whatdoyoumeme at SemEval-2024 Task 4: Hierarchical-Label-Aware Persuasion Detection using Translated Texts

Nishan Chaterjee [1,2,4]
[1]University of La Rochelle
La Rochelle, France

Marko Pranjić[2,4] and **Boshko Koloski**[2,4]
[2]Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Lidia Pivovarova[3]
[3] University of Helsinki
Helsinki, Finland

Senja Pollak[4]
[4]Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

In this paper, we detail the methodology of team *whatdoyoumeme* for the SemEval 2024 Task on Multilingual Persuasion Detection in Memes. We integrate hierarchical label information to refine detection capabilities, and employ a cross-lingual approach, utilizing translation to adapt the model to Macedonian, Arabic, and Bulgarian. Our methodology encompasses both the analysis of meme content and extending labels to include hierarchical structure. The effectiveness of the approach is demonstrated through improved model performance in multilingual contexts, highlighting the utility of translation-based methods and hierarchy-aware learning, over traditional baselines.

1 Introduction

Persuasion techniques in politics have a significant impact on democratic processes, which was particularly evident in contexts such as the 2020 US elections, where cognitive dissonance and media messages played a crucial role in influencing voter behaviour and attitudes (Perloff, 2013; Center, 2023). These techniques, which utilise psychological insights, align people's attitudes with their actions and thus influence political affiliations and opinions. The interplay of crises – pandemic, economic downturn, protests against racial justice, and debates over electoral legitimacy – has further highlighted the impact of persuasive narratives on public perception and democratic resilience (Jamieson et al., 2023). This complicated relationship underscores the crucial role of persuasion in political discourse and its potential to shape democratic outcomes at crucial historical moments in society.

Manually recognizing persuasion in textual content is increasingly challenging due to the vast amount of information generated daily and the nuanced nature of persuasion techniques. Efforts in this area have expanded to include the development of collaborative tasks(Da San Martino et al., 2019a) aimed at recognizing persuasion across languages and levels of hierarchy, reflecting the global and complex nature of persuasive communication in digital spaces.

In the past, researchers have used statistical text analysis methods that focused on lexical and syntactic features to identify patterns and markers of persuasive language (Jacobs, 1992). While these approaches provided basic insights, they were often not deep enough to fully capture the subtleties of human language and persuasion. The detection of persuasion in texts has shifted from statistical text analysis to the use of Large Language Models (LLMs), such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020). These models use deep learning to understand the context, semantics and complex interplay of language elements, providing more effective means of recognizing persuasive tactics in text.

The collective effort in data collection and the joint tasks have contributed significantly to belief detection, with initiatives such as the SemEval joint tasks fostering community-wide collaboration. The NLP4IF-2019 shared task(Da San Martino et al., 2019a) was another example of the collective effort to refine detection methods through standardised tasks. The task was divided into two parts: the identification of propagandistic text fragments and their specific techniques at fragment level and a binary classification at sentence level to recognise sentences containing propaganda. The joint task attracted a large participation and showed that most of the systems were able to significantly outperform the established baselines. Alhindi et al. (2019) found that for some propaganda techniques, it is not enough to look at just one sentence to make an accurate prediction (e.g. repetition) and therefore the whole article needs to be included as context. Da San Martino et al. (2019b) presented a novel method for detecting propaganda at the level of fragments in news articles that goes be-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1537–1543

yond traditional document-level detection. Their method addressed the need for more nuanced and explainable analysis by manually annotating news articles with specific propaganda techniques and developing a multi-granularity neural network model that outperformed BERT-based baselines. Koreeda et al. (2023) showed that cross-lingual and multi-task training combined with an external balanced dataset can improve genre recognition and framing, on a recently proposed task by Piskorski et al. (2023).

Recent studies have shown that translating texts from low-resource languages to a high-resource language, such as English, improves performance of end-to-end approaches on tasks such as classification (Ghafoor et al., 2021; Jauregi Unanue et al., 2023) and document similarity (Zosa et al., 2022). Koloski et al. (2023) show that cross-lingual validation can lead to improvement on classification performance for some tasks. Some earlier works also confirm this to be true for non transformer architectures (Moh and Zhang, 2012). The rest of this article is organised as follows: Section 2 describes the task, while Section 3 presents the proposed method and the results. Finally, the conclusion and proposed future work in Section 4.

