (2022) tries to reframe different categories from such previous studies into the fallacy of ethos, pathos and logos. In addition, they train models on a combination of re-grouped labels from different data-sources. We try to apply these models in our experiments. Note, these models uses data from the study of Martino et al. (2020) which has a data overlap with the current task.

Framing detection is concerned with analyzing how a news article is presented by the media in terms of perspective and choice of what is emphasised. It is previously computationally studied in Card et al. (2015) with a set of defined categories.

In the training of our system, we use oversampling. It is a technique that has been considered frequently to avoid class imbalance. When there is class imbalance, there is a bias towards the majority class and traditional models face difficulties in correctly classifying the minority class (Gosain

[1] https://github.com/AmaliePauli/semEvalPersuasion

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 847–855
July 13-14, 2023 ©2023 Association for Computational Linguistics

and Sardana, 2017). One mitigation approach is to balance the dataset before training either by undersampling which is to remove instances from the majority class or oversampling which is to repeat minority instances in the training set (Batista et al., 2004; Chawla et al., 2004).

2.2 Task data

We analyze the task training data and highlight the properties that form the basis of our experimental setup. Subtask 3 is concerned with detecting persuasion techniques at the paragraph level with a total of 23 possible classes. Subtask 2 is about framing detection on the article level with 14 different classes. Both subtasks are multi-label classification tasks, and the training data includes the following six languages English (en), German (ge), French (fr), Italian (it), Polish (pl)[2], and Russian (ru). The data consists of news and webpages collected between 2020-2022 covering a variety of topics. See Piskorski et al. (2023) for more details.

The training size samples per language for Subtask 3 range between 1555 (ge) and 9498 (en) paragraphs, with English having considerably more training data than the other languages. However, the average number of labels per paragraph on the English data is 0.6, which is lower than for the other languages, see also Table 5 in the appendix. The label distribution per class per language is displayed in Figure 1. We notice a different distribution over the labels per language, but mostly we note a high imbalance in labels both per language and in total. For English, the classes *Loaded Language* and *Name Calling Labelling* are best represented with 1809 and 979 labels, respectively, in the training set. But, at the same time, some classes are not represented in the English training data: *Appeal to Time*, *Appeal to Values*, *Consequential Oversimplification*, and *Questioning the Reputation*. Besides these three classes, the English training data has six more classes with under 50 labels per class in the training set. The English part of the dataset is common to the the dataset in Martino et al. (2020).

The training size samples per language for Subtask 2 range between 132 (ge) and 433 (en) articles, again with English being best represented. Also here, the labels are imbalanced (though not as extremely), as shown in Figure 3 in the appendix.

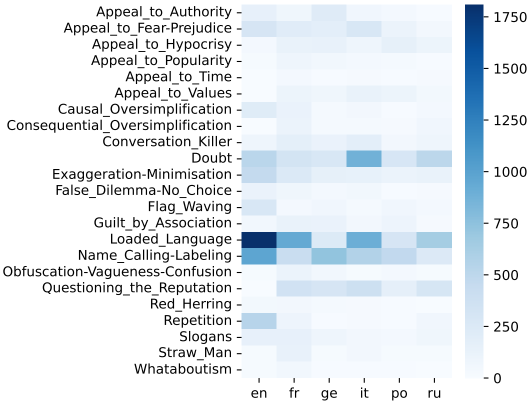

Figure 1: Distribution of labels per language per class for Subtask 3.

We display the total count of labels per class for English and contrast with the total count of labels per class for all Non-English data for Subtask 3 in Figure 2 and for Subtask 2 in the appendix, Figure 4. The figures illustrate the difference in training data available when using only English or training data on all languages.

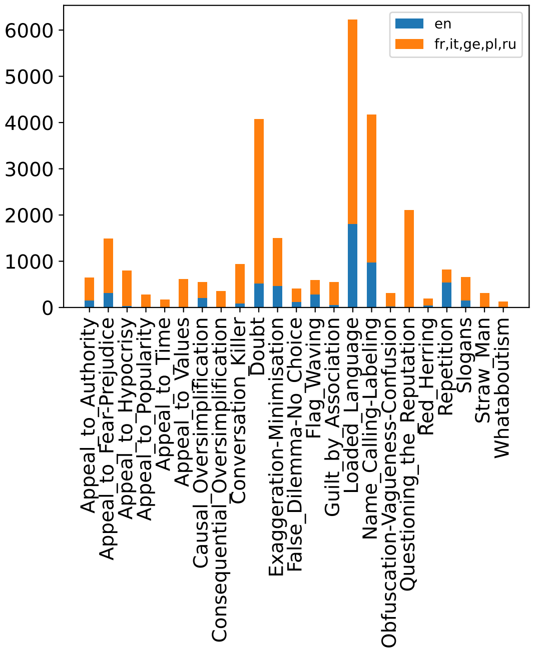

Figure 2: A stacked bar-chart of the total number of labels per class in Subtask 3, split in English and non-English.

