ShefCDTeam at SemEval-2024 Task 4: A Text-to-Text Model for Multi-Label Classification

Meredith Gibbons[1], Maggie Mi[1], Aline Villavicencio[1,2] Xingyi Song[1]
[1] Department of Computer Science, The University of Sheffield, UK
[2] Institute of Data Science and Artificial Intelligence, University of Exeter, UK
{magibbons1, zmi1, x.song}@sheffield.ac.uk
a.villavicencio@exeter.ac.uk

Abstract

This paper presents our findings for SemEval-2024 Task 4. We submit only to subtask 1, applying the text-to-text framework using a FLAN-T5 model with a combination of parameter efficient fine-tuning methods - low-rank adaptation and prompt tuning. Overall, we find that the system performs well in English, but performance is limited in Bulgarian, North Macedonian and Arabic. Our analysis raises interesting questions about the effects of label order and label names when applying the text-to-text framework.

1 Introduction

Social media platforms have become increasingly popular over time (Perrin, 2015). Whilst this enables greater public discourse, information and disinformation can also be presented purposefully to influence opinions online . Therefore, it is important to explore the detection of persuasion techniques. By fulfilling this goal, strategies that counteract false or misleading narratives can developed, and internet users can be empowered to think more critically about what they see online.

This paper describes our submission for SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes. We took a text only approach, and as such we only tackled subtask 1 - given only the "textual content" of a meme, our system must identify which persuasion techniques (of a possible 20) are used (Dimitrov et al., 2024). The labels are organized in a hierarchy (see figure 1) and multiple labels may apply to the same data point. For example:

> **Text:** HISTORY HAS SHOWN THAT THESE ARE THE FIRST TWO THINGS BANNED\\n\\nBY TOTALITARIAN GOVERNMENTS
>
> **Labels**: Loaded Language, Thought-terminating cliché

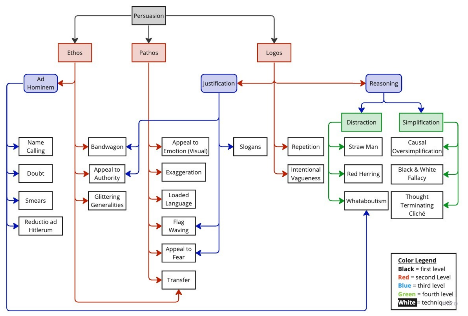

Figure 1: The hierarchical structure of the labels (Dimitrov et al., 2024).

In recognition of the diverse and intriguing use of language for manipulative communication, we target our exploration using a transformer-based architecture due to the ability of such models to capture linguistic intricacies (Plaza-del arco et al., 2023; Tenney et al., 2019). Specifically, we investigate this task using the text-to-text model FLAN-T5 (Chung et al., 2022).

2 Background

Research on identifying persuasion techniques in memes builds on the efforts of propaganda detection (Da San Martino et al., 2021; Dimitrov et al., 2021). Rashkin et al. (2017) trained models using n-gram TF-IDF feature vectors on a four category news reliability classification task. Barrón-Cedeño et al. (2019) both replicated the work of Rashkin et al. (2017) and applied n-grams to propaganda detection under binary classification. More recently, Da San Martino et al. (2019) took a more fine-grained approach. They developed a dataset of news articles with an annotation schema consisting of 18 propaganda techniques. They proposed a multi-granularity network using contextual embeddings derived with BERT (see also Da San Martino et al., 2020). Piskorski et al. (2023) presents a multilingual and multifaceted dataset of news articles, annotated with genre, framing and persuasion tech-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1860–1867

niques. They also evaluated the performance of a transformer model at various granularity levels - token-level, sentence-level, paragraph-level, and document-level.

To the best of our knowledge, there has been no work completed on exploring text-to-text (also known as sequence-to-sequence, or Seq2Seq) models for this multilingual, multi-label classification task in the domain of meme language. Text-to-text models take in text as input and output new text. Models such as T5 can be applied to many different tasks under the text-to-text framework (Raffel et al., 2019). They have also been shown to be effective in zero-shot settings (Chung et al., 2022; Plaza-del arco et al., 2023).

