NAP at SemEval-2023 Task 3: Is Less Really More? (Back-)Translation as Data Augmentation Strategies for Detecting Persuasion Techniques

Neele Falk*, **Annerose Eichel***, **Prisca Piccirilli***
Institute for Natural Language Processing, University of Stuttgart
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

Persuasion techniques detection in news in a multi-lingual setup is non-trivial and comes with challenges, including little training data. Our system successfully leverages (back-)translation as data augmentation strategies with multi-lingual transformer models for the task of detecting persuasion techniques. The automatic and human evaluation of our augmented data allows us to explore whether (back-)translation aid or hinder performance. Our in-depth analyses indicate that both data augmentation strategies boost performance; however, balancing human-produced and machine-generated data seems to be crucial.

1 Introduction

The SemEval 2023 Task 3 (Piskorski et al., 2023) aims at analyzing online news by detecting *genre*, *framing*, and *persuasion techniques*, i.e., *what* is presented *how* using *which rhetoric means*. Persuasion techniques detection (Subtask 3) aims to identify which rhetoric means are used to influence and persuade the reader. News articles are provided in English, German, French, Italian, Polish, and Russian. To foster the development of language-agnostic systems like our approach, the organizers additionally introduce three surprise languages – Spanish, Greek, and Georgian[1] – with test data only.

We build a system that successfully leverages (back-)translation as data augmentation approaches with multi-lingual transformer models to detect persuasion techniques in all 9 languages. We win the task for *fr*, achieve 2nd place for *ge*, *it* and *po*, and 3rd place for *es*, *el* and *ka*, while ranking midfield for *ru* and *en*. Our main contribution consists in exploring the extent to which data augmentation via (back-)translation boosts performance for the task at hand. Our findings suggest that *more* data does boost performance, especially for underrepresented labels. Our in-depth analyses however show that *less tends to be really more* w.r.t balancing natural vs. augmented data, as *more* (augmented) data can severely hurt performance.

So regardless of whether you, personally, participate, this will color popular sentiment to a **massive** degree. It will grow the cognitive dissonance that assures people of things like **"the government is your friend"** and that **"you don't need to protect yourself, the police will take care of you."**

Labels: Loaded Language, Slogans

Figure 1: Example of a multi-labelled paragraph and the corresponding relevant textual spans for the persuasion techniques *Loaded Language* and *Slogans*.

2 Background

Predicting a set of persuasion techniques given a piece of news text in a monolingual setting has been explored in previous shared tasks (Da San Martino et al., 2019, 2020; Dimitrov et al., 2021). Existing successful systems usually include monolingual transformer-based models and ensemble mechanisms to optimally aggregate predictions (Mapes et al., 2019; Jurkiewicz et al., 2020; Chernyavskiy et al., 2020; Tian et al., 2021). Approaches that additionally focus on provided or external data also show improvement, using techniques such as fine-tuning on additional in-domain data or augmenting the training data (Abujaber et al., 2021).

3 Data Description

3.1 Gold Data

The data consists of news and web articles, covering recent hot topics (such as COVID-19 and climate change) that are multi-lingual (*en, fr, it, ge, ru , po*) and multi-labelled amongst 23 fine-grained persuasion techniques (mapped to 6 coarse-grained categories). The relevant span-level annotations for each labeled paragraph are also provided. Fig. 1 illustrates a multi-labelled instance in *en*, and Table 3 offers an overview of the training data size for this

*All authors contributed equally to this work.

[1]Henceforth, we use the following official identifiers: *en*:English, *fr*:French, *it*:Italian, *ru*:Russian, *ge*:German, *po*:Polish, *es*:Spanish, *el*:Greek, and *ka*:Georgian.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1433–1446

gold dataset. This is a rather small and imbalanced dataset regarding both the language in consideration and the labels (cf. Tab. 7 in App. A.1). To increase our training data as well to provide additional examples of persuasion techniques for the low represented labels, we use data augmentation techniques.

3.2 Data Augmentation

We experiment with two data augmentation techniques by directly making use of the multi-lingual input that is provided. First, we generate **automatic translations** *from* and *to* all possible six languages. Not only does this technique allow us to increase text content, but the transfer of the persuasion techniques along with the corresponding text also increases label representation. We also experiment with **paraphrasing through back-translation**, to augment the data for each language *from* and *to* all possible six languages. As three surprise languages are added in the test set, we also generate (back-)translations to and from these languages, when possible. Table 1 provides an overview of the possible combinations we were able to explore depending on the MT models' availability. For both techniques, we use the translation system MarianMT, based on the MarianNMT framework (Junczys-Dowmunt et al., 2018a) and trained using parallel data collected at OPUS (Tiedemann and Thottingal, 2020). The purpose of data augmentation in this work is two-fold. While it enables us to substantially increase our training data size (Tab. 3), we are also aware that information contained in the original input can be "lost in (back-)translation" (Troiano et al., 2020), which leads to our research question: for the task of detecting persuasion techniques, to which extent can (back-)translation techniques help models' performance? We first conduct automatic and human evaluation on our obtained augmented data.

3.3 Evaluating Augmented Data

Automatic Evaluation Common metrics to evaluate automatic translations include BLEU (Papineni et al., 2002), ROUGE (Lin and Hovy, 2003), METEOR (Banerjee and Lavie, 2005) and require a reference *gold* (human) translation to be computed. We compute the BLEU scores for the paraphrases obtained via back-translations. Results vary greatly depending on the pivot language (e.g., en2*ru*2en vs. en2*es*2en) but on average, the scores are reasonable across languages (Tab. 2). Table 9 in App. A.2 provides the detailed BLEU scores across all combinations. BLEU does not account, however, for fluency nor the persuasion technique preservation. Moreover, we cannot use it to evaluate our automatically obtained translations as we do not have gold references. We therefore conduct a small-scale human annotation study to assess the quality of our (back-)translated data.