3 Overview of our approach

Our approach is to fine-tune models on the Transformer architecture (Vaswani et al., 2017), more concretely, the RoBERTa architecture (Liu et al., 2019). RoBERTa is an optimized pretraining of the BERT architecture (Devlin et al., 2019). A

[2] Note, we use the ISO two-letters language code where the task organizers shorten Polish to 'po'

RoBERTa model is preferred over a BERT model regarding performance on downstream-tasks in English shown by Liu et al. (2019), and Conneau et al. (2020) shows that the multilingual XLM-RoBERTa is prefered over mBert. We, therefore, use the English pre-trained model RoBERTa large (Liu et al., 2019) (Rob), and the multilingual pre-trained model XLM-RoBERTa (XLM-R) (Conneau et al., 2020) in our experiments. The English RoBERTa is pre-trained on 160GB of text from five different datasets. The multilingual XLM-R is trained on a total of 2.5 TB of data from filtered Common-Crawl data including 100 different languages with 301GB of English text data, 278GB of Russian, 67GB of German, 57Gb of French, 45GB of Polish and 30GB of Italian. Hence, we notice a difference for our target languages in how much specific language data the backbone model is trained on. We also note a difference in model size between XLM-R with 550M parameters and RoBERta large with 355M parameters (Conneau et al., 2020).

In our experiments, we examine when to use the English backbone with less annotated training data available and when to use the Multilingual backbone with more annotated data available.

We notice a large difference in the number of training samples for the different languages in the multilingual XLM-Roberta. In our experiments, we test the difference between oversampling the language-specific data for either the language with the highest presence in the pretraining versus the language for the matching target data at hand.

We also oversample the labels with a low count in the training data in an attempt to account for the class imbalance in the training data. When oversampling, we include every sample with the respective label twice in the training data.

We test a model on the English data, which first uses the models from Pauli et al. (2022) to predict the fallacies of ethos, pathos and logos on Subtask 3 paragraphs. We then include these predictions as features in the text input before fine-tuning on Subtask 3. We use the textual features from ethos as a 'credibility attack', from pathos as 'emotional appeal' and from logos as 'logical fallacy'. In this setup, the new text input becomes: "[credibility attack] [emotional appeal] [logical fallacy]: paragraph text", where the [] input depends on whether the respective prediction is positive or not.

4 Experiments

4.1 Training details

We run experiments based on the RoBERTa architecture (Liu et al., 2019) using the implementations from HuggingFace (Wolf et al., 2020) and the wrapper library simpletransformers [3].

In the experimental setup, our default model uses the EN training data when fine-tuning RoBERTa-Large (English setup), and en, fr, it, ge, pl and ru training data when fine-tuning the multilingual backbone XLM-RoBERTa (multilingual setup), (abbreviation XLM-R). In all experiments, we use a batch size of 8, a maximum sequence length of 512, and train for 10 epochs with the use of early stopping. For further details, see training script on our GitHub. The results reported in this Section 4 is on the development set (DEV set).

We test different learning rates. On Subtask 3, we find that an overly large learning rate gives unstable results on both backbone models. Using a learning rate of $4e-5$ and running the models on different seeds results in F1-scores of zero on the DEV set in some runs, and good performance for other runs. We find a learning rate of $1e-5$ to give stable results, and lower settings seem to not learn enough. On Subtask 2, we find a learning rate of $4e-5$ to be stable.

In the following, we study hypotheses on oversampling, backbone, and the use of fallacy predictions by training on the training part and evaluating on the development part of the datasets for Subtask 2 and Subtask 3, respectively. Unless stated otherwise, we examine the F1-micro score, which is the official evaluation metric.