2 Task description

SemEval-2023 Task 4 (Dimitrov et al., 2024) focuses on multilingual detection and classification of persuation techniques in memes.

It is composed of three subtasks. **Subtask 1** was a multi-label text classification task. The text was extracted from the image data that contained the original meme. Although the text contains less information than original image, the annotation procedure accounted for this and allows for differences between labels in the text-only data and image data that provides additional context. **Subtask 2a** was a multimodal extension of the *Subtask 1* by providing both a text and the image data. It is also a multi-label classification task, with the labels annotated based on both text and image data. **Subtask 2b** was also a multimodal task with the same inputs as *Subtask 2a*, but the task was a simpler binary classification task to detect if any persuasion technique is used.

Although only English dataset was provided for training, the evaluation was additionally done on three surprise test datasets in Bulgarian, Macedonian, and Arabic.

Dataset split	English	Bulgarian	Macedonian	Arabic
Train	7000	0	0	0
Validation	500	0	0	0
Development	1000	0	0	0
Test	1500	436	259	100

Table 1: Number of examples for each of the dataset split across languages. Notably, only English data contains data for train/val/dev split while other languages require a zero-shot approach.

We further focus only on the *Subtask 1* and its data as this was the only task we participated in.

2.1 Dataset

The input data for *Subtask 1* is the text extracted from the meme. The training, the development and the test sets were distributed as JSON files. Each of the files encoded a list of examples where each one contained text of a meme and a list of labels, together with additional metadata not used in the model (unique id of the example and URL).

Table 1 shows dataset sizes with numbers of examples in each of the dataset split for each of the languages present in the task.

Labels provided with the dataset were organized in a hierarchy that was not visible from the dataset files and the full overview of the relationships between labels was provided in the accompanying subtask description. Although 20 classes were present in the training data, their ancestors in the hierarchy provided 8 additional classes for a total of 28 classes. Submission files could provide any of the 28 classes as prediction. Predicting an ancestor class of the ground truth labels instead of the leaf-node ground truth label was counted as a partial match.

Figure 1 shows the distribution of label counts in the data. Most common labels like *Smears*, *Loaded Language*, and *Name Calling/Labeling* are almost two times more frequent than any other label. The least frequent labels like *Reductio ad hitlerum*, *Straw man*, *Red herring*, and *Obfuscation* contain less than 100 examples in train, development and validation sets combined.

2.2 Evaluation

The labels are organized in a hierarchy that can be represented as a directed acyclic graph (DAG)- a tree-like structure. Datasets presented in the shared task contained data annotated with leaf labels, but the prediction can take any of the DAG nodes: either leaf or parent. For this reason, a hierarchical

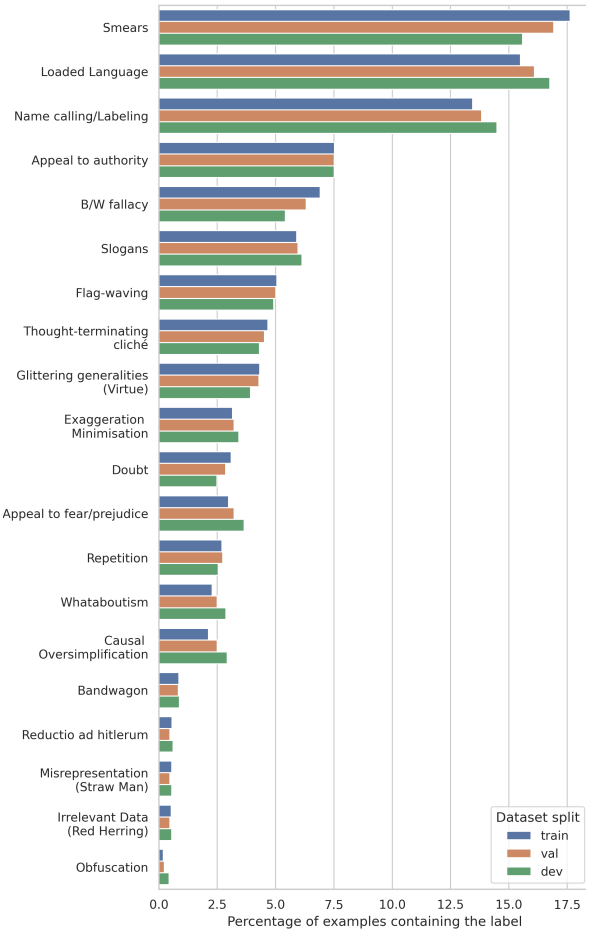

Figure 1: Label distribution shows a noticeable imbalance between class frequencies.