		Source Languages (SL)					
		en	fr	it	ru	ge	po
Target Languages (TL)	en	-	✓ ✓	✓ ✓	✓ ✓	✓ ✓	✓ ✓
	fr	✓ ✓	-	✓ ✗	✓ ✓	✓ ✓	✓ ✓
	it	✓ ✓	✗ ✗	-	✗ ✗	✓ ✓	✗ ✗
	ru	✓ ✓	✓ ✓	✗ ✗	-	✗ ✗	✗ ✗
	ge	✓ ✓	✓ ✓	✓ ✓	✗ ✗	-	✓ ✓
	po	✗ ✗	✓ ✓	✗ ✗	✗ ✗	✓ ✓	-
	es	✓ ✓	✓ ✓	✓ ✓	✓ ✓	✓ ✓	✗ ✗
	el	✓ ✗	✓ ✓	✗ ✗	✗ ✗	✓ ✗	✗ ✗
	ka	✗ ✗	✗ ✗	✗ ✗	✗ ✗	✗ ✗	✗ ✗

Table 1: Language pairs covered (✓) and not covered (✗) by marianMT models for translation (in blue) and back-translation (in red). The direction of *translation* is from SL to TL and back to SL for *back-translation.*

en	fr	it	ru	ge	po
48.06	32.80	45.79	20.84	33.34	21.23

Table 2: 4-gram BLEU score average per language.

Human Evaluation We perform human evaluation for *en*, *fr*, *it*, *ru* and *ge*. For **back-translations**, which we consider *paraphrases* of the original input, we ask one ((near-)native) volunteer per language to rate the quality of the generated paraphrases on a scale 1–5 for the following three aspects: *fluency*, *fidelity*, and *surface variability*, where 5 is the best score. For **translations**, as there is no reference input, the same annotators are asked to rate the quality of the generated input regard-

	Training Datasets				
	gold	+BT-sl	+T+BT	+T+BT-sl	+span
en	3,761	22,561	25,968	29,728	7,521
fr	1,694	11,852	17,700	21,086	3,387
it	1,746	6,981	10,248	11,993	3,491
ru	1,246	4,981	9,189	10,434	2,491
ge	1,253	7,513	14,691	15,943	2,505
po	1,233	3,697	6,642	6,642	2,465
total	**10,933**	**57,585**	**84,438**	**95,826**	**21,860**

Table 3: Training data size per language. *gold* is the original task data, to which are added all possible (back-)translations (+T and +BT) - with or without the surprise languages (*sl*) as pivot languages - and the relevant textual *spans*.

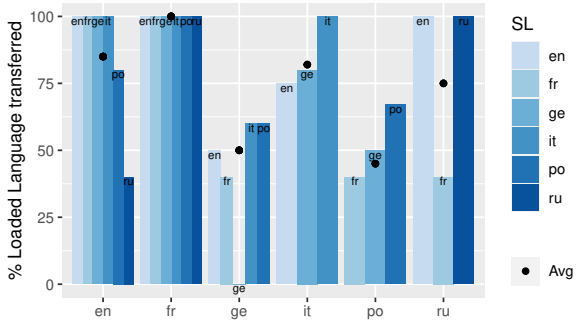

Figure 2: % of *translations* where *Loaded Language* is *transferred* irresp. of (**Avg**) and according to the **SL**, resp.

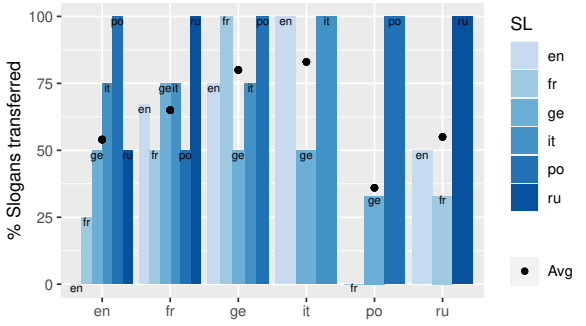

Figure 3: % of *translations* where *Slogans* is *transferred* irresp. of (**Avg**) and according to the **SL**, resp.

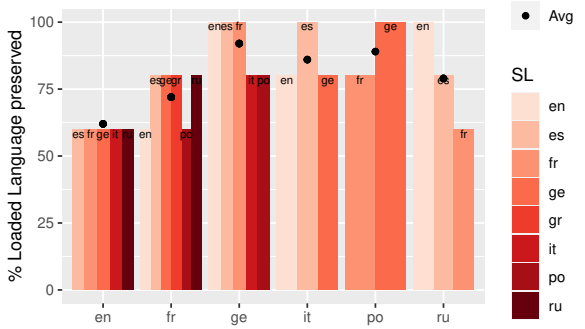

Figure 4: % of *back-translations* where *Loaded Language* is *preserved* irresp. of (**Avg**) and according to the **SL**, resp.

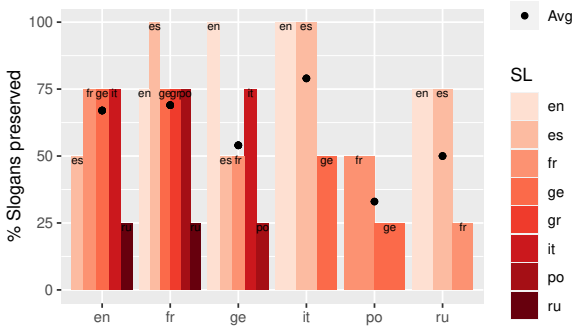

Figure 5: % of *back-translations* where *Slogans* is *preserved* irresp. of (**Avg**) and according to the **SL**, resp.

ing *fluency* (1-5), and whether they think this is a *human-produced* paragraph. Additionally, in both setups, we ask our annotators if they think the original assigned label is *preserved* (for paraphrases) or *fits* in the generated paragraph (for translations). See App. A.3 for a detailed explanation of data selection and annotation setup.

Findings Due to space restrictions, we provide a complete analysis of human judgements in App. A.3 and address in this section the question: to which extent can persuasion techniques be *transferred* or *preserved* in the process of *(back-)translation*, respectively? As shown in Fig. 2, *Loaded Language* easily transfers from *all* languages to *en* and *fr* (avg. 85% and 100%, resp.), but does not when its instances are translated into *ge* (avg. 52%). The opposite is however found in back-translation, as the label is mostly preserved in *ge* but not in *en* and *fr* (Fig. 4). Across target languages, Fig. 3 and 5 show that *Slogans* seems more impacted than *Loaded Language* by the source language from which it gets (back-)translated; however, this is not consistent per language across translation and back-translation. Indeed, a label can be 100% transferable from *po* to *ge* but gets lost the other way around. These findings give evidence that some persuasion techniques might be language- and culture-dependent, as their *transfer* and *preservation* vary depending on the *language pair* in consideration. Overall, the percentage of "lost labels" through (back-)translation, according to our human judgements, remains more or less low depending on the persuasion technique, indicating that they *do not completely* "get lost in (back-)translations". As a result, while we expect the detection of certain persuasion techniques to benefit from our augmented data, it would not be surprising if the detection of others gets hindered by it. We present the results of our regression analysis in Sec. 5.