3 System Overview

We use FLAN-T5 (Chung et al., 2022) as our base model. FLAN-T5 was created by fine-tuning T5 (Raffel et al., 2019) on a mixture of tasks including text classification, question answering, and translation. The model regards every task as a text-to-text task.

We train in two steps:

1. LoRA; Low-Rank Adaptation (Hu et al., 2021)
2. Prompt Tuning (Lester et al., 2021)

For both steps all of the original FLAN-T5 parameters are frozen, lessening training time and hardware requirements. As both methods introduce their own set of distinct parameters, the LoRA parameters do not need to be trainable during prompt tuning. We first train using LoRA, then freeze the values of the introduced LoRA parameters and train using prompt tuning to produce the final model.

3.1 LoRA

Neural networks contain many dense layers, which transform input x to output h via matrix multiplication. Without model adaption, the pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$ produces output as follows:

$$h = W_0 x$$

After model adaptation, the updated output can be represented as follows:

$$h_{adapted} = W_0 x + \Delta W x$$

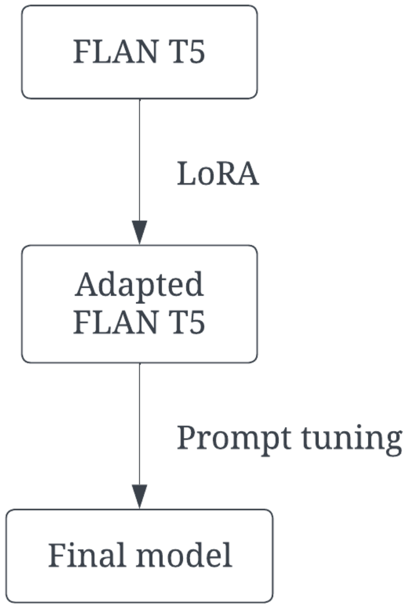

Figure 2: Training steps for our model.

where ΔW is the overall change to the weights, optimised during training. LoRA constrains ΔW by decomposing it into two low-rank matrices, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$, where $r << min(d, k)$:

$$h_{LoRA} = W_0 x + BAx$$

This process is summarised in figure 3. A and B are trainable parameters, initialised as a random Gaussian and 0 respectively to give an initial $BA = \Delta W$ of 0. $\Delta W x$ is scaled by $\frac{\alpha}{r}$, where α is a hyperparameter. Hu et al. (2021) applied LoRA to attention weights, achieving on par or better performance than full fine-tuning with only a fraction of the trainable parameters.

3.2 Prompt Tuning

In prompt engineering, a "hard prompt" is prepended to the input and used to guide the model to produce the desired output. Prompt tuning instead learns a "soft prompt", wherein the prompt tokens are taken as learnable parameters.

For input consisting of a token sequence $x_0, x_1, ..., x_n$, the tokens are first transformed to the embedding $X_e \in \mathbb{R}^{n \times e}$, where e is the dimension of the embedding space. The soft prompt, $P_e \in \mathbb{R}^{p \times e}$, where p is the length of the prompt, is concatenated to X_e to form new input matrix

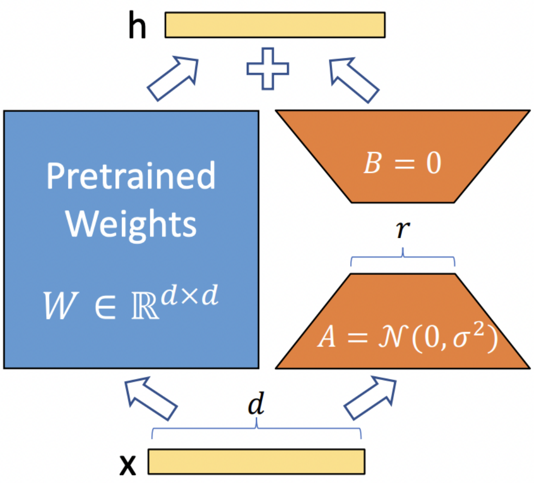

Figure 3: Overview of the LoRA method (Hu et al., 2021).

$[P_e; X_e] \in \mathbb{R}^{(p+n)\times e}$. During training, all model parameters are frozen and only P_e is optimised.