F_1 score (hF_1) (Kiritchenko et al., 2006) was used to take into account partially correct results, and leverage both the distance and the depth between true and predicted labels in the label hierarchy. The difference from standard, or flat, F_1 score is that the standard version considers each example as being a member of its assigned class. In contrast, the hierarchical version considers an example as a member of all parent classes in addition to its assigned leaf class. Formally, hierarchical (micro-average) version of precision (hP) and recall (hR) can be defined as:

$$hP = \frac{\sum_i |Y_A \cap \hat{Y}_A|}{\sum_i |\hat{Y}_A|} \quad hR = \frac{\sum_i |Y_A \cap \hat{Y}_A|}{\sum_i |Y_A|}$$

Where Y_A and $\hat{Y}_A$ represent a set of ground-truth and predicted labels, respectively, extended to contain all ancestors of included leaf nodes. Finally, we can define hierarchical F_1 score (hF_1) as:

$$hF_1 = 2 \cdot \frac{hP \cdot hR}{hP + hR}$$

This formulation effectively views the hierarchical classification as a multi-label setup by implicitly including hierarchy ancestor labels as additional labels.

3 Methodology and Results

Pre-trained language models, such as BERT (Devlin et al., 2018) and its variants, have shown remarkable performance across many NLP tasks. We evaluate several BERT-like models and their performance on the task. We are particularly interested in the impact of the hierarchy on model performances. We describe three approaches to understand the role of hierarchy information in the task.

First, we establish a baseline using BERT and mBART (Liu et al., 2020; Tang et al., 2020) models without using any hierarchy information and grid search to tune the hyperparameters. We explore different tokenization strategies and evaluate the model with micro-F_1.

Second, we evaluate approaches based on modifying the set of ground truth labels by extending the labels with ancestors to include hierarchy information and its influence on model performance.

Finally, we translate English train and validation data to Macedonian, Arabic, and Bulgarian with the NLLB-200 model (NLLB Team et al., 2022) and fine-tune our models for the multi-label classification task. We also compare the performance of the performance on translated data to zero-shot cross-lingual approaches using multilingual models and to translation of test sets.

3.1 Baseline Approach

We utilize distilBERT (Sanh et al., 2019) and mBERT (Devlin et al., 2018) for the baseline approach. The models are trained and evaluated without using any hierarchical information. Text data from the task was provided with escaped newlines (i.e. a newline character was represented with two characters '\' and 'n'). We evaluated a few approaches how to preprocess these data: directly tokenizing the provided text without any preprocessing (NoP), replacing the newlines with a space character (NL-Spc), or using a single newline ('\n') character (NL-L).

Both models are initialized with a linear classifier for multi-label classification. We use the AdamW optimizer (Loshchilov and Hutter, 2017) with binary cross-entropy loss and micro-F_1 score as evaluation metrics. We perform a grid search

over the learning rates (lr) [1e-4, 5e-5, 3e-5, 2e-5] and max sentence lengths of [128, 256] using a batch size of 16 over 6 epochs since BERT-based fine-tuning typically leads to decreased micro-F_1 and hF_1 scores after 6 training epochs (Sanh et al., 2019)

Results consistently favoured *lr*s of 5e-5 and 3e-5, with the highest micro-F_1 scores at lr 3e-5 with max length 128 (63.6%) and lr 5e-05 and max length 256 (63.1%) on the English validation set. However, the hF_1 score dropped to 56.9 on the English development set with the best model (see Table 2). Furthermore, we see a drop in the hF_1 on the development set when replacing the escaped newline with whitespace Table 2). We henceforth avoid any preprocessing of the input text.