4 Our System: Approach and Methods

Our approach combines predictions of several models in an ensemble, which differ in three main aspects: a) training data b) model architecture and c) input format to the model. We show in Table 3 the size of each training data we used for the final task submission, and report in Table 8 (App. A.1) all the training data we experimented with. In the following, used model architectures are presented.

	Official Results ($F1$)			Models + resp. training dataset(s) $\in$ Ensemble			
	Micro	Macro	Rank	XLM-R-base	XLM-R-large	Adapters	Additional
en	0.26	0.08	15/23	+BT-sl, +T+BT	+BT-sl, +T+BT-sl, spans	+T+BT-sl	setfit, heur.
fr	**0.47**	**0.33**	**1/20**	+T+BT	+BT-sl, spans	+T+BT-sl	-
it	0.54	0.27	2/20	+T+BT	+BT-sl, spans	+T+BT-sl	-
ru	0.31	0.19	8/19	+T+BT	+BT-sl	+T+BT-sl	-
ge	0.51	0.27	2/20	+T+BT	+BT-sl, spans	+T+BT-sl	-
po	0.42	0.25	2/20	+T+BT	+BT-sl, spans	+T+BT-sl	-
es	0.37	0.18	3/17	+T+BT	+BT-sl, +T+BT-sl, spans	+T+BT-sl	-
el	0.26	0.16	3/16	+T+BT	+BT-sl, +T+BT-sl, spans	+T+BT-sl	-
ka	0.41	0.31	3/16	+T+BT	+BT-sl, +T+BT-sl, spans	+T+BT-sl	

Table 5: Official test results and corresponding leaderboard rankings based on the official metrics $F1$ micro. Note that for each test language we experiment with different possible model combinations in an ensemble and pick a different combination, depending on which ensemble results in the the highest F1-micro score on the validation set. We report $F1$ macro for completeness and show which models and their respective training data were considered in the ensemble for a given language. For example, for *fr* we find the best results with combining predictions of 4 different models: XLM-R-base trained with +T+BT, XLM-R-large, one trained with +BT-sl and one trained with spans and predictions by the adapters trained on +T+BT-sl.

4.1 Model Architectures

XLM-RoBERTa-base/large We fine-tune all parameters of the XLM-RoBERTa (XLM-R) models with a multi-label classification head on top.

Adapter We train a label-specific adapter for each persuasion technique. Adapters (Houlsby et al., 2019) are a specific set of parameters inserted in every layer of a transformer. Instead of fine-tuning the parameters of the full pre-trained language model, these smaller parameters are updated for a specific task while the rest of the parameters is kept frozen. This makes them more efficient to train while still being compatible with the original transformer architecture. We use XLM-R-base as a back-bone model and the binary cross-entropy loss for each label. After training, we combine predictions of the 23 adapters for each paragraph. Note that the adapters are especially useful for low-frequency classes as the binary classification setup usually leads to a higher recall for such labels.

SetFit This few-shot learning method is based on sentence-transformers (Tunstall et al., 2022). As a first step, a pre-trained SBERT model is fine-tuned on a small number of labeled text pairs in a contrastive Siamese manner. This model can then be used to generate embeddings for sentences or paragraphs and to train a simple text classifier for the target task. The main advantages compared to other few-shot fine-tuning approaches are (i) efficiency and (ii) that it does not require prompts or verbalizers. Using a multi-target strategy, we train a distinct logistic regression classifier for each persuasion strategy with paragraph representations as inputs, which are then combined to output a prediction for each label for each paragraph.

4.2 Training data

Our data augmentation techniques (Sec. 3) allow us to train our models on different training sets of different sizes, which we report in Table 3. Besides the original *gold* training data, we obtain six additional training corpora as a result of *translations* (T) and *back-translations* (BT) techniques. Additionally, we experiment with injecting in the *gold* training data the relevant textual *spans* triggering the annotated persuasion techniques, in the hope that it helps the models to particularly focus on relevant information for each persuasion technique.

4.3 Post-processing and ensemble

After training each model we apply **threshold moving**, i.e., given the validation set for each language, we search the optimal classification threshold for each model. We tune the threshold for a range between 0.1 and 0.9 (step size=0.1) and pick the one that maximizes the F1-micro score on the corresponding validation set. We also develop rule-based **language-specific heuristics** for a small number of labels. For instance, for the persuasion technique *Doubt*, we overturn the model predictions iff a paragraph contains, for example, a question mark or question words.

We then compute the F1-scores on the validation set using an **ensemble**. For each instance in the validation set, all models that are part of the ensemble can vote (predict the classes according to the corresponding threshold). For each class, the votes are then summed up. For the final prediction,

	Df	explvar	sign
trainingSet:label	110	38.00	***
label	22	31.31	***
trainingSet	5	18.28	***
testLang:trainingSet	25	0.89	***
testLang	5	0.24	***
total explained variance		**88.73**	

Table 6: Terms of the most explanatory regression model for predicting **F1 (persuasion strategy)**, with degrees of freedom, statistical significance and explained variance. The best fit explains 88.73% of the variance.

a class is added if the sum of votes exceeds the voting threshold. We calculate the optimal voting threshold again based on the validation set F1-micro score.

4.4 Experimental Setup

We train XLM-R-base for 10 epochs on the different training datasets and apply early stopping, picking the model that achieves the highest F1-micro score on the provided development dataset. We do the same with XLM-R-large with a maximum of 5 epochs. The adapters are trained only on the largest training dataset (+T+BT-sl) for a maximum of 5 epochs. SetFit is trained on 1,000 instances sampled from +T+BT-sl. More details on implementation can be found in Table 11 in App. B.1[2]. We also report our non-submitted experiments in App. B.2, as we draw some insightful conclusions.

