914isthebest at SemEval-2024 Task 4: CoT-based Data Augmentation Strategy for Persuasion Techniques Detection

Dailin Li[1], **Chuhan Wang**[1], **Xin Zou** [1], **Junlong Wang**[2], **Peng Chen**[1]
Jian Wang [1†],**Liang Yang**[1],**Hongfei Lin**[1]
[1]School of Computer Science and Technology, Dalian University of Technology, China
[2]School of Software, Dalian University of Technology, China
{ldlbest,wangchuhan,zouxin,jlwang,pengchen}@mail.dlut.edu.cn
{wangjian,liang,hflin}@dlut.edu.cn

Abstract

Memes are commonly used in online disinformation campaigns, particularly on social media platforms. They are primarily effective on social media platforms since they can easily reach many users. Semeval2024-Task4(Dimitrov et al., 2024), "Multilingual detection of persuasion techniques in memes", focuses on detecting persuasive methods across four languages: English, Bulgarian, North Macedonian and Arabic. Subtask 1 aims to identify the given text fragments of memes and which of the 20 persuasion techniques it uses, organized in a hierarchy. For the difficulty of this task and the fundamental role of text in the artificial intelligence area, we concentrate solely on this task. We develop a system using CoT-based data augmentation methods,in-domain pretraining and ensemble strategy that combines the strengths of both RoBERTa and DeBERTa models. Our solution achieved the top ranking among **33** teams in the English track during the official assessments. We also analyze the impact of architectural decisions, data construction and training strategies. We release our code at https://github.com/ldlbest/semeval2024-task4

1 Introduction

In the present digital era, persuasive communication is pivotal across diverse arenas, from political discourses to the viral spread of content on social media platforms. A nuanced comprehension of the intricacies of persuasion is indispensable in safeguarding against misinformation, upholding the integrity of information, and nurturing a constructive digital discourse.

In online communication, memes have become decisive for disseminating information and influencing opinions. The focus of this task centres on addressing the intricate task of identifying persuasive techniques within the textual content of memes. This paper addresses the "Textual Persuasion Technique Identification" task, emphasizing recognizing persuasive techniques within meme text. Our approach aims to deliver a robust multi-label classification system tailored to navigate the intricate challenges posed by this task.

We employ the Transformer (Vaswani et al., 2017) architecture. We introduce ensemble learning (Breiman, 1996), integrating one DeBERTa (He et al., 2021) model and four RoBERTa (Liu et al., 2019) models, each trained with different random seeds. In the pretraining phase of our system development, we utilize the in-domain pretraining method to improve our model's context and semantic comprehension. To bolster our dataset, we incorporate additional data from similar past tasks. Furthermore, we implemented the data augmentation technique, enhancing data diversity by employing data augmentation techniques.

Below is a summary of our contributions:

- We augment the training dataset with a Chain-of-Thought (CoT) based data augmentation method and improve our model's performance.
- Our system utilizes in-domain pretraining to enhance performance and leverages ensemble learning to combine DeBERTa and RoBERTa for further improvements.
- In task 1, we achieve first place on the English test set among **33** participants with an F1 score of **0.752**.

2 Background

2.1 Persuasion Techniques

Persuasive communication wields a critical influence across various sectors, including political rhetoric and the spread of online content. Such communication is instrumental in guiding public discourse and moulding opinions, ensuring its significance in the modern digital landscape (Yu et al.,

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1315–1321
June 20-21, 2024 ©2024 Association for Computational Linguistics

2021). This task extends these concepts to analyzing memes, an increasingly prevalent medium on social media and internet platforms. With their distinctive blend of fun and brevity, memes deftly navigate the web to share insights, provoke conversation, and distribute knowledge.

In our task, we concentrate on identifying persuasive techniques within textual content. According to the work of Piskorski et al. (2023), this task involves categorizing textual persuasive techniques into three subtypes: ethos, pathos, and logos. This taxonomy is further amplified to include 20 subordinate precise methods, providing an extensive framework for understanding and interpreting the art of persuasion in digital content.