We additionally compute the performance of mBART-50 (same hyperparameters) without using hierarchical information, where it gets a low mF_1 score on the validation set, but it outperforms the BERT-based models on the development set on hF_1.

Model	**Val** mF_1	**Val** hF_1	**Dev** hF_1
Performance without hierarchical information			
distilBERT (NoP)	63.6	/	56.9
distilBERT (NL-Spc)	63.6	/	50.4
distilBERT (NL-L)	63.6	/	50.4
mBERT (NoP)	59.2	/	/
mBART-50 (NoP)	48.9	59.9	59.9
Performance with hierarchical information			
distilBERT	/	60.1	**61.5**
mBART-25	/	59.8	60.4
mBART-50	**/**	**61.2**	61.0

Table 2: Performance comparison of baseline models on English dataset. Text preprocessing approach is in parentheses - no preprocessing (NoP), concatenating the lines with a single space character (NL-Spc) or using newline-separated lines (NL-L). mF_1 represents the micro F_1 score, and hF_1 indicates the hierarchical F_1 score. mBART-50 was our final submission during the official test phase.

3.2 Hierarchical Label Encoding

To include the hierarchical structure of the labels, we use the persuasion hierarchy digraph to expand the list of target labels such that it also contains all ancestor nodes. For example, [*Loaded Language*, *Name calling/Labeling*] is extended into [*Pathos*, *Loaded Language*, *Ethos*, *Ad Hominem*, *Name calling/Labeling*]. Since some labels have multiple parents, we consider all possible ancestors, and therefore [*Bandwagon*] gets extended into [*Logos*, *Justification*, *Bandwagon*, *Ethos*]. We compare the results of distilBERT and mBART-50 models compared to the approach without hierarchical label encoding. Additionally, we compute the results for mBART-25.

We internally test two approaches, a) one that extends labels for all training and validation examples and b) the one that extends the labels for training examples but not the validation examples. As extending the labels with ancestors only for the training examples consistently leads to better results, we proceed with this version.

As reported in Table 2, when using hierarchical information (extending the labels with ancestors for the training examples), mBART-25 achieves a hF_1 score of 60.4 and mBART-large-50 achieves 61.0 hF_1, on the English development set. The hyperparameters for these specific models had a learning rate of 5e-05, input size of up-to 128 tokens, and a batch size of 64.

Our results show that the hierarchical label encoding strategy consistently leads to performance improvements in this task compared to our baseline approach without hierarchical encoding, as showcased by distilBERT and mBART-50 models (see Table 2 for development set results). Note that distilBERT with hierarchy encoding results were computed in post-evaluation phase.

We submitted mBART-50 results (as it achieved the highest score on the validation set from the models we tested) for official test set evaluation for English. The model achieved 61.7 hF_1 score. We show the performance of the final model on all different language combinations in Table 4.

3.3 Translation and Test Set Results

For the three surprise test sets, we use the NLLB-200's 3.3B model (NLLB Team et al., 2022) to translate the English train and English validation sets into each target language to mimic the test stage scenario. We evaluate three settings (on the English validation sets, see Table 3): training on English and validating on translated data, training on translated data and validating on English data, and training and validating on translated data. We use the same hyperparameter grid search over the learning rates of [3e-5, 5e-5] and max lengths of [128, 256] to produce models fine-tuned for each language. Results indicate that training and validating with both target language translated sets consistently yielded better results when compared

to the other two settings (see Table 3). We use these fine-tuned checkpoints to infer the final submissions achieving results shown in Table 4.

Train	Validation	hF_1 score
EN_{train}	EN_{val}	61.2
$EN_{train} \rightarrow BG$	EN_{val}	54.9
$EN_{train} \rightarrow BG$	**$EN_{val} \rightarrow BG$**	**55.5**
EN_{train}	$EN_{val} \rightarrow BG$	47.1
$EN_{train} \rightarrow MK$	EN_{val}	53.3
$EN_{train} \rightarrow MK$	**$EN_{val} \rightarrow MK$**	**56.5**
EN_{train}	$EN_{val} \rightarrow MK$	51.3
$EN_{train} \rightarrow AR$	EN_{val}	56.2
$EN_{train} \rightarrow AR$	**$EN_{val} \rightarrow AR$**	**56.4**
EN_{train}	$EN_{val} \rightarrow AR$	50.6

Table 3: Evaluating the influence of translation as a strategy for handling low-resource languages. We measure mBART-50 model performance on Bulgarian (BG), Macedonian (MK) and Arabic (AR) by training on the English (EN) dataset translated to the target language using NLLB-200. The model is trained and evaluated on English data, possibly translated to the target language (translated dataset is shown with $\rightarrow$ followed by a target language).