5 Results and Analysis

Subtask 3 official results and corresponding leaderboard rankings are shown in Tab. 5. Note that for each language we find an ensemble by computing the F1-micro score for different model combinations. The combination for each test language is listed in Tab. 5, including the respective training data. We win the task for *fr*, achieve 2nd places for *po*, *it*, *ge* and 3rd places for the surprise languages *el*, *gr*, *ka*. We rank roughly around mid-field for *ru* and *en*, but manage a 3rd place post-submission for the latter.

To identify trends in the effectiveness of the various data augmentation strategies employed, we train a linear regression model to predict the F1 scores of the different persuasion labels as dependent variables. Our regression model looks at the effect of *training data*, *test language* and *persuasion technique*, each coded as categorical independent

[2]Note that for *ru*, the *span* model was not included in the final ensemble due to a model error.

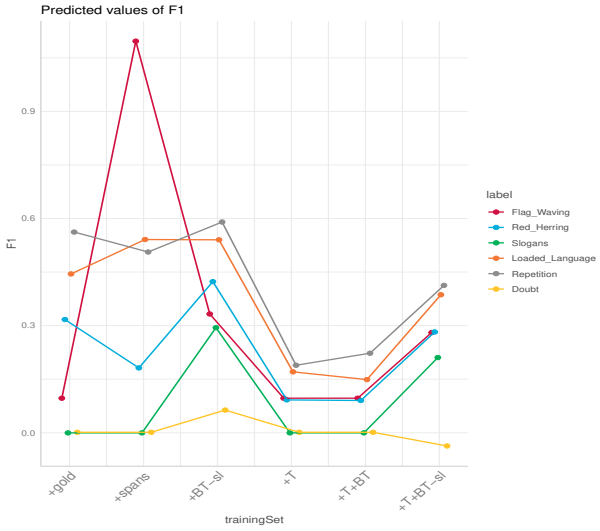

Figure 6: Effect of training data type (+size, in ascending order left to right) on predicted F1 scores for 6 persuasion techniques.

variables (IV)[3]. Regression results are presented in Tab. 6. We first look at which factors play an important role in predicting different persuasion strategies effectively. We look at the amount of explained variance by each IV: the best fit ($Adj.R^2 = 0.86$) contains each IV and the interactions between each training data with (i) test languages and (ii) persuasion techniques. Very little systematic differences across test languages are observed, as there is 1% explained variance by the interaction between training data and test languages. In contrast, most variance is explained by the interaction between training data and labels (38%) indicating that performance in detecting persuasion techniques is considerably impacted by the choice of augmentation strategy.

We now zoom into the extent to which the choice of augmentation strategy differs across persuasion techniques detection. We visualize the interaction term as an effect plot in Fig. 6. We select six labels of interest to show the overall trend, and present the full plots in App. C. We find the different augmented training sets on the x-axis (sorted by ascending size) and their impact on F1 scores' predictions on the y-axis. We observe that in general, data augmentation positively impacts the predicted performance, especially for the less-frequent persuasion techniques, e.g., *Flag Waving*. Training on the original *gold* data with the relevant textual *spans* can have extremely positive effects (*Flag Waving*, *Loaded Language*) but this effect is not

[3]App. C presents all details of the modeling set up.

observed across *all* labels, indicating that some persuasion techniques' relevant textual information is particularly compressed and the context is therefore less crucial. Results clearly point to *back-translation* (+BT-sl) as the most robust augmentation strategy with consistent improvement of F1 scores across all labels.

In contrast, *translations* consistently hurt the performance across all labels and is only effective if combined with *back-translations* (e.g., +T+BT-sl). This considerable difference in performance between injecting translations vs. back-translations, which are inherently the same processes, is surprising and not *on par* with our human evaluation findings (Sec 3.3); we plan to conduct further analyses to investigate the phenomenon. Overall, these analyses have shown that adding *some more* data, i.e., +BT only, does indeed improve performance but *too much* augmented data, i.e., +T, +T-BT tends to hinder it.

6 Conclusion

We tackled the task of detecting persuasion techniques in online news in a multi-lingual setup. We built a system that successfully combines natural with augmented data via (back-)translation with an ensemble of SOTA multi-lingual models. While we showed that using augmented data, i.e., *more* data, generally boosts performance, our results also indicate that it might be hurt when integrating *too much* augmented data. In conclusion, for the task of persuasion techniques detection, *more* data, obtained via (back-)translation, does help overall, but *less might be more* when it comes to adding automatically-generated translations.

Ethics Statement

In the context of our evaluation task, we collected ratings from human participants. For this, the participants were provided an Informed Consent Letter with the name and the contact of the investigators; the title, purpose and procedure of the study; risks and benefits for participating in the study; confirmation of confidential anonymous data handling; and confirmation that participation in the study is voluntary. The Informed Consent Letter was signed before the participants took part in the study.

Acknowledgements

We thank the volunteering annotators taking part in the annotation study. We are grateful for helpful feedback from Gabriella Lapesa and the anonymous reviewers.

Neele Falk is supported by Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB (Powering up e-deliberation: towards AI-supported moderation). Annerose Eichel receives funding by the Hanns-Seidel-Stiftung. Prisca Piccirilli is supported by the Studienstiftung des deutschen Volkes and the DFG Research Grant SCHU 2580/4-1 *Multimodal Dimensions and Computational Applications of Abstractness*.

References

Dia Abujaber, Ahmed Qarqaz, and Malak A. Abdullah. 2021. LeCun at SemEval-2021 task 6: Detecting persuasion techniques in text using ensembled pretrained transformers and data augmentation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1068–1074, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1462–1468, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China. Association for Computational Linguistics.

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.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018a. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018b. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Norman Mapes, Anna White, Radhika Medury, and Sumeet Dua. 2019. Divisive language and propaganda detection using multi-head attention transformers with deep learning BERT-based language models for binary classification. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 103–106, Hong Kong, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

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.

Junfeng Tian, Min Gui, Chenliang Li, Ming Yan, and Wenming Xiao. 2021. MinD at SemEval-2021 task 6: Propaganda detection using transfer learning and multimodal fusion. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1082–1087, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation*, Lisbon, Portugal.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. 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.