4.2 Oversampling Strategies

We experiment with different strategies for oversampling the training data in the multilingual setup for Subtask 3. We evaluate it on the five non-English DEV datasets, reported in Table 1. In these experiments, we study and compare the performance across the five languages.

Accounting for class imbalance We oversample training data for classes whose support falls under the 40 % fractile of training labels per class. We denote this strategy as 'Few'. We compare it against a model with no sampling (Table 1). We observe an improvement in the F1-macro score in 4 out of 5 languages, indicating that the average performance

[3] https://github.com/ThilinaRajapakse/simpletransformers

	fr		ge		it		pl		ru	
	F1-mic	F1-mac	F1-mic	F1-mac	F1-mic	F1-mac	F1-mic	F1-mac	F1-mic	F1-mac
XLM-R (no sampling)	0.474	0.325	0.462	0.295	0.521	0.247	0.397	0.246	0.408	0.202
XLM-R (few)	0.470	0.325	0.468	0.320	0.523	0.266	0.392	0.265	0.398	0.216
XLM-R (few + en)	0.467	0.325	0.449	0.296	0.527	0.262	0.383	0.286	0.383	0.286
XLM-R (few + ta*)	0.472	0.323	0.459	0.305	0.532	0.248	0.375	0.240	0.434	0.238
XLM-R (few+ta*+en)	0.485	0.345	**0.470**	0.324	0.525	0.255	0.384	0.275	0.418	0.208
Ensemble of above	**0.496**	0.337	0.468	0.305	**0.539**	0.232	**0.401**	0.265	**0.439**	0.228

Table 1: Oversampling strategies in the multilingual setup for subtask 3 on non-English DEV datasets.*ta means oversampling of the training data for the target languages

per class increases. In subsequent experiments, we thus use the ’Few’ strategy.

Priming the model We fine-tune a multilingual XLM-RoBERTa model for each language where we oversample the specific language, such that we use the training data of the target language twice. We compare this strategy with the strategy where we oversample the English data; the English data has most pre-training data for the backbone model (see Section 3; please note, though, that it has almost the same as Russian, but e.g. 10 times that of Italian). In Table 1, we see that oversampling the target language is preferable to oversampling English in 4 out of the 5 languages. However, there is no consistent indication that oversampling the target language should be better than not doing so: in our experiments, it is the case for Italian and Russian, but not for French, German, and Polish. We also experiment with oversampling both the target language and English by a factor of 2, which gives a higher F1-micro score on French and German, but again no consistent trend across languages. More investigation is needed to explain these results.

Ensemble We conclude our oversampling experiments on the non-English data by creating an ensemble over the five models for each language with majority vote. This gives the highest performance in 4 out 5 languages.

4.3 Multilingual or English Backbone

In the following experiments, we focus on the English target data, and we study whether to use the English backbone model with the task-annotated English training data, or instead the multilingual backbone model with the data from all languages. The differences in training sizes and label distribution are reported in Section 2.2 for both Subtask 2 and Subtask 3. In this set of experiments we use the oversampling strategy *Few* for all models. We run the models on five different seeds and report average results in Table 2 for Subtask 2 and Table 3 for Subtask 3.

	F1-Micro	F1-Macro
RoB (few)	0.665 (0.006)	0.382 (0.018)
XLM-R (few)	**0.720** (0.013)	0.433 (0.020)

Table 2: Results on Subtask 2 English DEV. Data averaged over five runs and standard deviation in brackets.

	F1-Micro	F1-Macro
RoB (few)	0.362 (0.004)	0.341 (0.025)
XLM-R (few)	0.344 (0.006)	0.217 (0.046)
Combination	**0.366** (0.005)	0.220 (0.049)
RoB fallacies	0.341 (0.012)	0.326 (0.006)

Table 3: Results on Subtask 3 English DEV. Results averaged over five runs and standard deviation in brackets.

For Subtask 2, we see that the multilingual setup is preferable to the English setup in terms of Micro-F1, but the opposite applies for Subtask 3.

We report the F1-Score per class average over the five runs in Subtask 3 in Table 6 in the appendix. For some classes, XLM-RoBERTa has higher F1-score than English RoBERTa, and vice versa. Not surprisingly, results seem to correlate with the number of English training examples available per class. We therefore adopt a thresholding approach to exploit the better performing models on Subtask 3: we use the prediction for a label from the English model when the number of English training data examples for that label exceeds 150, else, we use the prediction from the multilingual model. We refer to this as ’Combination’. Combination achieves a slightly higher F1-micro score on the DEV set.