Train	Validation	Test	hF_1 score
Final score on test data			
EN_{train}	EN_{val}	EN_{test}	61.7
$EN_{train} \rightarrow BG$	$EN_{val} \rightarrow BG$	BG_{test}	47.3
$EN_{train} \rightarrow MK$	$EN_{val} \rightarrow MK$	MK_{test}	36.2
$EN_{train} \rightarrow AR$	$EN_{val} \rightarrow AR$	AR_{test}	42.4
Zero-shot performance of the model			
EN_{train}	EN_{val}	BG_{test}	44.2
EN_{train}	EN_{val}	$BG_{test} \rightarrow EN$	44.2
EN_{train}	EN_{val}	MK_{test}	38.4
EN_{train}	EN_{val}	$MK_{test} \rightarrow EN$	33.8
EN_{train}	EN_{val}	AR_{test}	37.8
EN_{train}	EN_{val}	$AR_{test} \rightarrow EN$	36.4

Table 4: Evaluation of the model performance on the final test set. We measure mBART-50 model performance on Bulgarian (BG), Macedonian (MK) and Arabic (AR) by training on the English (EN) dataset. The model is trained and evaluated on English data translated to the target language (translated dataset is shown with $\rightarrow$ followed by a target language). We include final scores on test data achieved by the best-performing translation configuration. We additionally provide post-evaluation zero-shot performances of the model trained on the English data on the target language and the target language translated to English.

Using the approach where the model is trained on both train and validation data translated from English, we notice a significant degradation of the hF_1 scores for the three surprise test sets. The model for English did not use any translated data during training and validation and, as can be seen by comparing scores in Table 3 and Table 4, did not show signs of a similar degradation in performance on the test set.

This can be attributed to the distribution shift of persuasion label categories across the four languages. Our error analysis measuring accuracy for each label shows that the Arabic, Macedonian, and Bulgarian language models work well in identifying smaller classes with high accuracy while failing to generalize to the larger classes (see Table 5)[1]. These may arise from the model/training recipe failing to generalize over specific labels since different languages express persuasion strategies differently, and translations failing to capture some of these nuances.

Additionally, for our zero-shot performance evaluation, we use the model trained on English train and validation data to either directly predict the test datasets, or translating the test datasets to English and then predict. Here, we see that our translation approach works generally better than the zero-shot, except for the Macedonian dataset. This could be due to random seeds, however, larger comparable/parallel corpora are required to investigate this phenomenon.

Our final ranking on the SemEval Test Set Leaderboard are as follows: 17/32 for English, 8/20 for Bulgarian, 15/20 for Macedonian, and 4/17 for Arabic.

4 Conclusion and future work

In this paper, we describe the methods and models used by the *whatdoyoumeme* team in SemEval 2024 Subtask 1 to detect multilingual persuasion in memes. We combined two different approaches to solve this task: 1) machine translation, where we used the NLLB model (NLLB Team et al., 2022) to translate articles from English into the target languages and vice versa, and 2) including hierarchy information, where we extend a set of provided labels with labels corresponding to the ancestors nodes from the hierarchy DAG. We find that with the two proposed strategies, we can outperform both traditional encoder and decoder models, which emphasizes the importance of translation for downstream cross-lingual tasks.

In the future, we would like to extend our work

[1]All models generalize well over *Appeal to fear/prejudice* and *Distraction* but fail to generalize well over *Name calling/Labeling*.

in a few different directions. We would like to explore ensemble modelling techniques by building separate models for each belief category and using their joint predictions to improve overall performance. In addition, we would like to investigate the effects of translation quality and model size on the performance of this task.