A Appendix A: Data

A.1 Gold Label Distribution and Training Data Sizes

Tab. 7 lists the number of instances per label for each language for the *gold* training corpus as provided by the task organizers. The overview shows that there is quite some imbalance, e.g. for each *en* we observe a maximum label frequency of >1,800 instances for the label *Loaded Language*, while there is no data at all for the four labels *Appeal to Time, Appeal to Values, Consequential Oversimplification, Questioning the Reputation*. This picture changes depending on the language.

Tab. 8 presents an overview of all training data and their sizes we experiment with. *gold* is the original data provided by the organizers, which we augment with our automatically obtained (back-)translations and the relevant persuasion technique textual *spans* (provided for the task). We obtain (back-)translations from and to all possible six languages (*en, it, fr, ge, ru, po*). When the organizers release the three surprise languages (*es, el, ka*), we are able to obtain translations in *es* and *el*. We distinguish the back-translated augmented data containing - or not - the back-translations in the original six languages *from* the surprise languages. However, note that when combining *translations* and *back-translations* (+T+BT(-sl)), we do not include the *surprise language translations* (size 0). Overall, our data augmentation techniques allow us to considerably increase our training data by almost 900%.

A.2 Augmented Data: Automatic Evaluation

Tab. 9 presents BLEU scores for paraphrases that were obtained via back-translation for 1,2,3,4-grams. Scores are presented per language and language direction. Scores are rather low for back-translations in *fr* via *po* (11.19) and in *ru* via *fr* (15.62) but are overall reasonable across language pairs, reaching around 50 in *it* and *fr* and up to 60 for back-translation in *en* via *es*, giving us the intuition that the quality of these paraphrases are rather good. We conduct human evaluation to confirm this hypothesis.

A.3 Augmented Data: Human Evaluation

Data Selection For **back-translations**, we randomly extract 10 original input instances in each language, and their back-translations *from* all possible pivot languages, e.g., 10 original *en* instances where each instance was back-translated from *it, fr, ge, es* and *ru*: 10*5 = 50 *en* paraphrases to judge. For **translations**, we randomly extract 10 original input instances in each language and their translations *in* all possible target languages, e.g., 10 original *ge* instances which were translated in *en, fr* and *it*: these 10 translations to be judged in each target language (here *en, fr, it*) originate from the *same* source language. Additionally, in this *translations set*, we add six *control* original instances in each language.

Annotators We initially ask five volunteers - one for each language (*en, fr, it, ru, ge*) to partake in the study. One additional annotator (one of this paper's authors) finished annotations for *ge*. All six of them are based in Germany, are native or near-native speaker of the respective language. Each annotator submits two unique sets of answers for (i) translations and (ii) back-translations.

Setup The annotations are carried out in a remote setting using Google Forms. Annotators are provided detailed written guidelines including examples, first to complete back-translation judgements and then translation judgements (PART 1 & 2, resp. in Tab. 10.) In case of questions, annotators have the option to contact the authors of the paper. The evaluation can be completed flexibly in the course of two days. Annotators can take as much as time as they need for completing the evaluation. The collected data does not include any information that names or uniquely identifies individual people or offensive content. Letters of Consent are signed before participation and stored separately from the collected ratings.

Analyses Fig. 7 and 9 present the fluency scores attributed to translations and back-translations, respectively. For each target language, we report the average scores irrespective of the source/pivot language (**Avg**) and the average scores depending on which language it was (back-)translated from. Overall, fluency scores are rather high (avg. 4), which means that (back-)translation does not affect to a large extent the readability of the generated output. However, the percentage of instances that are judged "human-produced" (Fig. 8) drops with regards to *translations* in *fr, ge* and *ru* (around 30%). When zooming into language pairs, this percentages drops under 25% for *en2fr*, *it2fr*, *it2ge* and *en,fr2ru*. We also collected ratings regarding the *fidelity* and the *surface variability* aspects for the

Label	en	fr	ge	it	po	ru
Justification:						
Appeal to Authority	154	76	225	70	41	10
Appeal to Popularity	15	82	63	37	30	8
Appeal to Values	0	100	73	131	101	48
Appeal to Fear-Prejudice	310	210	182	285	108	54
Flag Waving	287	37	65	35	68	42
Simplification:						
Causal Oversimplification	213	125	33	50	12	39
False Dilemma-No Choice	122	73	41	61	12	28
Consequential Oversimplification	0	112	35	29	24	70
Distraction:						
Straw Man	15	135	15	51	15	21
Whataboutism	16	62	13	8	8	7
Red Herring	44	55	30	23	12	2
Call:						
Appeal to Time	0	41	11	27	14	28
Slogans	153	149	87	54	36	72
Conversation Killer	91	170	121	178	50	88
Manipulative Wording:						
Loaded Language	1,809	944	242	903	310	641
Repetition	544	92	8	22	13	69
Exaggeration-Minimisation	466	258	157	143	111	131
Obfuscation-Vagueness-Confusion	18	113	62	21	36	19
Attack to Reputation:						
Appeal to Hypocrisy	40	134	136	82	162	103
Doubt	518	327	288	882	295	509
Name Calling-Labeling	979	428	734	566	475	253
Guilt by Association	59	130	122	53	94	24
Questioning the Reputation	0	348	310	383	164	303

Table 7: Label distributions of *gold* training data (in absolute numbers), divided by coarse-grained categories.

	Training Datasets						
	gold	+T	+BT	+BT-sl	+T+BT	+T+BT-sl	+span
en	3,761	10,928	18,801	22,561	25,968	29,728	7,521
fr	1,694	10,928	8,466	11,852	17,700	21,086	3,387
it	1,746	6,758	5,236	6,981	10,248	11,993	3,491
ru	1,246	6,699	3,736	4,981	9,189	10,434	2,491
ge	1,253	9,683	6,261	7,513	14,691	15,943	2,505
po	1,233	4,178	3,697	3,697	6,642	6,642	2,465
es	0	9,696	0	0	0	0	0
el	0	6,706	0	0	0	0	0
ka	0	0	0	0	0	0	0
total	**10,933**	**65,576**	**46,197**	**57,585**	**84,438**	**95,826**	**21,860**

Table 8: Training data size per language. *gold* is the original task data, to which are added all possible (back-)translations (+T and +BT) - with or without the surprise languages (*sl*) as pivot languages - and the relevant textual *span*s.