4.4 Fallacies of Ethos, Pathos and Logos

We conduct experiments on English subtask 3 where we include the predictions of the fallacy of ethos, logos and pathos from the appeal models as textual features, as outlined in Section 3. The

	Task 2			Task 3			Task 3	
	Submitted (XLM-R Few)			**Submitted (*)**			**XLM-R Few**	
	F1-mic	**F1-mac**	**rank**	**F1-mic**	**F1-mac**	**rank**	**F1-mic**	**F1-mac**
en	0.56696	0.50961	**2**	0.32457	0.15768	7	**0.35951**	0.18471
fr	0.50558	0.47890	5	0.43442	0.30544	**2**	0.42943	0.29257
ge	0.63223	0.57266	4	0.47597	0.26610	5	**0.48811**	0.28106
it	0.59674	0.48268	4	0.52101	0.26355	4	**0.52597**	0.27214
pl	0.61386	0.55541	7	0.38918	0.23640	4	**0.41164**	0.26631
ru	0.40930	0.29380	4	0.37781	0.22740	**2**	0.36324	0.21927
ka	0.51667	0.37879	7	0.40816	0.25854	4	0.37778	0.27851
gl	0.54408	0.44372	**2**	0.23835	0.17135	4	**0.24098**	0.19388
es	0.50575	0.38650	4	0.38106	0.24366	**1**	0.36073	0.19979

Table 4: TeamAmpa's results from the official test leaderboard. We use the ISO two-letters language code. *The predictions come from different models depending on the target language, see Section 5

results are averaged over five runs and reported in Table 3. However, including these features does not improve performance, but to the contrary, deteriorates it. We compute the Pearson correlations between individual labels and the three fallacies categories predictions on the training set. We do see some correlation in Figure 5 in the Appendix. Thus, we speculate our method of including the textual features to be more disrupting for the model than helpful.

5 Results on Test set

We report our test results from the official leaderboard. The task organizers released the DEV labels with the encouragement to include them in the training of the final model. We thus retrained our models before submitting to the test leaderboard. All models are trained using the oversampling strategy *Few*. On Subtask 2, we train a single multilingual model and use this for predictions across all languages. On Subtask 3, the predictions in the official test phase come from different models: the submitted predictions for fr, ge, it, po and ru come from individual XLM-R models trained with oversampling of the target language. The predictions for the three new languages (gl, es and ka) come from a XLM-R model with oversampling of English. For English, they come from a combination of a RoBERTa model and a XLM-R model (oversampling English) (see Subsection 4.2 for abbreviation and explanations) [4]. The results for the leaderboard are replicated in Table 4. Our best ranking is nr. 1 for Spanish Subtask 3, otherwise we rank between nr. 2-7 with an average ranking of 4.33 on Subtask 2 and of 3.66 on Subtask 3.

After the official end of the test phase, we evaluate the prediction for all languages on Subtask 3 made by an XLM-R model with oversampling *Few* - the counterpart of the submitted model for Subtask 2 - to study the impact of opting for the simplicity of using a single model per subtask for all languages. In Table 4, we observe that this model gets higher F1-micro score on the test set in 5 out of 9 languages. This is confirming the DEV results that oversampling target language is not a robust strategy across languages.

6 Discussion

We have tried to account for label imbalance by oversampling. We see a small positive effect on F1-macro scores in the experiments in Table 1. However, predicting the labels of classes with few training data samples remains challenging: In Table 6 in the appendix, we observe an F1-score of zero for the English DEV set on several classes regardless of whether we use the English or multilingual setup. The low number of training examples for some classes along with the semantic complexity of the task, seems to be challenging for (fine-tuned) RoBERTa models.

In addition, Table 6 shows a low score on class 'Repetition' on the English DEV set which cannot be explained with a low number of training examples - however, the class is also semantically distinct from the remaining classes in the sense that contains recurring wording in the text instead of single expressions of a particular semantic mean-

[4] Oversampling on English in the XLM-R model used for es, gl and ka plus combination for en, was not the intended setting for submitting to the leaderboard since based on the dev results oversampling on 'en' did not yield higher performance

ing, which is likely not well capture by standard text classification in RoBERTa models, either.