2.2 Data Augmentation

Data Augmentation (DA) techniques are usually initially explored in computer vision (CV), but they have been relatively slow to gain traction in NLP. Challenges arise due to the discrete nature of language, which rules out continuous noise and makes it hard to maintain diversity (Feng et al., 2021). Although challenges exist, the evolution of NLP has led to an increasing demand for exploring tasks and domains with insufficient training data. Consequently, this trend has resulted in the proliferation of research studies utilizing DA techniques. One classical DA method is back translation (Sennrich et al., 2016), which involves translating the text into another language and then back into the original language. Wei and Zou (2019) proposed EDA to improve the performance of text classification tasks and exhibit solid results on smaller datasets. These techniques are helpful for augmenting data, but they only modify the original text in fundamental ways, sometimes even changing the entire meaning of the sentence. Additionally, Chen et al. (2023) proposed knowledge-guided data augment based on the semantic relations of the knowledge graph.

Recently, Large Language Models (LLMs) can provide a unified solution for various NLP tasks and achieve competitive performance (Zhao et al., 2023). For example, GPT-3 (Brown et al., 2020) and ChatGPT (Ouyang et al., 2022) have demonstrated strong performance in various NLP tasks and benchmark tests (Qin et al., 2023). Furthermore, LLMs play a role in data augmentation, enhancing their utility in multiple applications. Dai et al. (2023) introduced a text data augmentation approach based on ChatGPT, which can be used in downstream model training. Abaskohi et al. (2023) proposed Contrastive Paraphrasing-guided Prompt-based Fine-tuning of Language Models (LM-CPPF). To enhance the capacity of LLMs for intricate reasoning tasks, Wei et al. (2022) proposed Chain-of-Thought (CoT). Inspired by the effectiveness of the CoT method, we leverage CoT prompts to generate paraphrases used for data augmentation, ensuring the preservation of semantic consistency while significantly expanding our dataset. This augmentation strategy contributed to the enhanced performance of our model.

3 Our System

As depicted in Figure 1, our system comprises the following parts: Data Model, in-domain Pretraining, RoBERTa encoder, DeBERTa encoder and Soft Voting. The final prediction is obtained as $\hat{y}$. We ignore the hierarchical structure of the labels and define it as a multi-label classification problem (Tsoumakas and Katakis, 2007) for the labels of the training dataset are all final nodes of the graph.

3.1 Dataset Construction

We construct different data augmentation datasets based on various data augmentation strategies. In practice, while efforts to balance label distribution (such as using techniques like Nlpaug and CoT) aim to increase the number of samples for less frequent labels, it is essential to note that, since the data typically involves multiple labels, they can also result in the expansion of more frequent labels.

Nlpaug: We identify labels corresponding to train data with fewer than 1000 entries. Subsequently, we employ the nlpaug (Ma, 2019) library for these data points to implement data augmentation. Specifically, we utilize the method of synonym replacement, generating new training samples by substituting words in the text with their synonyms. This approach enhances the diversity of training data, thereby enhancing the model's robustness to different text inputs. Using this method, we augment more than 5700 data entries in total.

CoT-based Paraphrasing-Guided Data Augmentation: We filter data corresponding to labels that occupy less than 0.16 of the entire label distribution and rewrote these entries using GPT-3.5, generating 10,000 entries through this method. The LLM can fully understand the context and focus on improving the targeted content by explaining the labels and tasks and providing a specific description