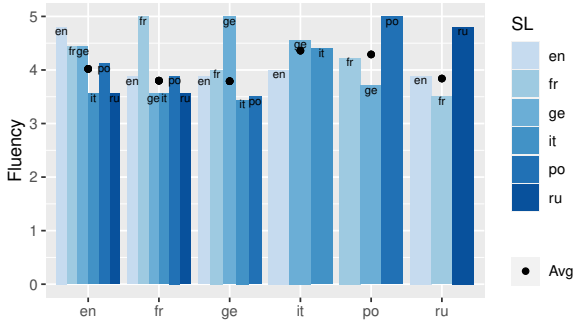

Figure 7: Average fluency scores of each TL (*translations*) irrespective of (**Avg**) and according to the source language (**SL**).

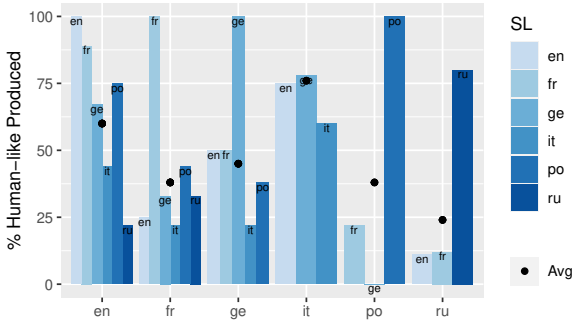

Figure 8: Percentage of TL *translations* judged "human-produced" irrespective of (**Avg**) and according to the **SL**, respectively.

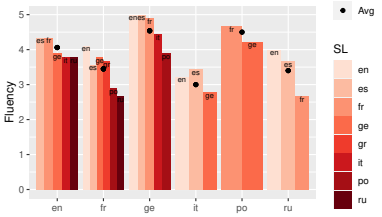

Figure 9: Average fluency scores of each TL (*back-translations*) irrespective of (**Avg**) and according to the (pivot) (**SL**).

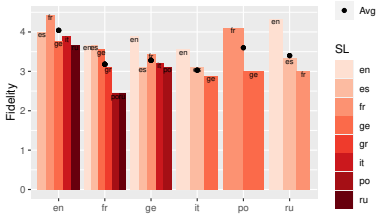

Figure 10: Average fidelity scores of each TL (*back-translations*) irrespective of (**Avg**) and according to the (pivot) (**SL**).

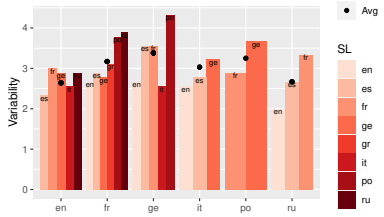

Figure 11: Average surface variability scores of each TL (*back-translations*) irrespective of (**Avg**) and according to the (pivot) (**SL**).

lang pair	1-gram	2-gram	3-gram	4-gram
en2ru2en	54.83	43.25	35.31	28.76
en2es2en	80.37	72.44	66.30	60.63
en2it2en	73.45	64.84	58.27	52.31
en2fr2en	68.01	59.47	53.03	47.23
en2ge2en	73.13	64.42	57.61	51.39
fr2ge2fr	60.33	50.37	43.26	37.04
fr2es2fr	67.23	59.52	53.63	48.16
fr2po2fr	20.98	16.47	13.57	11.19
fr2en2fr	69.32	61.22	55.04	49.33
fr2el2fr	48.15	40.03	34.26	29.21
fr2ru2fr	46.08	34.82	27.67	21.89
it2ge2it	55.55	45.12	37.96	31.86
it2es2it	71.05	63.80	58.23	52.99
it2en2it	73.52	64.90	58.44	52.52
ru2es2ru	48.52	35.68	27.68	21.39
ru2fr2ru	41.69	28.80	21.25	15.62
ru2en2ru	52.04	39.99	32.03	25.52
po2fr2po	40.33	30.60	24.29	19.23
po2ge2po	48.24	36.70	29.25	23.22
ge2it2ge	56.47	44.25	36.33	29.86
ge2po2ge	53.43	40.49	32.35	25.89
ge2es2ge	61.21	48.51	40.06	33.13
ge2fr2ge	57.65	45.59	37.57	31.04
ge2en2ge	72.92	61.86	53.80	46.75

Table 9: 1,2,3,4-grams BLEU scores for the paraphrases obtained via back-translation, per language and per language direction.

paraphrases obtained via back-translations. On average, and across target languages and language pairs, fidelity scores are found between 3 and 4, which confirms that information gets somewhat lost in back-translation. Regarding surface variability, scores do not go over 3.5: paraphrases do not diverge too much from their respective original inputs but are also not complete copy pastes, which is the desired outcome when paraphrasing. We reported in the main text findings regarding the *transfer* and the *preservation* of persuasion techniques in *translation* and *back-translation*, respectively (Fig. 2, 3, 4 and 5 in section 3.3). We stated that it depends in both cases on the *language pair direction* and the *label* in consideration, but that observations were not always similar for the same *language pair* and *label* depending on translation or back-translation, making the point that certain persuasion techniques may be language- and culture-dependent. From a machine-translation point of view, this finding need to be accounted for when dealing with persuasive text data. However, for the translation setup, we had also included six *control* original instances, i.e., in the original language, therefore not translations, to assure that potential label "loss" between all other SL to TL was not an artifact of the human judgements. It appears that identifying persuasion

techniques, even when provided with a definition and examples, is a difficult task, even for humans: even though these techniques were present in the *gold data*, our annotators judged that *Loaded Language* was "lost" in over 70% in *ge* of the cases and that *Slogans* was 100% "lost" in *en* and half the time in *fr* and *ge*. This raises the limitations of our small-scale annotation study to evaluate the quality of our automatically-obtained (back-)translations. We pave the way for an interesting project regarding the transfer and the preservation of persuasion techniques through (back-)translation, and a larger-scale human evaluation study could be conducted to confirm our findings.

B Appendix B

B.1 Model Architectures and Hyper-Parameters

Hyper-Parameter Tab. 11 show the hyper-parameter setup for each model architecture. *#Shots* is the number of instances used to train the SBERT model in a contrastive manner for `SetFit`. For each instance, 5 triples (original instance, positive, negative) were created.