This method drastically reduces the number of required parameters, while achieving comparable performance to full fine-tuning when applied to very large models.

4 Experimental Setup

For hardware reasons, we use a sharded version of FLAN-T5-XXL[1] loaded in 8-bit precision.

The training set (size = 7000) was used for the LoRA training and the validation set (size = 500) was used for the prompt tuning.

Preprocessing was required to transform the data into an appropriate format for text-to-text training. We transform the input text to lower case, and for LoRA we prepended a simple task prompt. For example:

> NEW POLL\\n\\n82 percent of voters support TERM LIMITS ON CONGRESS\\n

becomes

> which persuasion techniques are in this text? text: new poll\\n\\n82 percent of voters support term limits on congress\\n

When preprocessing the labels, we observed that many original labels were metaphorical and/or lengthy, such as 'Glittering generalities (Virtue)'. Theorising that these sequences would be more difficult for the model to generate, we replace each label with a simplified (if applicable), lower case version. Finally, we concatenate the labels into a comma-separated list. Some examples are listed in table 1 - see Appendix A for a full list of simplified labels.

Original	**Preprocessed**
`['Bandwagon']`	`'appeal to popularity'`
`['Repetition', 'Name calling/Labeling']`	`'repetition, labeling'`
`[]`	`'none'`

Table 1: Examples of preprocessed labels for text-to-text training.

We use the PEFT implementation of LoRA and prompt tuning (Mangrulkar et al., 2022). For LoRA, we train for 5 epochs with a learning rate of 0.001. We mostly use the same hyperparameters for prompt tuning as Mozes et al. (2023) on T5-XXL. We initialise the prompt as:

> 'which persuasion techniques are in this text? text: '

More details on hyperparameters for both training steps can be found in Appendix B.

The evaluation measure used in this task is hierarchical F1 (Kiritchenko et al., 2006), which takes into account the tree structure of the labels when calculating model performance.

5 Results

Our final results are summarised in table 2[2]. Our English language result places us slightly above the centre of the leaderboard. Our Bulgarian result places lower, but is still superior to the baseline. Our North Macedonian result is below baseline performance. While FLAN-T5 was fine-tuned on a small number of Bulgarian language tasks during training, no North Macedonian language tasks were included. Likely due to the absence of Bulgarian and North Macedonian data in our training data and the small size of the corresponding test sets (size = 436 and 259 respectively), our results on these languages are much more variable than our English results.

[1] `https://huggingface.co/philschmid/flan-t5-xxl-sharded-fp16`

[2] All reported results obtained after the original task deadline.

	Hierarchical Precision
English	0.6701 ± 0.0025
Bulgarian	0.4631 ± 0.0069
N. Macedonian	0.4804 ± 0.0007
	Hierarchical Recall
English	0.6142 ± 0.0057
Bulgarian	0.2575 ± 0.0307
N. Macedonian	0.1882 ± 0.0160
	Hierarchical F1
English	0.6409 ± 0.0020
Bulgarian	0.3302 ± 0.0271
N. Macedonian	0.2700 ± 0.0164

Table 2: Hierarchical precision, recall, and F1 for our model on the test sets; average and range across two repeats.

The model failed to generalize to the fourth language, Arabic, despite its presence in the FLAN-T5 training data - we did not make a submission for this language as the model predicted no labels for all inputs.

5.1 Error Analysis

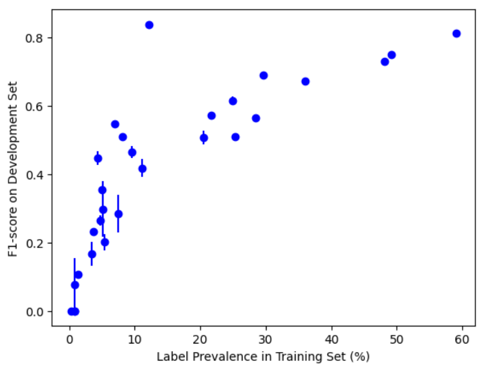

Figure 4: Prevalence of each label in the training set versus average F1 score on the English development set (size = 1000) over two repeats. Error bars show the range of values[3].