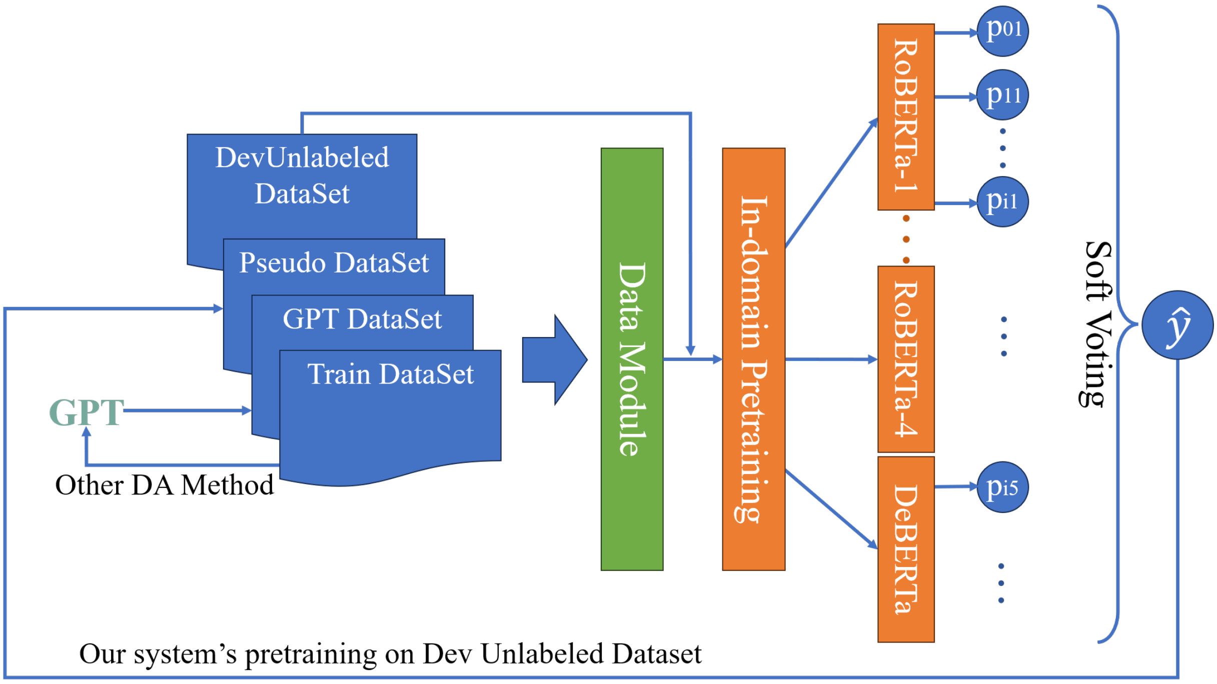
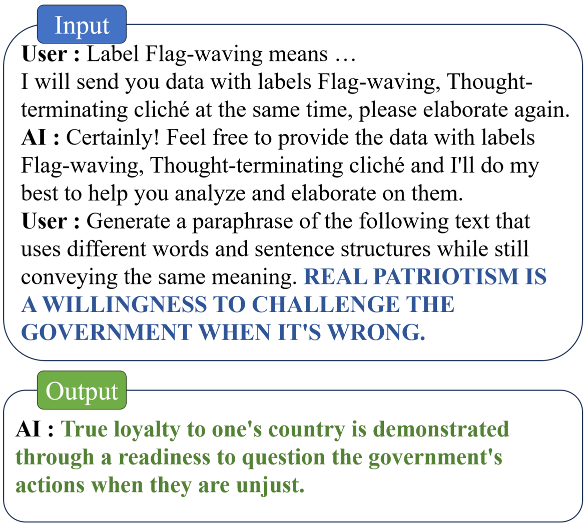


Figure 1: The overall architecture of our system.

of the problem and data that need to be rewritten. Applying the CoT technique enables the model to acquire more information and generate improved augmented data.


Input

User : Label Flag-waving means …
I will send you data with labels Flag-waving, Thought-terminating cliché at the same time, please elaborate again.
AI : Certainly! Feel free to provide the data with labels Flag-waving, Thought-terminating cliché and I'll do my best to help you analyze and elaborate on them.
User : Generate a paraphrase of the following text that uses different words and sentence structures while still conveying the same meaning. **REAL PATRIOTISM IS A WILLINGNESS TO CHALLENGE THE GOVERNMENT WHEN IT'S WRONG.**

Output

AI : True loyalty to one's country is demonstrated through a readiness to question the government's actions when they are unjust.

Figure 2: An example of using CoT for data augmentation.

We illustrate Figure 2. In the third round of our conversation with GPT-3.5, we use the instruction "Generate a paraphrase of the following text using different words and sentence structures while still conveying the same meaning" because it accurately describes the task with its instructions. Abaskohi et al. (2023) proved its effectiveness as an instruction template.

Pseudo-labeling: We use our model to classify 1000 data points on the dev dataset and 1500 on the test dataset. Pseudo-labelling (Lee, 2013) employs labelled data for training and utilizes information from unlabeled data to enhance the model's performance. The objective is to make more complete use of available data resources and improve the model's performance.

3.2 In-domain Pretraining

We utilize Masked Language Model (MLM) pertaining to all data, including data from SemEval 2023 task 3, which injects in-domain knowledge of our training datasets, thereby encouraging better learning outcomes for the model.

For a given input text x, we first tokenize it to obtain the tokenized representation $x_{tokenized}$ and truncate or pad them according to the maximum sequence length. For each text, a certain proportion of tokens are randomly masked based on the MLM probability and replaced with the masked token $[MASK]$. We use cross-entropy loss as the loss function for the masked language model. We compare the model's predicted probabilities for each token position with the actual token's one-hot encoding and calculate the cross-entropy loss.