Libraries We use XLM-RoBERTa-base/large implementations by `huggingface` (Wolf et al., 2020), the setfit implementation by Tunstall et al. (2022), and adapaters provided by Pfeiffer et al. (2020). `scikit-learn` (Pedregosa et al., 2011) is leveraged for SVM and Random Forest implementations as well as pre-processing and metrics. We use the MarianNMT implementation by Junczys-Dowmunt et al. (2018b) for data augmentation.

B.2 Negative Insights: What did not work

Considering XLM-RoBERTa-base trained on *gold* data as a threshold to be passed, we discard the following models and architectures.

Models SetFit and adapters trained on *gold* only did seriously under-perform our threshold. This observation was one trigger for using data augmentation techniques. Not surprisingly, classic approaches to multi-label classification problem such as SVMs or Random Forest Classifiers (Pedregosa et al., 2011) did not outperform transformer-based approaches even in the case where more training data was added.

Training Data Strategically playing around with which relevant training data would lead to better performance was a focus on this work. We showed in Sec. 5 that training with both *gold* and the *spans* increased performance on certain labels. We also tried to train a XLM-RoBERTa-base model on *spans* only, i.e, discarding the rest of the paragraph's context. This however seriously harmed performance, indicating the importance of larger textual context to detect persuasion techniques.

To account for label imbalance in the *gold* training data (Tab. 7), we experiment with injecting translations and/or back-translations only for the paragraphs whose labels are under-represented (< 100) for the respective language, e.g., *Appeal to Time* in *en*, *Repetition* in *ge*. Performance drops for most languages, but this is in retrospective not a surprising finding. Not only are persuasion techniques imbalanced *intra-language*, but also *inter-language*: injecting (back-)translations, and therefore transferring persuasion techniques from SL to TL, wrongly fills the gaps of label imbalance in the TL, mistakenly introducing labels that do not *fit* in that TL. Finally, we also attempt training on languages grouped by their language families (*en-ge*, *fr-it*, *ru-po*), with (i) only the *gold* data and (ii) injecting *translations*. The results vary between languages, but we note improvements on performance, indicating a potential promising direction to take in further experiments.

C Appendix C: Quantitative Analysis

Modeling Setup We train a XLM-RoBERTa-large model for 5 epochs on six different training sets, including the *gold* dataset. We measure the F1 score for each of the 23 persuasion strategies on the development set for each language. This results in a data frame of size ($n = 6 \times 23 \times 6 = 828$).

We add a categorical variable as independent variable (IV) step-by-step, starting with the label. We compare whether the more complex model improves the fit significantly. We then add two-way interactions.

Analyses We presented in Sec. 5 the effect of training data on prediction scores for six persuasion techniques (Fig. 6). We report this effect on *all* persuasion techniques, from Fig. 12–17. Similarly to our findings on six techniques, results clearly point to back-translation (+BT-sl) as the most robust augmentation strategy with consistent improvement of F1 scores across all labels.

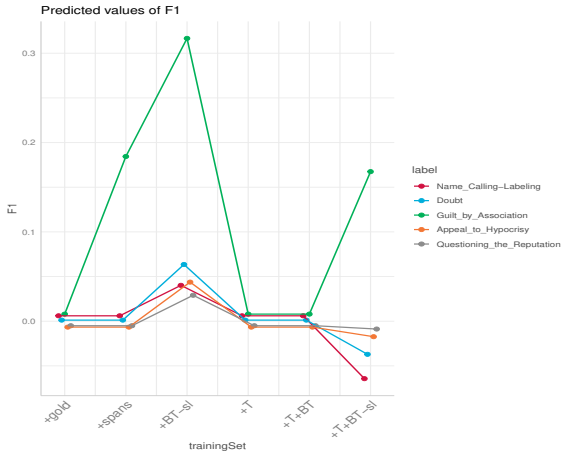

Figure 12: Effect of training data type (+size, in ascending order left to right) on predicted F1 scores for persuasion techniques falling under *Attack of Reputation*.

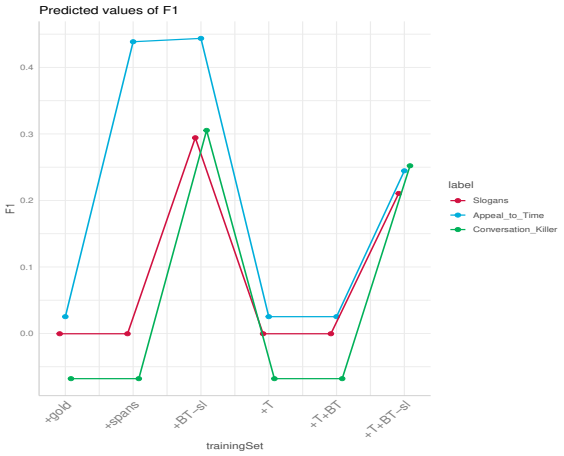

Figure 13: Effect of training data type (+size, in ascending order left to right) on predicted F1 scores for persuasion techniques falling under *Call*.

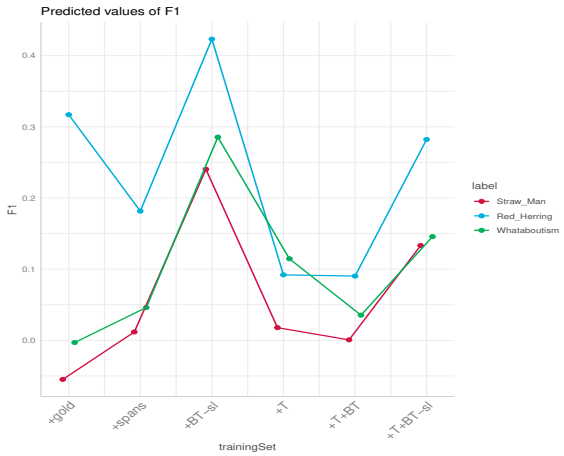

Figure 14: Effect of training data type (+size, in ascending order left to right) on predicted F1 scores for persuasion techniques falling under *Distraction*.