To investigate the errors of our model, we analysed the data on a per-label basis using our best performing language, English. Instead of using hierarchical F1, we split the multilabel task into 20 binary tasks (one for the prediction of each label) and calculated the average F1 score for each. In general, our system performed better on labels that were common in the training data (see figure 4). Several labels with very low training set prevalence had F1 scores of zero.

A notable result was the label 'Appeal to authority', which achieved a very high average F1 score of 0.838 while appearing in only 12.14% of the training data. Most data labelled with 'Appeal to authority' contains a quote, leading to the simplification of the label to 'quoting'. This clear pattern may have contributed to the higher average F1 score.

Other than 'Appeal to authority', the highest performing labels were non-leaf labels such as 'Ethos'[4]. These categories are very prevalent in the training data, so higher F1 scores are expected.

5.2 Further Analysis

We investigated two features of our system which may have affected the performance:

1. Ordered labels

2. Simplified label names

5.2.1 Ordered Labels

The text-to-text format necessitates that the labels be placed in an order (see table 1). This trains the model to associate an order with the labels - however, the order that the labels appear holds no semantic significance. For instance, "smears, slogans" is equivalent to "slogans, smears". In the data, there is a bias in the lists of labels in which certain labels ('Appeal to authority', 'Loaded Language', and 'Doubt') usually occur at the start. Labels such as 'Smears' usually occur at the end of the list, although the bias is not as strong as that of 'Appeal to authority'. Therefore, superfluous information may have been introduced to the model, decreasing the performance.

Alternatively, the model may leverage label order to reduce the number of possibilities while decoding, improving the performance. The typical positioning of 'Appeal to authority' at the start of the label list is another factor that may have made it an easier label to predict.

To investigate the effect of label order, we trained a separate version of our model, in which the labels of the training and validation sets (used for LoRA and prompt tuning respectively) were randomly shuffled. Our results are outlined in table 3[5], showing a slight increase in English hierarchical F1 and a much greater increase for Bulgarian and North Macedonian. This suggests that the bias in the label order may be detrimental to overall performance.

[3]As the range of F1 scores for some labels was zero or close to zero, not all error bars are visible.

[4]The model does not predict these labels directly. For the error analysis, the ancestor labels of each predicted label were added to the prediction in post-processing.

[5]All reported results obtained after the original task deadline.

	Hierarchical Precision
English	0.6978 ± 0.0031
Bulgarian	0.4362 ± 0.0117
N. Macedonian	0.4355 ± 0.0081
	Hierarchical Recall
English	0.6037 ± 0.0039
Bulgarian	0.3443 ± 0.0218
N. Macedonian	0.2724 ± 0.0228
	Hierarchical F1
English	0.6473 ± 0.0036
Bulgarian	0.3847 ± 0.0181
N. Macedonian	0.3349 ± 0.0196

Table 3: Hierarchical precision, recall, and F1 on the test sets for our model trained using shuffled labels; average and range across two repeats.

5.2.2 Simplified Label Names

Simplified labels (see Appendix A) were manually determined and focused on semantic simplicity and length. Despite this, many simplified labels were long in order to convey the concept of the persuasion technique, and some labels could not be easily simplified, being left with metaphorical or vague meanings.

To investigate the effect of the label names on performance, we compared the simplified label names with the per-label F1 scores. Table 4 shows the average per-label F1 score for the English development set and the prevalence of each label in the training set. As is also shown in figure 4, there is a correlation between average F1 score and training set prevalence. However, there are exceptions - 'virtue', the simplification of 'Glittering generalities (Virtue)', is a short and semantically obvious label and performs better than expected. Meanwhile, the longer and more metaphorical 'black and white thinking' has a lower average F1 score than expected.

This evidence suggests that longer and more complex labels may compromise text-to-text model performance, but more study is needed to reach a definitive conclusion. For example, the unusually high performance of 'quoting' is likely influenced by other factors. Some persuasion techniques may be easier or harder to detect regardless of label name.

6 Conclusion

In this paper we present a case study for the application of the text-to-text framework to multilabel classification. While our model exhibits some strengths, it did not achieve performance on par with top-ranking results. However, our analysis shows the potential for label names to affect performance, and suggests that shuffling labels during training may lead to increased performance.