Model	Label	Acc	Supp	Freq
mBART-50 Eng. Dev	Appeal to authority	95%	136	13.6%
	Repetition	94%	46	4.6%
	Distraction	92%	72	7.2%
	Simplification	79%	215	21.5%
	Name calling/Labeling	79%	262	26.2%
mBART-50 Arb. Test	Appeal to fear/prejudice	92%	8	8.0%
	Exaggeration/Minimisation	82%	18	18.0%
	Justification	79%	11	11.0%
	Name calling/Labeling	73%	26	26.0%
	Loaded Language	61%	24	24.0%
mBART-50 Mac. Test	Appeal to fear/prejudice	95%	13	5.02%
	Distraction	95%	11	4.25%
	Simplification	90%	10	3.86%
	Name calling/Labeling	63%	83	32.05%
	Loaded Language	61%	110	42.47%
mBART-50 Bul. Test	Appeal to authority	97%	18	4.13%
	Flag-waving	93%	28	6.42%
	Name calling/Labeling	68%	140	32.11%

Table 5: Error analysis of the model performance. Classes with higher accuracy are highlighted in green, while classes with lower accuracy in red. Only a selection of classes is shown, but a similar trend exists across all classes. Accuracy (*Acc*) is calculated using the standard binary accuracy measure. Support (*Supp*) is the number of instances from the dataset (*Train/Val/Dev/Test*) where the labels occur, and where the labels have been extended to include all ancestor nodes. Frequency (*Freq*) is calculated as support divided by the length of the dataset.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research and Innovation Agency through core research programme Knowledge technologies (No. P2-0103) and research projects: Embeddings-based techniques for Media Monitoring Applications (No. L2-50070) and Hate speech in contemporary conceptualizations of nationalism, racism, gender and migration (No. J5-3102). A Young Researcher Grant (No. PR-12394) supported the work of BK.

References

Tariq Alhindi, Jonas Pfeiffer, and Smaranda Muresan. 2019. Fine-tuned neural models for propaganda detection at the sentence and fragment levels. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 98–102, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Annenberg Public Policy Center. 2023. Democracy amid crises: How polarization, pandemic, protests, and persuasion shaped the 2020 election.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019a. 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, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Abdul Ghafoor, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Rakhi Batra, Mudasir Ahmad Wani, et al. 2021. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490.

Paul S Jacobs. 1992. Joining statistics with nlp for text categorization. In *Third Conference on Applied Natural Language Processing*, pages 178–185.

Kathleen Hall Jamieson, Matthew Levendusky, Josh Pasek, R Lance Holbert, Andrew Renninger, Yotam Ophir, Dror Walter, Bruce Hardy, Kate Kenski, Ken Winneg, et al. 2023. Democracy amid crises: Polarization, pandemic, protests, and persuasion.

Inigo Jauregi Unanue, Gholamreza Haffari, and Massimo Piccardi. 2023. T3l: Translate-and-test transfer learning for cross-lingual text classification. *Transactions of the Association for Computational Linguistics*.

Svetlana Kiritchenko, Richard Nock, and Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: Application to text categorization. volume 4013, pages 395–406.

Boshko Koloski, Blaž Škrlj, Marko Robnik-Šikonja, and Senja Pollak. 2023. Measuring catastrophic forgetting in cross-lingual transfer paradigms: Exploring tuning strategies. *arXiv preprint arXiv:2309.06089*.

Yuta Koreeda, Ken-ichi Yokote, Hiroaki Ozaki, Atsuki Yamaguchi, Masaya Tsunokake, and Yasuhiro Sogawa. 2023. Hitachi at SemEval-2023 task 3: Exploring cross-lingual multi-task strategies for genre and framing detection in online news. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1702–1711, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Teng-Sheng Moh and Zhang Zhang. 2012. Cross-lingual text classification with model translation and document translation. In *Proceedings of the 50th Annual Southeast Regional Conference*, pages 71–76.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Richard M Perloff. 2013. Political persuasion. *The SAGE handbook of persuasion: Developments in theory and practice*, pages 258–77.

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-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Elaine Zosa, Emanuela Boroş, Boshko Koloski, and Lidia Pivovarova. 2022. EMBEDDIA at SemEval-2022 task 8: Investigating sentence, image, and knowledge graph representations for multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1107–1113.