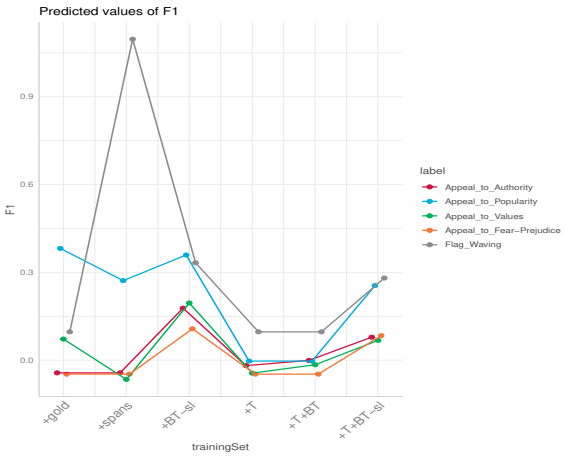

Figure 15: Effect of training data type (+size, in ascending order left to right) on predicted F1 scores for persuasion techniques falling under *Justification*.

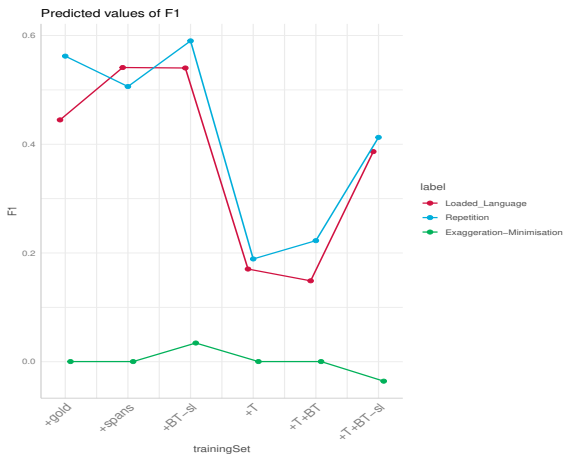

Figure 16: Effect of training data type (+size, in ascending order left to right) on predicted F1 scores for persuasion techniques falling under *Manipulative Wording*.

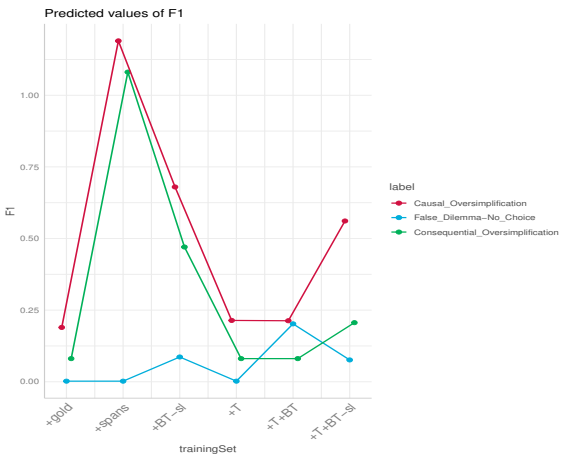

Figure 17: Effect of training data type (+size, in ascending order left to right) on predicted F1 scores for persuasion techniques falling under *Simplification*.

PART 1: You will evaluate the **quality of generated paraphrases.**

You will be given the original sentence and several paraphrases for that original sentence. For each paraphrase:

1) On a scale from 1 to 5, rate the **fluency** of that paraphrase.
Irresp. of the original sentence, how readable is the paraphrase?
1 means the sentence is not readable/plausible at all, and 5 is fully fluent.

2) On a scale from 1 to 5, rate the **fidelity** of that paraphrase.
Compared to the original sentence, how much information is preserved? -
How semantically consistent is the paraphrase?
1 is when the information is fully lost, and 5 is fully semantically consistent.

3) On a scale from 1 to 5, rate the **surface variability** of that paraphrase.
How much difference does the paraphrase have in the form of expression compared to the original sentences?
1 is a word-per-word paraphrase, and 5 is a fully new constructed sentence.

4) A **persuasion technique** is assigned to the original sentence:
Does it apply to the paraphrase as well? Yes or No.

We show an example below.
You will encounter only two different persuasion techniques: **Loaded_Language** and **Slogans**.
We provide below a short definition and a couple of examples in English for you to get an overall idea of the techniques.

Reference Text	Paraphrases	Label	Fluency	Fidelity	Surface Variability	Label preserved?
How can I find the girl for me?	How can you find your cat?	Doubt	5 (fully fluent)	1 (not readable)	2	Yes
	How can I find the girl for me?	Doubt	5 (fully fluent)	5 (fully fluent)	1 (same sentence)	Yes
	Is there any way to find my right girl?	Doubt	4	5 (fully fluent)	5 (new sentence)	Yes

Loaded_Language: Using specific words and phrases with strong emotional implications (either positive or negative) to influence and convince the audience that an argument is valid/true.
This fallacy is also known as euphemisms, appeal to/argument from emotive language, or loaded language.

Examples:
- "How ***stupid*** and ***petty*** things have become in Washington"
- "They keep feeding these people with ***trash***. They should stop."

Slogans: A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

Examples:
- "Our ***'unity in diversity'*** contrasts with the divisions everywhere else."
- ***"Make America great again!"***
- ***"Immigrants welcome, racist not!", "No border. No control!"***

PART 2: You will evaluate the **quality of generated paragraphs**.

You will be given sentences, for each of them:

1) On a scale from 1 to 5, rate the **fluency** of that paraphrase:
How readable is the paraphrase?
1 means the sentence is not readable/plausible at all, and 5 is fully fluent.

2) Do you think this sentence was **human-produced** (vs. automatically generated)? Yes or No.
3) A **persuasion technique** is assigned to the sentence:
Do you think it fits? Yes or No.
We show an example below.
If needed, we provide the info on the persuasion techniques we consider once again (above in this paper).

Paragraph	Label	Fluency	Human-produced	Label fits?
How can I find the girl for me?	Doubt	5 (fully fluent)	Yes	Yes
I don't. Know but, why?	Slogans	1 (not readable)	No	No

Table 10: Annotation guidelines for the human annotation study.

	XLM-R-base	XLM-R-large	Adapters	SetFit
learning rate	1e-5	1e-5	1e-4	-
max epochs	10	5	5	2
num of text pairs	-	-	-	5
#shots	-	-	-	1000
seed	42	42	42	42
batch size	16	16	16	32
loss	Binary Cross Entropy	Binary Cross Entropy	Binary Cross Entropy	Cosine Similarity
metric for best model	F1 macro	F1 macro	F1 macro	F1 macro

Table 11: Overview of hyper-parameter for each model architecture used for the submission on the final test set.