Simplified Label	F1	Prevalence (%)
`quoting`	0.838	12.14
`loaded language`	0.616	25.00
`labeling`	0.574	21.69
`smears`	0.564	28.43
`virtue`	0.547	6.97
`appeal to identity`	0.509	8.16
`slogans`	0.464	9.53
`repetition`	0.447	4.36
`black and white thinking`	0.418	11.14
`doubt`	0.355	5.00
`exaggeration or minimisation`	0.298	5.09
`shutting down discussion`	0.285	7.54
`appeal to fear or prejudice`	0.265	4.81
`whataboutism`	0.232	3.69
`causal oversimplification`	0.167	3.43
`appeal to popularity`	0.108	1.39
`guilt by association`	0.077	0.90
`straw man`	0.000	0.89
`red herring`	0.000	0.84
`obfuscation`	0.000	0.30

Table 4: Simplified label names, average F1 score on the English development set over two repeats, and the prevalence of each label in the training set. The labels are ordered by average F1 score.

Limitations

Our paper has several limitations. Firstly, we only report results for our model across two repeats. This means that by chance, our results may appear to be better or worse than they would be on average. We only use English training data, which likely led to lower performance on the Bulgarian, North Macedonian, and Arabic test sets. Finally, we did not use a full-precision version of FLAN-T5-XXL due to hardware concerns. This likely led to decreased performance across all languages.

Acknowledgements

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. We also wish to thank the reviewers for their feedback and efforts.

References

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, Online. Association for Computational Linguistics.

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

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.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Svetlana Kiritchenko, Stan Matwin, Richard Nock, and A. Fazel Famili. 2006. Learning and evaluation in the presence of class hierarchies: application to text categorization. In *Proceedings of the 19th International Conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence*, AI'06, page 395–406, Berlin, Heidelberg. Springer-Verlag.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Maximilian Mozes, Jessica Hoffmann, Katrin Tomanek, Muhamed Kouate, Nithum Thain, Ann Yuan, Tolga Bolukbasi, and Lucas Dixon. 2023. Towards agile text classifiers for everyone.

A. Perrin. 2015. *Social Media Usage: 2005-2015: 65% of Adults Now Use Social Networking Sites–a Nearly Tenfold Jump in the Past Decade*. Pew Research Trust.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

A Simplified Labels

This appendix contains the simplified labels used in preprocessing. We did not remove all metaphorical references, leaving those which are relatively common (e.g. 'red herring') as FLAN-T5 is likely to have encountered them during training. As 'whataboutism' is difficult to explain succinctly, we left it as-is. All simplified labels are listed in table 5.

B Training Hyperparameters

Table 6 shows the training hyperparameters used in LoRA and prompt tuning. For our final output, we limit the length of the generated text to 20 tokens.

	Hyperparameter	**Value**
LoRA	Epochs	5
	Learning Rate	0.001
	Rank	16
	α	32
	Dropout	0.05
	Target modules	q,v
Prompt Tuning	Epochs	1
	Learning Rate	0.1
	Weight decay	0.00001
	Batch size	32
	Prompt tokens	10

Table 6: Hyperparameters used in LoRA training and prompt tuning.

Original Labels	Simplified Labels
Black-and-white Fallacy/Dictatorship	black and white thinking
Loaded Language	loaded language
Glittering generalities (Virtue)	virtue
Thought-terminating cliché	shutting down discussion
Whataboutism	whataboutism
Slogans	slogans
Causal Oversimplification	causal oversimplification
Smears	smears
Name calling/Labeling	labeling
Appeal to authority	quoting
Exaggeration/Minimisation	exaggeration or minimisation
Repetition	repetition
Flag-waving	appeal to identity
Appeal to fear/prejudice	appeal to fear or prejudice
Reductio ad hitlerum	guilt by association
Doubt	doubt
Misrepresentation of Someone's Position (Straw Man)	straw man
Obfuscation, Intentional vagueness, Confusion	obfuscation
Bandwagon	appeal to popularity
Presenting Irrelevant Data (Red Herring)	red herring

Table 5: Labels before and after simplification.