Lastly, we note as limitation of the above analysis for the F1-macro in the DEV set, that some classes have very few samples, which makes the results and their discussion less certain.

7 Conclusion

We experiment with fine-tuning different RoBERTa models for multi-label classification to achieve robust results on both subtasks and across languages. We noted that small changes in the learning-rate could lead to unstable results where the models either scored high or zero depending on the seed. We tried to mitigate class imbalance in the training data. However, we notice that some classes have so few training instances in comparison to their semantic complexity, that finetuning RoBERTa models is not a good method fit for these classes.

We explored different oversampling strategies and tried to include textual features from other models in the training. We got inconclusive results for the first and negative results for the second. However, we ended up getting robust results on both subtasks.

Acknowledgements

This work was supported by the Danish Data Science Academy – which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516)

References

Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.

Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

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.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Anjana Gosain and Saanchi Sardana. 2017. Handling class imbalance problem using oversampling techniques: A review. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 79–85.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*.

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*.

Giovanni Da San Martino, Alberto Barrón-Cedeno, 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.

Amalie Pauli, Leon Derczynski, and Ira Assent. 2022. Modelling persuasion through misuse of rhetorical appeals. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 89–100.

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 multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation, SemEval*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. “nice try, kiddo”: Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Appendix

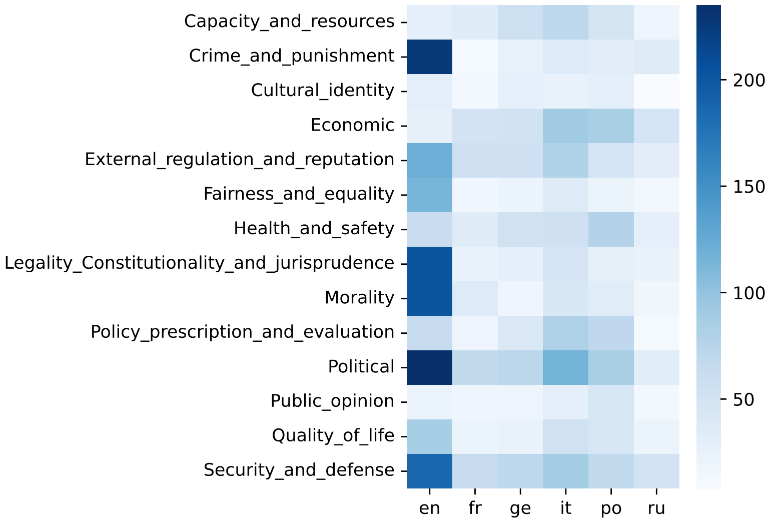

Figure 3: Distribution of labels per language per class for Subtask 2.