Method	Recall	Precision	F1-score
BCAmirs	0.732	0.668	0.699
OtterlyObsessedWithSemantics	0.755	0.648	0.697
TUMnlp	0.714	0.638	0.674
GreyBox	0.688	0.652	0.670
BCAmirs	0.690	0.640	0.664
LomonosovMSU	0.632	0.674	0.652
NLPNCHU	0.706	0.604	0.651
Baseline	0.300	0.477	0.369
Our System	**0.836**	**0.684**	**0.752**

Table 1: Comparison of the performance between other team's models on Task 1 English test dataset.

$$L(t, \hat{t}) = -\frac{1}{N}\sum_{i=1}^{N} t_i \log(\hat{t}_i) \quad (1)$$

where t is the encoding of the true token, and $\hat{t}$ is the probability distribution of the model's predictions.

3.3 Ensemble Learning

We train four RoBERTa models and one DeBERTa model using different random seeds. We integrate them using the soft voting approach, which averages the predicted probabilities of each label from all five models. Given predictions $p_{i1}, ..., p_{iN}$ for class i from these models, we employ the following formula to obtain the final prediction $\hat{p_i}$

$$\hat{p_i} = \frac{1}{N}\sum_{j=1}^{N} p_{ij} \quad (2)$$

We then set a threshold of 0.25, where $\hat{p_i}$ greater than the threshold is chosen as the predicted label.

3.4 Low-Resource Languages

Since our training data is limited to English, we utilize GPT-3.5 to translate Bulgarian, North Macedonian and Arabic datasets into English. Subsequently, we perform inference on the translated data. The results obtained from these experiments can be found in the A.1. For the loss of information during translation, our system gets a relatively low F1 score in these languages.

4 Experimental Setup

The completion is based on PyTorch, Transformers and Pytorch-Lighting. During training, we set the batch size as 16, the learning rate as 3e-5, and the warmup steps ratio as 0.3. Five seeds(42,3407,114514,4096,1234) are used for the label ensemble. We use the AdamW optimizer and the cosine decay scheduler with a power of 0.01. We set a maximum epoch of 7. All experiments are run on one RTX 4090 GPU.

We create three additional datasets for the experiment: GPTDataset, PseudoDataset, and GPT-PseudoDataset (GPT-PDataset). The GPTDataset contains 7,500 training data and 10,686 sentences generated by LLM. PseudoDataset contains 7500 training data and test dev dataset labelled by the Ensemble model. GPT-PDataset is a union of GPT-Dataset and PseudoDataset.

5 Result and Analysis

In this section, we display our results and analyze the impact of each component through ablation studies.

5.1 Results

In this competition with 33 teams, we achieve first place with a hierarchical F1 score of 0.75247. We outperform the official baseline by 0.38382. The result is shown in Table 1.

We conduct a comparative analysis between our system and other models, including LLMs on the Task 1 English dev set, revealing the superior performance of our approach.

As illustrated in Table 2, GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023) utilized zero-shot learning, where only label meanings were provided in textual form without specific examples. We compare ourselves with other participating teams; the results are shown in Table 2. We achieve fourth place with a Hierarchical F1 score of 0.67833. Our performance is significantly better than the official baseline by 0.32010.

Method	Recall	Precision	F1-score
GPT-3.5	0.457	0.385	0.418
GPT-4	0.432	0.482	0.456
CLaC	**0.967**	**0.808**	**0.881**
OtterlyObsessedWithSemantics	0.754	0.636	0.690
GreyBox	0.716	0.657	0.685
EURECOM	0.702	0.650	0.675
Baseline	0.291	0.466	0.358
Our System	0.727	0.636	0.678

Table 2: Comparison of the performance between LLM and other team models on Task 1 English dev.

5.2 Ablation Study

We also conduct ablation experiments to validate our designs, including the encoder model, data modules, training strategy and ensemble.

Encoder Model We build our baseline model with BaselineDataset and BCE loss and run experiments to find out the best encoder model among RoBERTa (Liu et al., 2019), BERT (Devlin et al., 2019) , DeBERTa (He et al., 2021) , etc. As shown in Table 3, the large version of DeBERTav3 achieves the best score. Due to limited computility, we chose RoBERTa as our base model.

Method	F1-score
$BERT_{base}$	0.542
$BERT_{large}$	0.576
$RoBERTa_{base}$	0.614
$RoBERTa_{large}$	0.632
$DeBERTav3_{large}$	**0.649**

Table 3: F1 score of different Transformer-based models.

Training strategy We apply the in-domain pretraining on all encoder-based models to facilitate their performance on the downstream task. The result is shown in Table 4. For the five models, the F1 score improved by 0.2.

Method	F1-score
$BERT_{base}$	0.599
$BERT_{large}$	0.613
$RoBERTa_{base}$	0.630
$RoBERTa_{large}$	0.664
$DeBERTav3_{large}$	**0.667**

Table 4: F1 score of models after MLM training.

Data Module We use the best encoder model based on the result of dev datasets for ablation experiments on different datasets, including PseudoDataset, GPTDataSet, and GPT-PseudoDataset. The results are shown in Table 5.

Method	dataModule	F1-score
$RoBERTa_{large}$	GPTDataSet	0.685
$DeBERTav3_{large}$	GPTDataSet	0.700
$RoBERTa_{large}$	PseudoDataset	0.707
$DeBERTav3_{large}$	PseudoDataset	0.704
$RoBERTa_{large}$	GPT-PDataset	0.718
$DeBERTav3_{large}$	GPT-PDataset	**0.719**

Table 5: results on different dataset.

Ensemble Our ensemble approach can significantly improve performance. We integrate the results of different seeds and models based on their performance on the dev set. The result is shown in A.2, where we can see our ensemble approach outperforms the best single model by 0.15 F1 score over the dev set.

6 Conclusion

This paper details the architecture and performance of our multi-label classification system designed for the Persuasion Techniques Detection task. Our system achieves the highest rank for English in the leaderboard, signalling a notable accomplishment in the competitive framework. A comprehensive analysis of the data characteristics and model dynamics informs the strategic modifications we institute to the dataset construction and model training strategy. The efficacy of these refinements is corroborated by extensive empirical evaluation.

For future research, exploring methods to integrate the informational richness of hierarchical labels within the multi-label classification framework and fully exploiting LLMs to identify persuasion techniques remain promising avenues for further exploration.

References

Amirhossein Abaskohi, Sascha Rothe, and Yadollah Yaghoobzadeh. 2023. LM-CPPF: Paraphrasing-guided data augmentation for contrastive prompt-based few-shot fine-tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 670–681, Toronto, Canada. Association for Computational Linguistics.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher 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. *ArXiv*, abs/2005.14165.

Peng Chen, Jian Wang, Hongfei Lin, Di Zhao, Zhihao Yang, and Jonathan Wren. 2023. Few-shot biomedical named entity recognition via knowledge-guided instance generation and prompt contrastive learning. *Bioinformatics (Oxford, England)*, 39(8).

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

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. 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.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dong-Hyun Lee. 2013. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

OpenAI. 2023. Gpt-4 technical report. Available at `https://arxiv.org/abs/2303.08774`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

J. Piskorski, N. Stefanovitch, V-A Bausier, N. Faggiani, J. Linge, S. Kharazi, N. Nikolaidis, G. Teodori, B. De Longueville, B. Doherty, J. Gonin, C. Ignat, B. Kotseva, E. Mantica, L. Marcaletti, E. Rossi, A. Spadaro, M. Verile, G. Da San Martino, F. Alam, and P. Nakov. 2023. News categorization, framing and persuasion techniques: Annotation guidelines. Technical Report JRC132862, European Commission, Ispra.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *ArXiv*, abs/2302.06476.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6383–6389, Hong Kong, China. Association for Computational Linguistics.

Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. 2021. Interpretable propaganda detection in news articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1597–1605, Held Online. INCOMA Ltd.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

A Appendix

A.1 Model Result on Multilingual Datasets

Method	Recall	Precision	F1-score
English	**0.836**	**0.684**	**0.752**
Bulgarian	0.450	0.477	0.463
North Macedonian	0.340	0.401	0.369
Arabic	0.436	0.285	0.345

Table 6: Performance Metrics on Multilingual Datasets

A.2 Ensemble Model Result

Method	seeds	F1-score
$RoBERTa_{large}$	42	0.698
$RoBERTa_{large}$	3407	0.697
$RoBERTa_{large}$	4096	0.694
$RoBERTa_{large}$	1234	0.695
$RoBERTa_{large}$	1145145	0.696
$RoBERTa_{large}$	42,3407,4096	0.709
$RoBERTa_{large}$	42,3407,114514	0.710
$RoBERTa_{large}$	3407,4096,114514	0.710
$RoBERTa_{large}$	42,4096,114514	0.712
$RoBERTa_{large}$	42,3407,4096,114514	0.713
$RoBERTa_{large}$	42,3407,4096,114514	**0.718**
$DeBERTav3_{large}$	42	

Table 7: Ensemble of different methods and seeds