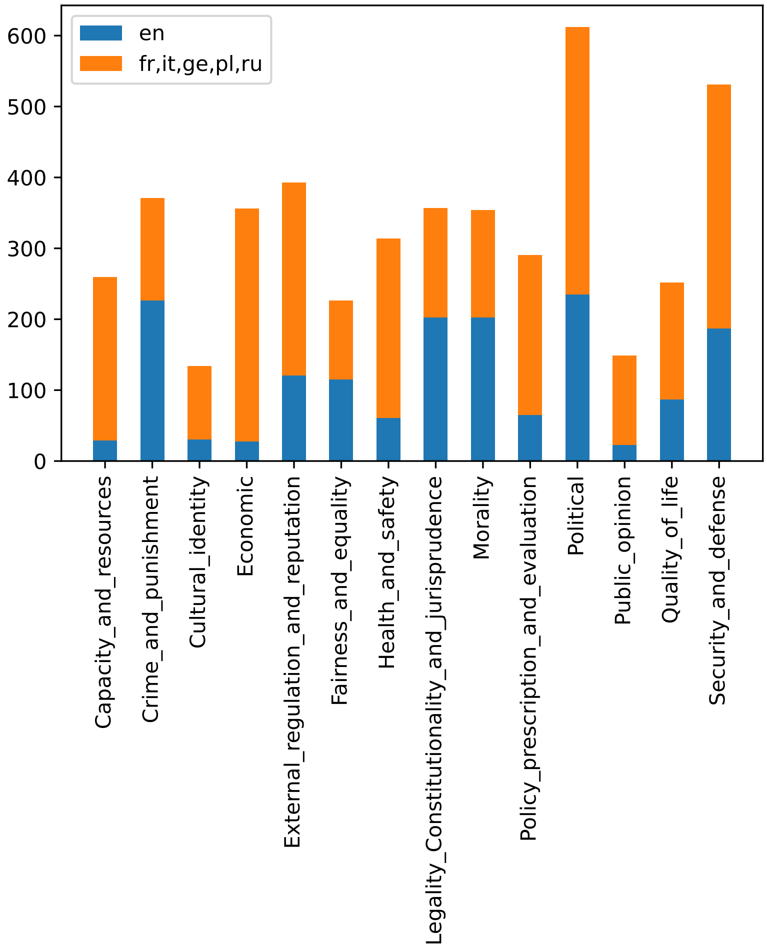

Figure 4: The total number of labels per class for English vs. non-English in Subtask 2.

Lang	Samples	Avg. Labels	No Labels
en	9498	0.6	5738
it	2623	1.6	878
ru	1962	1.3	717
fr	2259	1.9	566
ge	1555	2.0	303
pl	2310	0.9	1078

Table 5: Samples, average number of persuasions, and paragraphs without persuasion for Subtask 3 training data.

Label	F1-score (XLM-R few)	F1-score (RoB few)	Support
Appeal_to_Authority	0.046 (0.006)	0.058 (0.011)	28
Appeal_to_Fear-Prejudice	0.307 (0.016)	0.255 (0.015)	137
Appeal_to_Hypocrisy	0.000 (0.000)	0.000 (0.000)	8
Appeal_to_Popularity	0.081 (0.042)	0.000 (0.000)	34
Causal_Oversimplification	0.170 (0.023)	0.124 (0.053)	24
Conversation_Killer	0.165 (0.033)	0.094 (0.025)	25
Doubt	0.233 (0.028)	0.224 (0.020)	187
Exaggeration-Minimisation	0.209 (0.011)	0.236 (0.008)	115
False_Dilemma-No_Choice	0.276 (0.015)	0.121 (0.022)	63
Flag_Waving	0.448 (0.017)	0.504 (0.018)	96
Guilt_by_Association	0.247 (0.134)	0.527 (0.184)	4
Loaded_Language	0.537 (0.012)	0.549 (0.015)	483
Name_Calling-Labeling	0.509 (0.010)	0.549 (0.014)	250
Obfuscation-Vagueness-Confusion	0.000 (0.000)	0.000 (0.000)	13
Red_Herring	0.027 (0.033)	0.000 (0.000)	19
Repetition	0.047 (0.006)	0.039 (0.011)	141
Slogans	0.300 (0.051)	0.286 (0.051)	28
Straw_Man	0.000 (0.000)	0.000 (0.000)	9
Whataboutism	0.000 (0.000)	0.000 (0.000)	2

Table 6: F1-scores per class for Multilingual backbone (XLM-RoBERTa) and English backbone (RoBERTa). Data collected over 5 runs on English DEV dataset and averaged with standard deviation in parenthesis.

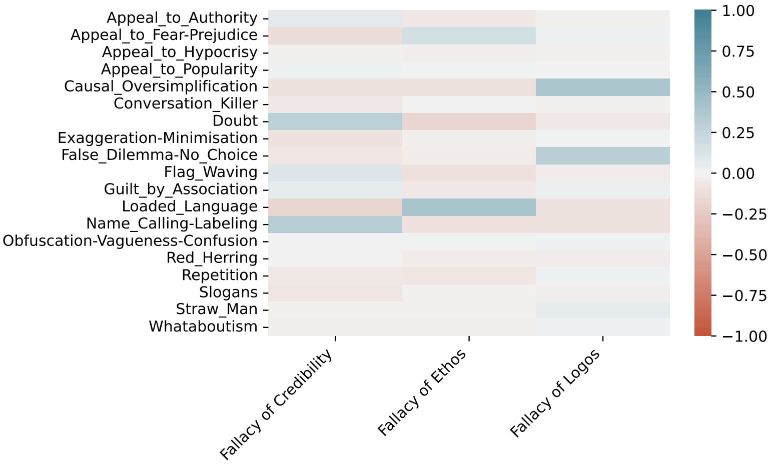

Figure 5: Pearson correlation plot between labels and predicted categories of ethos, pathos and logos on English subtask 3.